DinG – a corpus of transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of Catan

Maria Boritchev, Maxime Amblard

▶ **To cite this version:**

Maria Boritchev, Maxime Amblard. DinG – a corpus of transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of Catan. Journées LIFT 2021 - Linguistique informatique, formelle et de terrain, GDR lift, Dec 2021, Grenoble, France. hal-03537970

HAL Id: hal-03537970
https://inria.hal.science/hal-03537970

Submitted on 20 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DinG – a corpus of transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of *Catan*

Maria Boritchev* Maxime Amblard*
(*) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
{maxime.amblard,maria.boritchev}@loria.fr

MOTS-CLÉS : corpus, dialogue, transcription, questions, oral, français.

KEYWORDS: corpus construction, dialogue, transcription, multilogue, questions, oral, French.

1 General presentation of the corpus

We introduce a new corpus of manual transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of *Catan*[1], called Dialogues in Games (DinG), first presented in (Boritchev, 2021). *Catan* is a board game for three to four players in which the main goal for each participant is to make their settlement prosper and grow, using resources that are scarce. Bargaining over these resources is a major part of the gameplay and constitutes the core of DinG's data. The corpus has been designed to showcase the SLAM corpus (Amblard et al., 2014a,b, 2015), a corpus of interviews of patients with schizophrenia, while being widely available.

Dialogues from DinG are unconstrained, as the players don't have to follow any rule or specific guideline, apart from playing the game. As bargaining over the resources is part of the gameplay, the players have to speak in order to play, so the dialogues are the ones naturally occurring in this particular setting. As the players have to speak to play, they do not discuss personal subjects outside the game setting, which makes it possible to completely anonymize the corpus by removing the players' names (de-identification).

The recordings took place during university game nights. As we wanted the participants to feel as relaxed and natural as possible, the recordings were conducted in the room where the rest of the game night took place. Recording during the game nights raised some technical challenges, in particular, because different people were playing different games in the same physical space. Yet, it allowed us to record in a way that made the participants very comfortable: most of them report afterward that they forgot the recording devices after the first fifteen minutes of playing. All recordings were conducted by a non-player observer, whose duties were to explain the experiment, find volunteers and supervise the smooth running of the process. In particular, the observer had to manage the microphone and monitor the level of surrounding noise. We needed to address the technical challenge of recording our participants in a clear enough way for transcription, without recording other people's conversations. In order to do so, we used *H2 next handy recorder* by ZOOM[2], in XY (90° recording mode) setting.

Some of the participants knew each other as friends and/or colleagues, but in most of the games

[1]Copyright ©2017 CATAN Studio, Inc. and CATAN GmbH. All rights reserved.

[2]https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h2n-handy-recorder/

at least one player did not know the others at all. All participants are native French speakers. 33 people participated in the recording process, 12 women and 21 men. All participants but 3 had a master's degree or higher. Each participant only appears once in the corpus. We collected as little personal data as possible, but we can say that the average age of the participants is around 25 years old, and all the participants are native French speakers. All the participants signed an informed consent sheet, acknowledging they were giving us the right to record personal data (their voices) and share transcriptions of it.

The corpus was transcribed by paid transcribers, resulting in a high quality transcription. 6 transcribers took part in the project. 5 of them were recruited among natural language processing students, one is an expert in production and synchronization of subtitles. The transcription guide sets the norms to follow. The guide is an adaptation of (Blanche-Benveniste and Jeanjean, 1987). The main modifications are adaptations to the subject of our observation and the object of our research: (1) we specified the noise tags in order to adapt them to the board game context by adding tags such as [dice], [tokens]; (2) we added an explicit transcription of interrogative marks in order to account for utterances that were perceived (by the transcribers) as questions (rising intonation, answers given in the following dialogue turns). The transcribers who participated in the project have all received training on the same 5 minutes excerpt. Everyone did an individual segmentation and transcription before pooling and comparing the results.

The inter-annotator agreement for transcriptions is calculated on the transcription of a 5 minutes excerpt of DinG2, pre-segmented. Two independent annotators[3] (not working on the project before) have received empty segments for the excerpt and filled them with transcriptions, following the transcription guide. First, we computed the agreements for the full transcriptions, see the first two lines of table 1. It is important to stress that inter-annotator agreement on transcriptions is always low, as the amount of possible transcriptions is very large; yet, even taking this into account, the results we got were very low (under 0.3). Then, we computed the agreements for the transcriptions from which we removed the noises and the pauses. This produced lines 3 and 4 of table 1, with results higher than 0.5, which is usually considered to be a good agreement for transcriptions. This difference leads us to the conclusion that the quality of the recordings might be insufficient to grant an objective transcription of noises, on one hand, and also that transcriptions of the duration of pauses can vary from one transcriber to another.

	κ_{ipf}	**Raw agreement**
With noise, including unlinked/unmatched annotations	0.28	0.28
With noise, excluding unlinked/unmatched annotations	0.28	0.28
Without noise, including unlinked/unmatched annotations	0.52	0.55
Without noise, excluding unlinked/unmatched annotations	0.53	0.55

Table 1: Interrater agreement for transcription before/after noise tags and pauses removal, calculated with ELAN, following (Holle and Rein, 2013).

Transcribers are asked to respect scrupulously what is recorded/heard/said – they are not supposed to correct the language but to produce a faithful written version of French as it is used by the speakers. The writing of onomatopeias is normalized via lists (« euh », « hum », *etc.*). Pauses are explicitly marked with their approximate duration (ex: `(0.2s)`). Dysfluencies are also kept, in particular repetitions and beginning of words, that are marked with a hyphen (ex: « ca-(interrup-) carte » //

[3]We thank greatly Amandine Lecomte and Samuel Buchel for their contribution to our work.

"ca-(interrup-) card"). The transcription does not contain any punctuation except the interrogation point, which is used to annotate rising intonations that correspond to questions in the recording's oral context. As we are interested in questions and answers in dialogue, having an explicit annotation of questions is particularly useful for us.

It was of major importance for us to be able to distribute our resource while preserving the participants' private data. The last step in the transcription process is anonymizing the transcription. Each of the players is identified with the colour of their game pieces: Red (**R**), White (**W**), Yellow (**Y**) or Blue (**B**). If a name is pronounced out loud, it is replaced in the transcription by the name of the corresponding color, in upper case. Outside noises and speakers are assigned to an outside speaker called Other (**O**).

An average game of *Catan* lasts at least 30 minutes, thus DinG contains long interactions, going beyond informative exchanges. The corpus was originally designed to study human-human dialogue based on attested, spontaneous, and unconstrained oral data in French. Its nature allows for large dissemination and high cross-domain reusability. Its length allows for a study from different perspectives. The following shows an excerpt from the corpus[4]:

```
009 Y   j'aimerais bien faire 7 pour une fois
         00:00:14.438 - 00:00:15.880
         (0.64)
```

009 Y I would like to get a 7 for once

```
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
```

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   ouais c'est ça
         00:00:21.712 - 00:00:22.718
```

011 Y yeah that's it

The corpus is available on Gitlab:
`https://gitlab.inria.fr/semagramme-public-projects/resources/ding/`.
It is distributed under the Attribution ShareAlike Creative Commons license (CC BY-SA 4.0). Each game is available as a numbered .txt file, exported from ELAN[5] (Wittenburg et al., 2006).

2 Corpus description

DinG is composed of 10 recordings of games that last 70 minutes on average. The shortest recording is almost 40 minutes long (DinG8), the longest lasts a little over 1h44m (DinG1). Table 2 shows the first corpus measurements.

DinG1 is the longest both with respect to time and amount of speech turns; it also contains the biggest amount of questions. While DinG9 and 10 are not the shortest in terms of time, their amount of

[4]The participants are designated by the colour of their tokens: **Red** (**R**), **White** (**W**), **Yellow** (**Y**), **Blue** (**B**).
[5]`https://archive.mpi.nl/tla/elan`

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 2: DinG data – observations per game, on average and coefficients of variation (CV).

speech turns and questions are significantly (more than 10%) smaller than DinG8's (shortest in terms of time). This observation is supported by the fact that DinG9 and 10 present the smallest amount of speech turns per minute, while DinG8 presents the greatest: DinG8 lasts less time but DinG8's players talked at least twice more than DinG9 and DinG10's ones. Similarly, DinG8 presents the highest amount of questions per minute while DinG9 and DinG10 show the smallest ones.

The focus returns on DinG1 when we look at the percentage of questions among all the speech turns, as this game presents the highest percentage (the smallest one is shown by DinG3). DinG is homogeneous in terms of all the measures used in table 2, as all the coefficients of variation stay under 60%. While the amount of questions (the utterances marked with a '?') varies quite a lot from one recording to another, the percentage of questions among turns stays very similar (under 30%).

3 Future perspectives

We envision three perspectives for further development of DinG: its extension, its annotation, and its usage. A first step would be to transform it to fit the TEI format[6] (Parisse and Liégeois, 2020). Another path we envision is through the anonymization of the recordings through approaches such as the ones described in (Qian et al., 2017). Once the transcriptions, the oral data, and the participants' consent are available, a synchronization work would have to take place to enrich the resource. Then, transcription constitutes a first level of linguistic annotation. We would like to offer other annotations, at different linguistic levels: morphosyntactic, part-of-speech, disfluencies, syntactic (through universal dependencies, for example). We would also like to annotate on layers specific to dialogue: dialogue transactions, connectives, argumentation structures, in particular, throughout the annotation schemata that were developed for the STAC project (Asher et al., 2016).

Finally, this corpus can be used as a starting point for fine-grained analysis on the mechanisms underlying the articulations of questions and answers in French, such as the ones presented in (Boritchev and Amblard, 2021). A first step would be the inclusion of DinG in the French Question banks (Judge et al., 2006; Seddah and Candito, 2016).

[6] https://tei-c.org/Guidelines/

References

Amblard, M., Fort, K., Demily, C., Franck, N., and Musiol, M. (2015). Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues*, 55(3):91 – 115.

Amblard, M., Fort, K., Musiol, M., and Rebuschi, M. (2014a). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

Amblard, M., Musiol, M., and Rebuschi, M. (2014b). L'interaction conversationnelle à l'épreuve du handicap schizophrénique. *Recherches sur la philosophie et le langage*, 31:1–21.

Asher, N., Hunter, J., Morey, M., Benamara, F., and Afantenos, S. (2016). 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.

Blanche-Benveniste, C. and Jeanjean, C. (1987). *Le français parlé: transcription et édition*. Didier érudition.

Boritchev, M. (2021). *Dialogue Modeling in a Dynamic Framework*. PhD thesis.

Boritchev, M. and Amblard, M. (2021). Picturing questions and answers—a formal approach to slam. In *(In) coherence of Discourse*, pages 65–89. Springer.

Holle, H. and Rein, R. (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. *Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour, H. Lausberg, Ed. Frankfurt am Main: Peter Lang Verlag*, pages 261–277.

Judge, J., Cahill, A., and Van Genabith, J. (2006). Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 497–504.

Parisse, C. and Liégeois, L. (2020). Utiliser les outils CORLI de conversion TEI pour l'analyse de corpus de langage oral. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4: Démonstrations et résumés d'articles internationaux*, pages 64–65. ATALA; AFCP.

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X.-Y., Wang, Y., and Deng, Y. (2017). Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*.

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

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.