# A Multi-Party Dialogue Ressource in French

Maria Boritchev*, Maxime Amblard⊢

* Institute of Mathematics of the Polish Academy of Sciences Warsaw, Poland
⊢ LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France

## Dialogues in Games – DinG

A corpus of manual **transcriptions** of **real-life**, **oral**, **spontaneous multi-party dialogues** between **French**-speaking players of the board game *Catan*.

- Quality resource in view of studies such as [1];
- Study of questions in dialogue through annotation.

## Process

- Manual Segmentation in Speech Turns
- Transcription
- Anonymization
- Super-Annotatation

## Excerpt from DinG transcription, DinG6

```
009 Y j'aimerais bien faire 7 pour une
fois
      00:00:14.438 - 00:00:15.880
      (0.64)
010 R en fait t'as (te-) t'étais
contente parce que juste tu as fait
un double 6 et qu'en général c'est cool
dans les jeux [rire]
      00:00:16.518 - 00:00:21.910
011 Y ouais c'est ça
      00:00:21.712 - 00:00:22.718
012 R [rire]
      00:00:21.915 - 00:00:23.219
```

009 Y I would like to get a 7 for once
010 R in fact your have (y-) you were happy because simply you got a double 6 and generally it's cool in games [laugh]
011 Y yeah that's it
012 R [laugh]

## Quantitative data

| Name | Length (min) | Length (turns) | # questions | # turns /minute | # questions /minute | % questions among turns |
|---|---|---|---|---|---|---|
| DinG1 | **104.33** | **3,572** | **506** | 34.24 | 4.85 | **14.17** |
| DinG2 | 86.31 | 2,969 | 290 | 34.40 | 3.36 | 9.77 |
| DinG3 | 53.7 | 1,716 | 126 | 31.96 | 2.35 | **7.34** |
| DinG4 | 75.93 | 2,985 | 333 | 39.31 | 4.39 | 11.16 |
| DinG5 | 78.41 | 3,012 | 362 | 38.41 | 4.62 | 12.02 |
| DinG6 | 84.02 | 3,130 | 265 | 37.25 | 3.15 | 8.47 |
| DinG7 | 96.34 | 3,293 | 340 | 34.18 | 3.53 | 10.32 |
| DinG8 | **39.92** | 1,627 | 196 | **40.76** | **4.91** | 12.05 |
| DinG9 | 41.71 | **795** | **69** | **19.06** | **1.65** | 8.68 |
| DinG10 | 41.13 | **476** | **41** | **11.57** | **1.00** | 8.61 |
| Global data | 701.8 | 23,575 | 2,528 | 33.59 | 3.60 | 10.72 |
| CV | 34% | 47% | 57% | 29% | 40% | 20% |

Table 1:DinG data – observations per game, average on whole corpus and coefficients of variation (*CV*).

## Question annotation

| Tag | Name |
|---|---|
| YN | yes/no-question |
| WH | *wh*-question |
| DQ | disjunctive question |
| CS | completion suggestion |
| PQ | phatic question |
| N/A | non-assigned |

Table 2:Question tags [2].

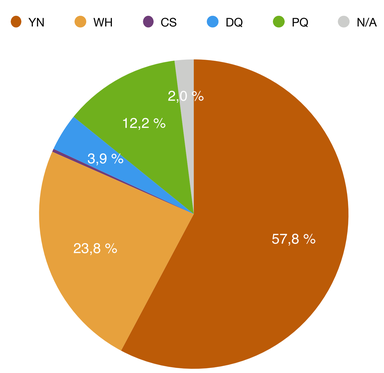


Figure 1:Distribution of the question annotation tags for DinG, on average, in percentage.

## Automatic annotation

| File | ID | Question | YN | WH | CS | DQ | PQ | N/A |
|---|---|---|---|---|---|---|---|---|
| ding3-1.txt.ufo | 961 | R:6 | | | | | | |
| ding3-1.txt.ufo | 962 | R:c'était limite hein | | | | | | |
| ding3-1.txt.ufo | 963 | Y:**combien j'en ai ?** | | 1 | | | | |
| ding3-1.txt.ufo | 964 | Y:4 | | | | | | |
| ding3-1.txt.ufo | 965 | Y:hum | | | | | | |

**R:** 6 **R:** it was borderline eh **Y: how many do I have? Y:** 4 **Y:** hum

(a)

| File | ID | Question | YN | WH | CS | DQ | PQ | N/A |
|---|---|---|---|---|---|---|---|---|
| ding3-1.txt.ufo | 985 | O:[dés] | | | | | | |
| ding3-1.txt.ufo | 986 | W:[rire] | | | | | | |
| ding3-1.txt.ufo | 987 | W:**pourquoi c'est toujours comme ça ?** | | 1 | | | | |
| ding3-1.txt.ufo | 988 | O:[dés] | | | | | | |
| ding3-1.txt.ufo | 989 | R:10 | | | | | | |

**O:** [dice] **W:** [laugh] **W: why is it always like that? O:** [dice] **R:** 10

(b)

Figure 2:Screen-shots and translations of the spreadsheet used to annotate questions from DinG with an automatic annotation of the central utterance.

## Conclusion & future work

For the corpus:

- Now publishing .txt and .eaf files
- Anonymize the oral data
- Comparative studies: STAC [1], ESLO [3], FQB [4]

For the question annotations:

- Gold annotation corpus
- Use more complex annotation schema such as ones from [1]

## References

[1] Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos.
Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus.
In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727, Portoroz, Slovenia, May 2016.

[2] Maria Andrea Cruz Blandon, Gosse Minnema, Aria Nourbakhsh, Maria Boritchev, and Maxime Amblard.
Toward Dialogue Modeling: A Semantic Annotation Scheme for Questions and Answers.
In *Proceedings of the 13th Linguistic Annotation Workshop (LAW XIII)*, 2019.

[3] Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua, and Isabelle Tellier.
Un grand corpus oral « disponible »: le corpus d'Orléans 1 1968-2012.
*Traitement automatique des langues*, 53(2):17–46, 2011.

[4] Djamé Seddah and Marie Candito.
Hard time parsing questions: Building a questionbank for French.
In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

## Acknowledgements

with the support of ANR-15-IDEX-04-LUE.

## Where to find the data?

The corpus is available on Gitlab, under CC BY-SA 4.0 license:
`https://gitlab.inria.fr/semagramme-public-projects/resources/ding/`