Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 814–823
Marseille, 20-25 June 2022

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
mboritchev@impan.pl, maxime.amblard@loria.fr

Abstract

We present *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*. Our objective is to make available a quality resource for French, composed of long dialogues, to facilitate their study in the style of (Asher et al., 2016). In a general dialogue setting, participants share personal information, which makes it impossible to disseminate the resource freely and openly. In DinG, the attention of the participants is focused on the game, which prevents them from talking about themselves. In addition, we are conducting a study on the nature of the questions in dialogue, through annotation (Cruz Blandon et al., 2019), in order to develop more natural automatic dialogue systems.

Keywords: Dialogue, Transcription, Annotation, Questions

1. Introduction

We envision our corpus as a corpus of spontaneous dialogues in French with a quality transcription. Its nature allows for large dissemination and high cross-domain reusability. Its length allows for a study from different perspectives. As the cost of producing resources is very high, we designed our corpus aiming for the widest possible use. Furthermore, we want to present it as a corpus that follows the good practices of corpus collection, usable for the collection of other corpora, especially ones that cannot be disseminated largely because of the nature of the data they contain.

One of the main inspirations in the design of DinG was the STAC project[1] corpus (Asher et al., 2016). The corpus of STAC is composed of chat logs from an online version of the board game *Catan*, played by English speakers. STAC is annotated in Segmented Discourse Representation Theory (SDRT, (Asher and Lascarides, 2003)). The proximity our two corpora share is a great opportunity to progress on comparisons between written and oral discourse on the same topics while opening on dialogue.

DinG is now used to study questions in dialogue in approaches such as (Boritchev and Amblard, 2020; Boritchev, 2021). As DinG is composed of long human-human interactions, it can be very largely used for dialogue studies in numerous fields.

Links with other Studies of Dialogue Another motivation for developing DinG is to showcase the work we are conducting on highly sensible health-related data. Studying DinG can help not to tag a phenomenon as a conversational disorder by acknowledging the fact that this phenomenon occurs in a non-pathological setting as well. Data from DinG is also used in the Sémagramme team (LORIA, Inria Nancy Grand-Est) to help develop and showcase SLODiM[2], a tool for dialogue analysis in a medical (psychiatric) setting, developed in the line of Schizophrenia and Language, Analysis and Modeling[3] (SLAM) (Amblard et al., 2014b), (Amblard et al., 2014a), (Amblard and Fort, 2014). The ODiM corpus follows the same transcription process as DinG. For the sake of science reproducibility, we developed DinG as a free-to-share corpus to showcase our studies, our tools, and our formal models.

To work on dialogic interaction, we need to study the dynamics of dialogic exchanges. Thus, we wish to have the transcription, but also the time codes corresponding to the dialogue. We have therefore developed a fine segmentation in addition to the transcription.

Link with other Researches with Catan The corpus was designed to study human-human dialogue based on attested, spontaneous, and unconstrained oral data in French. We want to study different types of dialogue-encountered phenomena; on one hand, the mechanisms underlying the combination of dialogue turns, in order to produce a computational model of dialogue, on the other hand, the dynamics of interactions, in order to account for meta-levels of human-human interactions. The proximity we share with the STAC corpus (Asher et al., 2016) appears to us as a great opportunity. Indeed, modeling natural language at discourse level is a task for which annotated resources are scarce. The Parallel Meaning Bank[4] constitutes a notable exception and is annotated in DRT (Kamp, 1981). The STAC corpus is annotated in SDRT (Asher and Lascarides, 2003), which means that the proximity our two corpora share is an opportunity to study fine-grained modeling phenomena, as much regarding the correspondence be-

[1] `https://www.irit.fr/STAC/`

[2] `https://slodim.loria.fr/`

[3] `https://team.inria.fr/semagramme/fr/slam/`

[4] `https://pmb.let.rug.nl/`

tween the two languages as for the logical models.

Ethics Recording during a game of *Catan* allows us to capture long spontaneous interactions that (almost) do not contain personal data. (Amblard et al., 2014a; Grouin et al., 2015) show that long interactions contain enough information to reduce the identification process to a (very) small amount of people. As the production of transcriptions of oral data is a very costly process, we want our corpus to be as widely sharable as possible. Therefore, following the recommendations of (Leidner and Plachouras, 2017), we decided to take care of ethical aspects by thinking them through the entire process of data collection and publication. The process was developed under the supervision and the validation of the Operational Committee for the Evaluation of Legal and Ethical Risks[5] (OCELER) of the INRIA, in due respect of the General Data Protection Regulation[6] (GDPR). 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. It was important to us to stress the fact that their consent was retractable at any point in the process. For now, we remove all mention of the names of the participants in the transcriptions and we publish transcriptions only.

2. General Presentation of DinG

Dialogues in Games (DinG) is a corpus of manual transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of *Catan*[7]. *Catan*, or *Settlers of 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.

(1) **A dialogue from DinG**

Yellow$_1$ Tu veux bien me donner un mouton?
Yellow$_1$ Would you like to give me a sheep?

Blue$_2$ Je veux bien euh un blé
Blue$_2$ I would like uh a wheat

Yellow$_3$ Et tu me donnes un mouton?
Yellow$_3$ And you give me a sheep?

Blue$_4$ Et je te donne un mouton
Blue$_4$ And I give you a sheep

Yellow and **Blue** are players of *Catan*, designated by the colour of their game tokens. In example 1, **Yellow** and **Blue** are bargaining over resources: **Yellow** seeks a sheep; **Blue** offers to give them one, but only in exchange for wheat. **Yellow** secures the bargain in speech turn **Yellow**$_3$; **Blue** confirms. 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.

A game of *Catan* A typical game of *Catan* takes place between 3 to 4 players and lasts between 30 and 90 minutes. The board is built using 19 hexagon terrain tiles of different colors: bright green pastures, yellow fields, grey mountains, brown carriers, dark green forests, and a desert tile. These tiles constitute the island of Catan, surrounded by water. See figure 1 for an example.

Figure 1: The game board during a game of *Catan*.

Most of the players we recorded for DinG never played *Catan* before. Therefore, each recording was preceded by a phase of rules explanation conducted by an observer. After this phase, the rule book was handed to the participants for them to be able to play autonomously.

Corpus Description The corpus 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). Most of the recordings (all but n°4, 5, and 6) were split into two parts because of a food break occurring during the game, as the recordings took place during university game nights. This division was kept in the transcription process, as it was easier for transcribers to work with shorter pieces of audio, as well as in the published data. The data presented in table 1 is computed on the merged recordings, one per game. We compute CV, the ratio of the standard deviation σ to the mean μ. A $CV < 100\%$ corresponds to a dataset with low variance.

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 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,

[5] https://www.inria.fr/en/operational-committee-assesment-legal-and-ethical-risks

[6] https://gdpr-info.eu/

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

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).

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 percentage is shown by DinG3. DinG is homogeneous in terms of all the measures we considered in table 1, as all the coefficients of variation stay under 60%. While the amount of questions (identified as the utterances marked with a '?') varies quite a lot from one recording to another, the percentage of questions among turns stays very similar (around 10%).

3. Building the Corpus

The recording part of the corpus collection took place during university game nights. 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.

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 afterwards 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.

3.1. Data Processing

Once a game is recorded, the raw audio file is given to transcribers. As our corpus was recorded in a noisy environment, transcribers had to pre-process the audio signal before starting to properly work on it. The pre-processing is done using Audacity[8] and aims to reduce the peaks in the sound signal (corresponding to loud noises such as rolling the dices) in order to then be able to uniformly amplify the whole signal and make the voices clearer. These treatments diminish the background noise while sharpening the voices of the participants. The next step is manual transcription.

Before choosing a transcription tool, we conducted a comparative study based on ergonomy, quality, and general characteristics of different specialized software. We also took into consideration the free availability of the software and their codes. We tested ExpressScribe[9], Astali[10], Youtube[11], TranscriberJS[12], oTranscribe[13] and ELAN[14]. These software were evaluated on each of the following aspects: sound manipulation, navigation inside the recording, supported formats, text manipulation, transcription tools, speaker annotation, dialogue act annotation, management of overlaps, noises, and inaudible fragments. The tool that got the best evaluation was ELAN (Wittenburg et al., 2006) because of its overall ergonomic design (for segmentation and transcription) in one tool while giving access to a visualization of the sound signal. Thanks to ELAN, DinG contains timecode alignment and disambiguation of speakers' overlaps. Among the different possibilities that ELAN offers, we used mainly segmentation and transcription to

[8] https://www.audacityteam.org/

[9] https://www.nch.com.au/scribe

[10] http://ortolang108.inist.fr/astali/

[11] https://support.google.com/youtube/answer/2734796?hl=en

[12] https://ct3.ortolang.fr/trjs/

[13] https://otranscribe.com/

[14] https://archive.mpi.nl/tla/elan

produce the final version of DinG.

3.2. Process

6 transcribers took part in the project, 5 of them were recruited among NLP students, one is an expert in production and synchronization of subtitles. They were trained for the task on a 5 minutes excerpt from DinG, that they all annotated and got to discuss with us and, when possible, between them. The transcribers were paid for the task. We counted 30 hours of work for the transcription of a recording of 1.5 hours. The transcribers followed different strategies, from minute-by-minute segmentation & transcription in parallel to the full segmentation and then full transcription, speaker by speaker, of the whole recording.

Manual Segmentation in Speech Turns The first step in the process of getting the transcription is a manual segmentation of the recording in speech turns, called *segments* in ELAN. We define a segment as a speech turn that is composed of a pseudo-sentence, an onomatopoeia, a noise, or any combination of the above. A speech turn is a theoretical linguistic unit corresponding to the verbal production of a speaker, (Schegloff et al., 1974). ELAN allows us to process overlaps in a very simple and visual way. Thus, each segment constitutes a coherent linguistic unit.

Transcription The transcription guide sets the norms to follow. The guide is inspired by (Blanche-Benveniste and Jeanjean, 1987), which has inspired many others transcription guides such as the one used within Transcriber(Barras et al., 1998). The (Blanche-Benveniste and Jeanjean, 1987) transcription guide advises not to use punctuation marks, so we explicitly added pauses duration and interrogative marks for our purposes, which makes us closer to guides such as Transcriber's one for French[15]. 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 transcription guide will be made available online. Furthermore, we produced a transcription and a segmentation, to preserve the dynamic aspect of interaction, for example by explicitly visualizing overlaps. Several automatic annotations were produced using SLODiM[16], in particular for disfluencies and syntax.

As mentioned before, 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, by two independent annotators (not working on the project before). They 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 2. κ_{ipf} is a modified version of Cohen's κ, computed using an *iterative proportional fitting* algorithm, it includes the unmatched annotations in the agreement calculation. The raw agreement is computed by dividing the number of agreeing cases by the total number of cases (Holle and Rein, 2015). 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 2, 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.

Anonymization 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 anonymization. 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 (e.g. "your turn, BLUE"). Outside noises and speakers are assigned to an outside speaker called Other (**O**).

Super-Annotatation Once the transcription is done, it is given to a super-annotator, whose goal is to proofread, homogenize the corpus via the correction of typos, standardization of the noise tags, of the transcription of onomatopoeias, of the writing of numbers[17] and of the anonymization. Table 2 shows the raise in transcriptions quality obtained after super annotation, computed on non-specialists' transcriptions: the last two lines correspond to the results obtained by calculating the agreement on the proofread transcriptions.

The Corpus The corpus is available on Gitlab[18]. 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. We export the transcriptions as `Traditional`

[15] `http://trans.sourceforge.net/en/transguidFR.php#r241`

[16] `https://academia.slodim.fr/`

[17] In French, « un » corresponds both to "a" and to "one". The guide asks for disambiguation, as much as possible. Following this idea, we transcribe « 1 », « 2 », etc. for amounts and « un » when the usage corresponds to the determiner.

[18] `https://gitlab.inria.fr/semagramme-public-projects/resources/ding/`

noise	unlinked/unmatched annotations	κ_{ipf}	**Raw agreement**
+	+	0.28	0.28
+	-	0.28	0.28
-	+	0.52	0.55
-	-	0.53	0.55
After super-annotation			
+	+	0.35	0.35
+	-	0.35	0.35

Table 2: Interrater agreement for transcription before/after noise tags and pauses removal, /after super-annotation, calculated with ELAN, following (Holle and Rein, 2013).

`Transcript Text` in order to generate text files that would be both readable by human observers and could easily be manipulated by scripts.

The files correspond to linearised versions of the games. Each segment appears on one line, that starts by the number of this segment in the transcription, then the letter that identifies the speaker (**R**, **W**, **Y**, **B** or **O**), then the transcription of what has been said. The next line contains the time codes of the beginning and end of the segment. When two following speech turns do not overlap, the gap between the two is calculated automatically and written in brackets on the next line. These elements are shown in Figure 2, see in particular the gap between `009` and `010`.

```
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'étai
        s 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]

Figure 2: Excerpt from DinG transcription and translation, DinG6.

4. Question Annotation

As we have indicated, the objective of the development of this resource is to make available long dialogues in French, but also to study their functioning. As dialogues are characterized by the use of questions and answers, we focused on analyzing them in DinG. As the questions were explicitly transcribed through question mark, we were able to automatically retrieve all of them along with a small context with the two preceding and the two following utterances. The questions were first automatically assigned tags, then annotated by hand to correct and augment this first annotation.

4.1. Annotation Scheme

Several annotation schema for different dialogue phenomena have been developed over the years.Following the developments we presented in the previous sections, we annotated the questions from DinG using an annotation schema for questions adapted from (Cruz Blandon et al., 2019), presented in table 3.

Our annotation schema is an easy-to-use one. We would like to extend the annotation of questions and answers through more detailed annotation schemas such as the ones developed for STAC (Asher et al., 2016) or inspired by insights from (Bazillon et al., 2011; Abeillé, 2013; Smirnova and Abeillé, 2021).

Our beginning assumption is that the corpora would contain at least two well-known and well-defined categories of questions: *yes/no*-questions and *wh*-questions. Some questions are similar to *wh*-questions or *yes/no*-questions in usage but have a different form: e.g., *wh-in-situ* questions such as "You saw what?", or *yes/no*-questions without inversion such as "You saw him?". We decided not to introduce new categories for these based on their semantics and pragmatics.

Some questions containing a disjunction (e.g. "Do you go on Monday or on Tuesday?") are semantically and pragmatically similar to *wh*-questions, but are syntactically closer to *yes/no*-questions. This kind of questionexhibits subject-auxiliary inversion (in English) but does not ask for the confirmation or denial of the proposition that it expresses. Instead, it expects the addressee to provide some missing information within the set of options to choose from. We call this type of questions *disjunctive questions*.

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

Table 3: Question tags.

Some questions have the syntactic characteristics of a *yes/no*-question or a *wh*-question, but are used with different pragmatics and/or semantics. For example, the speaker of the question can suggest a way to complete the utterance of the previous speaker, and the expected answer would confirm or deny this suggestion. This is subtly different from a prototypical *yes/no*-question because the speaker of the question does not necessarily ask their interlocutor to confirm the truth value of the semantic content of the suggestion. We call these types of questions *completion suggestions*.

Other questions take the appearance of a *yes/no*-question or a *wh*-question, respectively, but the context and intonation of the utterance make clear that the speaker is not actually interested in the confirmation or denial of the proposition. Instead, such questions can have various so-called *phatic* functions, i.e. their semantic content is less important than their social and rhetorical functions (Freed, 1994; Senft, 2009)). We call this type of questions *phatic questions*.

After first experiments with human annotators, we added a last category of automatic annotation: questions that we cannot assign a category to: `N/A`. It is used in particular if the complete utterance is « (xxx) ? », which is interpreted as "the person who was transcribing could not figure out any words but still picked up an interrogative/rising intonation".

4.2. Automatic Annotation

As phatic questions and completion suggestion are categories that are highly context-dependant, we did not pre-annotate them automatically. Some of the utterances contain multiple interrogative marks, corresponding to several questions asked in a row, without pauses between them, with several rising intonation points. The automatic annotation only annotated once.

The automated annotation of questions from DinG, called hereafter *utterance number n*, follows the next rules, in this order:

1. If the utterance number $n+1$ is affirmative (starts with the word « oui », « ouais » or « ok ») or negative (starts with the word « non »), the question in utterance n is automatically tagged as a *yes/no*-question: `YN`.

2. If the utterance contains a French *wh*-word , the question is automatically tagged as a *wh*-question: `WH`. (Boritchev and Amblard, 2021) gives a list of *wh*-words in French.

3. If the utterance contains « ou », the question is automatically tagged as a disjunction: `DQ`.

4. `N/A` otherwise.

The automatic annotation was able to assign a tag to 772 out of 2504[19] utterances containing at least one interrogation point, which corresponds to $\sim 31\%$.

The automatic annotation was systematically wrong in several cases: i) Several times, a *wh*-question is directly followed by a negation (see example 2). Following 1. in the automatic annotation rules, the question was assigned the `YN` tag in this case, while it should have been a `WH`. ii) A lonely « quoi ? » ("what?") is most of the time phatic (if it is the only content of the utterance). Following 2. in the automatic annotation rules, the question was assigned the `WH` tag in this case, while it should have been a `PQ`.

(2) ***Wh*-question followed by a 'no', DinG4**

W_{1150} **qui veut une pierre contre un mouton?**
W_{1150} **who wants a rock for a sheep?**

R_{1151} **non** mais vraiment
R_{1151} **no** but really

R_{1152} oui oui oui je euh je prends totalement
R_{1152} yes yes yes yes I uh I'll totally take [it]

A way to interpret example (2) is that **R** started the utterance $\mathbf{R}_{1151}$ without listening to $\mathbf{W}_{1150}$, to answer $\mathbf{B}_{1149}$. This hypothesis is supported by the content of $\mathbf{R}_{1152}$, which is an answer to $\mathbf{W}_{1150}$. This has to do with multi-thread conversation phenomena that we do not take into account in this annotation.

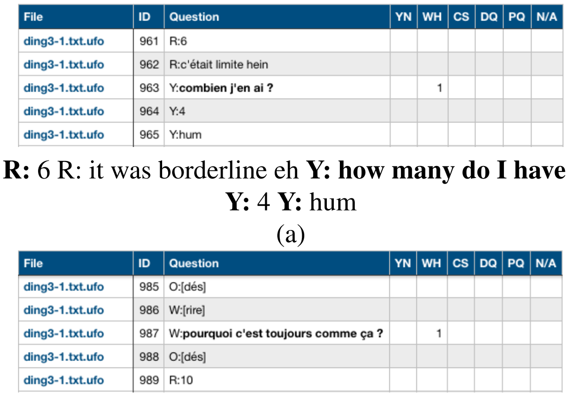
4.3. Human Annotation

The annotations was conducted by 10 people, among which 3 did the full annotation and 7 annotated subparts of the corpus. All the annotators but 2 are native speakers of French. Annotating the whole corpus took 6 hours to one human annotator, which is why most of the annotators only went through part of the task.

The annotations were performed using spreadsheets. Part of the questions were automatically pre-annotated, through the process presented in section 4.2.

In example (a), figure 3, **Yellow** is asking a *wh*-question introduced by the French *wh*-word « combien » ("how much"). The *wh*-word is identified by the automatic tagger, the '1' is put in the `WH` column. The human

[19] Note that 24 questions were removed from the process for technical issues.



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

annotator can validate this tagging by not changing it and proceeding to the next question.

In example (b), figure 3, **White** is asking a *wh*-question introduced by the French *wh*-word « pourquoi » (“why”). The *wh*-word is identified by the automatic tagger, the ‘1’ is put in the `WH` column. Yet, the human annotator can see, from the surrounding context, that **White**’s question, while *wh*- in its form, is actually phatic. **White** is in fact (fake?) complaining about something related to the game, most likely the result of the roll of the die. The human annotator needs to correct this tagging by moving the ‘1’ to the `PQ` column.

4.4. Results of the Human Annotation

Consequences of the Annotation The human annotators encountered several difficulties throughout the annotation. Annotator 1 was the first to conduct the full annotation. Their experience led us to implement the following decisions in the next annotations.

- Annotate question-tags (« j’ai fini, c’est ça ? » / “I finished, isn’t it?”) as `YN`.

- Separate the « ou pas » (or not) question tag from the others and tag questions finishing by « ou pas » as disjunctive, because they contain an “or” (« ou »). This decision is arguable and could be replaced by a tagging of this type of questions as `YN`.

- We annotate as phatic, unless clear from context that it’s not, the following questions: short questions such as « quoi ? » (“what?”), « c’est bon ? » (“all good?”), rhetorical/theatrical questions such as « sérieusement ? » (“seriously?”), « encore ? » (“again?”), « pourquoi c’est encore le 7 ? » (“why is it again the 7?”).

- A question in speech turn n is tagged as a completion suggestion only if the speech turn $n-1$ is uttered by a speaker different from n’s one. This comes from the fact that if a person pauses in the middle of a question, it can be split into two speech turns and create the false impression of a completion suggestion.

These decisions eased the work of the following annotators, however, they didn’t solve all the difficulties. The following section presents some of the cases that raised difficulties during the annotation process.

Annotation Results The annotation results obtained by the three annotators who worked on the whole corpus are presented in table 4. The inter-annotator agreement scores are shown in table 5. Annotator 1 was the first to annotate the whole corpus; after they turned in their annotation, the annotation guide was adjusted, as presented above. In particular, the definition of completion suggestions (CS) in the context of DinG was clarified. This explains the fact that Annotator 1 tagged no questions as CS, while the other found a few.

It is also interesting to note that the annotators did not annotate the same amount of questions even though they were all working with the same corpus. Our hypothesis is that this results from the utterances that contain two interrogation points: even though the convention was to annotate each of the questions separately, some examples seemed to be open to interpretation. Example 3 shows an utterance transcribed with two interrogation marks. $\mathbf{R}_{1074}$ was attributed two tags: a ‘1’ in the YN column, because the expected answer for the “and some clay?” part of the utterance is a ‘yes’ or a ‘no’; and a ‘1’ in the PQ column, as the second part of the utterance is used for metacommunication purpose, to put an emphasis on the first part or perhaps on the fact that the consent of the addressee matters.

(3) **Utterance transcribed with two ‘?’, DinG2**

Y_{1073} 1 mouton et autre chose oui
Y_{1073} 1 sheep and something else yes

R_{1074} **et de l’argile ? ça te va ?**
R_{1074} **and some clay? is that okay for you?**

Y_{1075} ça me va parfaitement
Y_{1075} it is perfectly okay for me

In general, it seems that tagging the questions as polar or *wh*- is an easier task than assigning the `CS`, `DQ`, or `PQ` tags. `CS` and `PQ` are categories that correspond to the pragmatics of dialogue, they are highly open to interpretation. The case of disjunctive questions is more interesting: they constitute on average less than 4% of the corpus, and they are quite likely to be confused with *yes/no*-questions because of a frequently encountered mixed form such as the one in example (4).

(4) ***Yes/no*-question with an embedded disjunction, DinG1**

	YN	WH	CS	DQ	PQ	N/A	Total
Automatic	494	262	-	16	-	-	772
Annotator 1	1,588	608	0	71	89	100	2,456
Annotator 2	1,389	602	17	115	364	26	2,513
Annotator 3	1,345	572	7	106	458	23	2,511
Average amount	1,441	594	8	97	304	50	2,493
Average percentage	57.78 %	23.82%	0.32 %	3.90 %	12.18%	1.99%	100%

Table 4: Annotation results for DinG's questions annotations of the 3 annotators who did all annotations.

R_{569} non moi je je
R_{569} no I I I

R_{570} **j'achète du mouton quelqu'un veut du (1s) blé ou du bois?**
R_{570} **I'm buying sheep does anyone want (1s) wheat or wood?**

B_{571} non
B_{571} no

In example (4), $\mathbf{R}_{570}$ starts as a *yes/no*-question with a do-support ("does anyone want"), but continues with a disjunctive part ("wheat or wood?"). The decision was taken to follow the top-most form and thus annotate this type of question with the YN tag.

Inter-annotator Agreement Table 5 presents the inter-annotator agreement scores for the three annotators that annotated all of the questions. As Cohen's κ measures agreement between two annotators only, we also computed Fleiss' κ for all three annotators. All the scores are quite high ($\kappa > 0.61$), but it is particularly interesting to notice that the agreement between annotator 2 and annotator 3 is higher than 0.8. Annotators 2 and 3 performed the annotation after the aforementioned modifications of the annotation guidelines, inspired by annotator 1's experience.

	Average Cohen κ	Fleiss κ
A1 + A2	0.651	-
A1 + A3	0.615	-
A2 + A3	**0.804**	-
A1 + A2 + A3	-	0.693

Table 5: Inter-annotator agreement scores for DinG's questions annotations, where A1, A2 and A3 are the three annotators that annotated all of the questions.

Three annotators annotated only the first half of the questions. Partial annotator 2 annotated in parallel with annotator 1, so the adjustments in the annotation guide mentioned above took place after partial annotator 2 turned their annotation in. Table 6 presents the inter-annotator agreement scores for the three partial annotators. All the scores are high ($\kappa > 0.77$), but it is particularly interesting to notice that the agreement between partial-annotator 1 and partial-annotator 3 is higher than 0.87. Partial-annotators 1 and 3 performed the annotation after the aforementioned modifications of the annotation guidelines, inspired by annotator 1's experience, while partial-annotator 2 annotated with the same guidelines as annotator 1.

	Average Cohen κ	Fleiss κ
P1 + P2	0.797	-
P1 + P3	**0.872**	-
P2 + P3	0.771	-
P1 + P2 + P3	-	0.813

Table 6: Inter-annotator agreement scores of the annotator which annotated the first half of DinG's questions.

5. Conclusion

The next step for questions in DinG is to produce a golden version. It will straightforwardly contain all the annotations for which the 6 annotators agree (from the first half of the questions). Then, we need to examine the annotations from the first half of the questions for which all annotators but A1 and P2 agree. In practice, we are waiting for more complete annotations before building and publishing the gold corpus.

Another aspect that we could not develop in this article due to lack of space is the comparison with other existing resources for French on the one hand, and for the Catane game on the other hand. For the French resources, the comparison with a large corpus such as ESLO[20] (Eshkol-Taravella et al., 2011) shows important similarities, with a smaller volume of questions. Another comparison is with the French QuestionTreebank[21] (FQB, (Seddah and Candito, 2016)), the reference corpus for questions in French. This corpus is built from governmental websites' FAQs. We expect to find multiple differences because DinG is intended to highlight spontaneous production. Finally, the comparison with the STAC corpus remains an important step for the analysis of the interaction dynamics.

In future work, we are considering anonymizing the oral data, following approaches such as (Qian et al., 2017). When we manage to do so, we will contact again the participants as they will have to sign a new consent for their data to be published. This step is very time-consuming, but can be done in parallel with the linguistic analysis. However, getting this data will increase the interest of the resource and its dissemination.

[20] http://eslo.huma-num.fr/

[21] http://alpage.inria.fr/Treebanks/FQB/

Acknowledgements

This work was supported partly by the french PIA project "Lorraine Universite d'Excellence", reference ANR-15-IDEX-04-LUE. We thank immensely Esteban Marquer, Julien Botzanowski, Laurine Jeannot, Léa Dieudonat, Srilakshmi Balard, Lucille Dumont, Amandine Lecompte and Samuel Buchel for their great transcription work. We thank greatly Amandine Decker, Bruno Guillaume, Pierre Lefebvre, Chuyuan Li, Léo Mangel, Siyana Pavlova, Guy Perrier, and Valentin Richard for their time and their annotation to our work. We also thank our reviewers for all the useful comments and references.

6. Bibliographical References

Abeillé, A. (2013). French questioning declaratives in question.

Amblard, M. and Fort, K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. In *TALN - Traitement Automatique des Langues Naturelles*, pages 292–303, Marseille, France, July.

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, November.

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. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

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, May.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Bazillon, T., Maza, B., Rouvier, M., Béchet, F., and Nasr, A. (2011). Qui êtes-vous ? catégoriser les questions pour déterminer le rôle des locuteurs dans des conversations orales (who are you? categorize questions to determine the role of speakers in oral conversations). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 83–93, Montpellier, France, June. ATALA.

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

Boritchev, M. and Amblard, M. (2020). There is as yet Insufficient Data for a Meaningful Answer. In *SemDial - WatchDial The 24th Workshop on the Semantics and Pragmatics of Dialogue*, Brandeis, United States, July.

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

Boritchev, M. (2021). *Dialogue Modeling in a Dynamic Framework*. Ph.D. thesis, Université de Lorraine.

Cruz Blandon, M. A., Minnema, G., Nourbakhsh, A., Boritchev, M., and Amblard, M. (2019). Toward Dialogue Modeling: A Semantic Annotation Scheme for Questions and Answers. In *Proceedings of the 13th Linguistic Annotation Workshop (LAW XIII)*.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., and Tellier, I. (2011). Un grand corpus oral « disponible »: le corpus d'Orléans 1 1968-2012. *Traitement automatique des langues*, 53(2):17–46.

Freed, A. F. (1994). The form and function of questions in informal dyadic conversation. *Journal of Pragmatics*, 21(6):621 – 644.

Grouin, C., Griffon, N., and Névéol, A. (2015). Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39.

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.

Holle, H. and Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior research methods*, 47(3):837–847.

Kamp, H. (1981). A theory of truth and semantic representation. *Formal semantics – the essential readings*, pages 189–222.

Leidner, J. L. and Plachouras, V. (2017). Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40.

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*.

Schegloff, E., Jefferson, G., and Sacks, H. (1974). A simplest systematics for the organization of turntaking for conversation. *Language*, 50(4):696–735.

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)*.

Senft, G. (2009). Phatic communion. In Gunter Senft, et al., editors, *Culture and language use*, pages 226–233. Amsterdam/Philadelphia.

Smirnova, A. and Abeillé, A. (2021). Question particles ça and donc in french: A corpus study. *Linguistic Research*, 38(2):239–269.

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.