DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter

Julian Hough[1], Ye Tian[2], Laura de Ruiter[3], Simon Betz[1], Spyros Kousidis[4],
David Schlangen[1], Jonathan Ginzburg[2]

[1]Dialogue Systems Group, Bielefeld University,
[2]Laboratoire de linguistique formelle, Université Paris-Diderot,
[3]School of Psychological Sciences, University of Manchester,
[4]Carmeq GmbH
julian.hough@uni-bielefeld.de

Abstract

We present the DUEL corpus, consisting of 24 hours of natural, face-to-face, loosely task-directed dialogue in German, French and Mandarin Chinese. The corpus is uniquely positioned as a cross-linguistic, multimodal dialogue resource controlled for domain. DUEL includes audio, video and body tracking data and is transcribed and annotated for disfluency, laughter and exclamations.

Keywords: disfluency, laughter, exclamations, multi-lingual, multimodal, spontaneous

1. Introduction

Natural, spontaneous dialogue corpora are rich resources for a variety of linguistic research. In this paper, we present the DUEL ('Disfluency, exclamations and laughter in dialogue' (Ginzburg et al., 2014b)) corpus,[1] consisting of 24 hours of natural, face-to-face, loosely task-directed dialogue in German, French and Mandarin Chinese.

The corpus is uniquely positioned as a cross-linguistic, multimodal dialogue resource controlled for domain, including audio, video and body tracking data and is transcribed and annotated for disfluency, laughter and exclamations.

To ensure cross-linguistic comparability, the experimental tasks were designed to be culture-neutral, the data in three languages were recorded using near-identical technical setups, and our transcription and annotation protocol is designed to be language-general.

In this paper, we give a summary of the tasks, the recording procedure and the transcription and annotation protocol. Then we discuss briefly the characteristics of our corpus, possible use cases and implications for natural dialogue research.

2. Existing Spontaneous Speech Corpora in Target Languages

Previous corpus work on spontaneous speech in German has focused on small domains and/or on speech data that does not generalize well to natural face-to-face dialogue. Kohler (1996) elicited dialogues using an appointment making scenario, but had speakers press a button to speak, eliminating any turn-overlaps (and potential disfluencies resulting from these). This is similar to (Burger et al., 2000), who used a similar scenario and instructed speakers not to interrupt each other. Schiel et al. (2012)'s non-intoxicated spontaneous control data was obtained by having participants talk to the experimenter in a car. Schmidt et al. (2010) and the Berlin Map Task Corpus (BeMaTaC)[2] both used map tasks, with the latter recording only non-native speakers. Peters (2005) collected a corpus of spontaneous speech by having two friends talk about video sequences via headset without eye-contact.

For French, there are several corpora for spontaneous speech. Several projects collected spoken French for studying prosody, for example, PFC (Durand et al., 2009), C-PROM (Avanzi et al., 2010) and Rhapsodie (Lacheret et al., 2014). Because of their research interests, these corpora cover a variety of discourse genres and do not focus on face-to-face dialogues. Bonneau-Maynard et al. (2005) collected the MEDIA corpus, containing roughly 70 hours of French dialogues on the topics of tourist information inquiry and hotel booking. It was recorded using a Wizard-of-Oz system where the participants interact with a human wizard they believe to be a machine. The C-ORAL-ROM corpus (Campione et al., 2005) contains 300,000 words of French formal and informal speech (along with the same amount of data in Italian, Portuguese and Spanish) in a variety of contexts, dialogue structures and text genres. There are also corpora where the speech is not completely spontaneous, for example, the French oral narrative corpus (Carruthers, 2013) is a collection of stories told by storytellers.

For Mandarin Chinese, work on spontaneous speech is sparser. The NCCU corpus of spontaneous Chinese (Chui et al., 2008) contains face-to-face conversations (not necessarily between two speakers) in three languages: Mandarin, Hakka, and Southern Min. The Mandarin sub-corpus contains about 3.5 hours of conversations. The Lancaster Los Angeles Spoken Chinese Corpus (Xiao and Tao, 2007) is a collection of dialogues and monologues in Mandarin Chinese, both spontaneous and scripted. Recently, The Chinese Academy of Social Science initiated the on-going project "Spoken Chinese Corpus of Situated Discourse", aiming to collect 1000 hours of spoken Chinese, covering different discourse genres and major dialects in China (cf.(Gu, 2000)).

[1] The DUEL project website is at: http://www.dsg-bielefeld.de/DUEL

[2] https://u.hu-berlin.de/bematac

The DUEL corpus is the first to provide French, Chinese and German sub-corpora in comparable spontaneous dialogue domains with a unified disfluency and laughter mark-up, making it of potentially great interest to the dialogue and speech research communities.

3. Corpus Construction

We recorded 10 dyads per language. Each dyad participated in three tasks, with the whole interaction lasting roughly 45 minutes in total.

3.1. Task design

We devised the tasks with three goals in mind, for them to: 1) be specific enough so participants do not spend significant time in silence working out what they should do, but unconstrained enough to allow free speech; 2) help elicit laughter and exclamations (we assume that as long as the conversations are spontaneous, disfluencies occur regularly); and, 3) elicit different types of laughter depending on the nature of the roles the participants have in the tasks: laughter of pleasure (*Duchenne* laughter) and laughter of embarrassment and other interaction management (social laughter). The three tasks used were as follows:

Dream Apartment First used in (Kousidis et al., 2013), the participant pairs are told that they are to share a large open-plan apartment, and will receive a large amount of money (500,000 Euros) to furnish and decorate it. The two participants are allowed their own bedroom but will share the rest of the apartment. They discuss the layout, furnishing and decoration decisions for 15 minutes. The participants sit across from each other in comfortable chairs.

Film Script This more open task requires the participants to spend 15 minutes creating a scene for a film in which something embarrassing happens to the main character. They are told that they can draw on their own experience. Again the participants sit across from each other in comfortable chairs.

Border Control This role-play interview task is the most constructed. One participant plays the role of a traveler attempting to pass through the border control of an imagined country, and is interviewed by an officer. The traveler has a personal history and situation that disfavours them in this interview (for example having a criminal record and carrying illegal substances). The officer asks questions that are general as well as specific to this traveler. In addition, the traveler happens to be parent-in-law of the officer. For this task, the two participants receive separate information regarding their character roles, and the task is not timed – it ends when they feel that the interview is finished, though no pairs exceeded 20 minutes. The purpose of this task is to bring in an element of power asymmetry in the roles of the participants, while the other two tasks can be considered symmetrical. To effect an interview setting, the participants sit either side of a table.

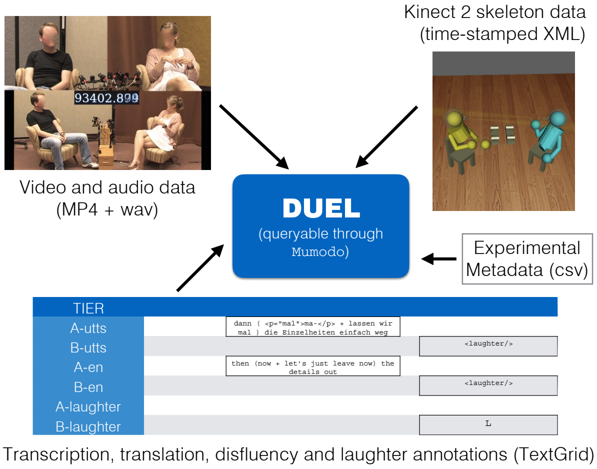

Figure 1: The schematic structure of the DUEL corpus

After the three tasks, the participants complete a questionnaire about the pair's relationship (whether or for how long they know each other, and the frequency of contact) and how they felt about the tasks: how much they understood each other, and to what extent they felt uncomfortable or embarrassed during each task. This meta-data is available with the corpus.

3.2. Languages and participants

There were 10 pairs of native speakers for each of the three languages: German, French, and Chinese. The German speakers were all students at Bielefeld university where 3 pairs were friends/acquaintances and the remaining 7 strangers. The French speakers were students at Université Paris Diderot– 5 pairs were friends/acquaintances, and 5 were strangers. Among the 10 pairs of Chinese speakers, 7 pairs were university students in Paris, and 3 pairs were recruited via a local Chinese forum and, again, 5 pairs were friends/acquaintances and 5 were strangers. Participant gender was not controlled for.

3.3. Recording setup

The German data was recorded at Bielefeld University. The French and Chinese data were recorded at Université Paris Diderot. MINT tools (Kousidis et al., 2013), a toolkit for multi-modal recording was used to ensure synchronization of the various data sources, which included high-quality audio, video and body tracking data.

The video data was filmed using two cameras to capture the gesture space and face of both participants, close lapel microphones were used to capture excellent audio quality without being intrusive, and the body movement was tracked by a Microsoft Kinect 2– the decision to track body movements without using a motion capture system requiring a suit was again to maximize naturalness.

The body tracking data was logged as time-stamped skeleton coordinates for the two participants into a standardized XML format using the *Venice.hub* logger (Kennington et al., 2014), a format which can be easily interpreted and queryable through the freely available Mumodo analysis tool kit.[3]. The schematic overview of the corpus can be seen in Figure 1

[3] Available from:
https://github.com/dsg-bielefeld/mumodo

4. Transcription and Segmentation

Transcription was done from the WAV audio files using Praat (Boersma and Weenink, 2010), following the instructions of the DUEL transcription and annotation manual (Hough et al., 2015). The manual specifies language general practices such as segmentation, disfluency annotation and laughter annotation, as well as language specific instructions regarding filled pauses, exclamations, and non-standard orthography.
For the transcription, the following four tiers are available for a given participant X. The transcriptions are thus aligned with the audio in terms of turns, utterances and laughter episodes.

X-turns tier containing the turn boundaries for participant X

X-utts tier used for segmentation and transcription of X's utterances

X-en tier containing English paraphrase translation for X's utterances

X-laughter tier containing the laughter and laughed speech duration intervals for participant X

Segmentation: In the *X-turns* tiers, all continuous stretches of speech by one speaker until the other speaker takes over, modulo small overlaps, are considered one turn. The only exception is when it really sounds as if the speaker finishes and waits for the interlocutor, and only resumes speaking after a silence during which the interlocutor does not start speaking. Cases like this would be annotated as two consecutive turns by the same speaker.
In the utterance tiers *X-utts*, we follow Meteer et al. (1995)'s notion of a *slash unit*, defining the notion of utterance as "maximally a sentence but can be a smaller unit [...] Intuitively, slash-units below the sentence level correspond to those parts of the narrative which are not sentential but which the annotator interprets as complete". Leading discourse markers such as "ok", "right" and "so" are regarded as individual utterances (unless "so" can be replaced by "hence"). Leading filled pauses such as "um" or "uh" are not cut off but considered part of longer utterances. Restart disfluencies are not cut off but considered part of longer utterances. For example, "(because he +) well I think I need to call him". Here the speaker abandoned the disfluency "because he", and restarted the utterance with "well I think I should call him". In this case, the abandoned disfluency is considered part of the utterance.

5. Disfluency, Laughter and Exclamation Annotation

Our annotations follow the light-weight inline method of dialogue annotation described by Hough et al. (2015).

Disfluency: We consider disfluencies anything that leads to an audible deviation from expected speech production. We annotated the following phenomena: silent pauses, lengthening, filled pauses and editing terms, repairs, abandoned utterances and restarts.
For silent pauses, we transcribed pauses of short, medium and long duration, using one, two and three dots respectively. Lengthening was transcribed using the symbol "`:`" following the lengthened syllable(s), e.g. `u:m:`.
We mark filled pauses by a {`F` }, bracketing other fillers and editing terms simply with { } - e.g. `I { you know } like her`.
The inventory of editing phrases and filled pauses differ depending on the language. For example, in German, the common filled pauses are {F äh}, {F ähm} and {F hm}; in French they are {F euh}, {F mmh} and {F euhm}; in Chinese, they are {F en}, {F eh}, as well as demonstratives {F nage} (literally "that") and {F zhege} (literally "this").
For repairs, restarts and abandoned utterances, we mark the structure according to this scheme, consistent with the Switchboard repair mark-up (Meteer et al., 1995):

(*reparandum* + { *editing term* } *repair*)

Both the editing term (which can be a filled pause) and the repair are optional. The structure can be nested and can appear in any position in an utterance, as in the following examples:

(1) *Standard repair:* I went to (the: + {F um } the) garden
(2) *Nested:* (I + (I + I)) want to go to Berlin
(3) *Restart:* (I + {F uh }) yesterday someone said yes to that

For partial words, transcribers were encouraged to guess the complete standard form of the word where possible, using a simple XML-style tag structure `<p s="standard form">partial form</p>`, as below:

(4) (<p s"Wohnzimmer">Wohn-</p> + . { ja also } (die + (die + das)) {F äh} ... Wohnzimmer)
(<p s="living room">liv-</p> { yes well } (the + (the + the)) {F uh } living room)

Laughter: We distinguish laughter concurrent with speech (laughed speech) and standalone laughter bouts. The former is transcribed again with simple XML-style tags spanning the affected speech, e.g. `<laughter>...</laughter>`, and the latter is marked `<laughter/>`. A `<laughterOffset/>` tag as in (5) is used for the often audible deep inhalation of breath after laughed speech or a bout.

(5) (Und mit einem +) mit vielleicht Sachen die nicht
<laughter> auseinander brechen < /laughter>
<laughterOffset/> -
(And with a +) with perhaps things that don't
<laughter> fall apart
</laughter> <laughterOffset/> -

In addition, in the *X-laughter* tiers, the intervals over stretches of laughed speech and laughter bouts for a given participant are marked.

Exclamations: We mark any exclamative short utterances with a simple bracketing as with filled pauses and editing phrases, for example, {`X ohlala` } in French. Compared to disfluencies and laughter, exclamations were sparse in our corpus, but the investigating the differing forms and contexts of occurrence between languages is a fruitful area of cross-linguistic research.

Example 1 (Chinese): Chaining repeat repair:

A	就	感觉	客厅	是	(((公+公)+公)+公共:)	{F jiushi }	休息	啊	或者
A-en	*then*	*feel*	*living room*	*is*	*(((public+public)+public)+public:)*	*{F that is }*	*relax*	*PRT*	*or*

Example 2 (German): Chaining substitution repair after laughed speech:

A dann hat jeder genug Privatsphäre .. mit seinem <laughter> Partner </laughter>
(und die Küche + (und die + {F ähm } (und die + ... und das Wohnzimmer))) ist quasi so ... mittig

A-en *then everyone has some privacy ... with their <laughter> partner </laughter>*
(and the kitchen + (and the + {F um } (and the + ... and the living room))) is kind of ... central

Example 3 (French): Restart disfluency within laughed speech:

A bah quand même <laughter> c'est (un chien +) deuxième étage </laughter>

A-en *well still <laughter> it is (a dog+) second floor </laughter>*

Figure 2: DUEL's disfluency and laughter mark-up in the three languages in the Dream Apartment task

Non-standard pronunciation: In the data of all three languages, there are frequent pronunciations that deviate from the standard forms of words. Very often, the deviation is conventionalized in everyday speech, but when the pronounced form is noticeably unconventional the word is annotated by the transcription of both the pronounced form as well as the form in standard orthography as in `<v s="standard form">pronounced form</v>`. For example, <v s="auf der">aufer</v> (German), and <v s="il faut">'faut</v> (French). Similarly, for obvious mis-pronunciations the mis-pronounced form and the standard form are annotated as in `<m s="standard form">pronounced form</m>` (e.g. <m s="angry">angly</m>).

Non-verbal utterances For monosyllabic functional utterances we use `hm` (with no brackets) and for the disyllabic equivalent we use `mhm`. As regards other non-verbal vocalizations, apart from laughter and breathing and these functional utterances, for non-linguistic contributions such as coughing, sneezing and lip-smacking, we use a `<nonverbal/>` tag.

6. Use cases

DUEL's light-weight and consistent mark-up of the above phenomena allows for fast searching of the utterance tiers, and example utterances from the Dream Apartment task with the mark-up across the three languages can be seen in Figure 2. The mark-up exhibits good inter-annotator agreement and it is compatible with several existing schemes – see Hough et al. (2015) for details.

The annotations can be used for doing fine-grained qualitative analysis of these phenomena, for example on formal characterizations of editing phrases as in (Tian et al., 2015), or forward and backward-looking disfluencies (Ginzburg et al., 2014a) in addition to quantitative work, for example on repair rates (Hough and Purver, 2013).

The audio and transcription and annotation data can be used in conjunction with the body-tracking data – the Dream Apartment task has been used in a study on the multimodal aspects of laughter by Kousidis et al. (2015) and continues to be used for multimodal dialogue and laughter studies. Given the exact timings of laughter are available, phonetic analysis of laughter is possible, in line with the requirements listed by Truong and Trouvain (2012).

For automatic disfluency processing, approaches such as (Zwarts et al., 2010; Hough and Purver, 2014; Hough and Schlangen, 2015) can be employed due to the consistency with the Switchboard disfluency mark-up.

7. Limitations

Our setup also has limitations. The cooperative (rather than competitive or argumentative) interaction between the interlocutors may restrict the type of disfluencies, exclamations and laughter exhibited. Particularly, in the power-asymmetric *Border Control* task, we note the prevalence of meta-situational laughter about the strangeness of the task, rather than laughter deriving from the situation of the characters the participants were role-playing. Furthermore, for quantitative analyses, the small number of stranger/friends pairs as well as the small number of male/female, male/male and female/female pairs make it hard to analyse the role of acquaintance and gender in any communicative phenomena. The DUEL corpus would benefit from a larger numbers of dyads which balances these features.

8. Availability and Searchability

The anonymised transcripts and movement data are available under a public PDDL license (`doi:10.4119/unibi/2901458`). For access to the audio files, an individual license agreement is needed– please contact one of the first two authors.

We have a Python interface for searching through the whole corpus, which is the latest version of the Mumodo analysis toolkit, available at `https://github.com/dsg-bielefeld/mumodo`.

9. Conclusion

We have presented the DUEL corpus, a multi-lingual, multi-modal data-set that is uniquely positioned for dialogue and spontaneous speech research, both in terms of the consistency of the domain across languages, its standardization of disfluency and laughter mark-up and its synchronized multimodal data.

10. Acknowledgements

We thank Israel de la Fuente for his assistance the setting up of the multi-modal recording of the Chinese and French corpora. We thank Oliver Eickmeyer, Michael Bartholdt, Gerdis Anderson, Florian Steig, Angelika Maier, Charlotte Hauser, Emmanuelle Bonnet, Nicolas Benichou, Jingshu

Liu and Yunfang Lu for their work on transcription and annotation.

This work is supported by the DUEL project, supported by the Agence Nationale de la Research (grant number ANR-13-FRAL-0001) and the German Research Foundation (DFG) (grant number SCHL 845/5-1). This work is supported by the German Research Foundation (DFG) through the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University.

References

Avanzi, M., Simon, A.-C., Goldman, J.-P., and Auchlin, A. (2010). C-prom: An annotated corpus for french prominence study. In *Proceedings of Prosodic Prominence, Speech Prosody 2010 Workshop*.

Boersma, P. and Weenink, D. (2010). Praat: doing phonetics by computer.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005). Semantic annotation of the french media dialog corpus. In *Ninth European Conference on Speech Communication and Technology*.

Burger, S., Weilhammer, K., Schiel, F., and Tillmann, H. G. (2000). Verbmobil data collection and annotation. In *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 537–549. Springer.

Campione, E., Véronis, J., and Deulofeu, J. (2005). C-oral-rom, integrated reference corpora for spoken romance languages, édité par e. cresti et m. moneglia, chapitre 3. the french corpus.

Carruthers, J. (2013). French oral narrative corpus. Commissioning Body / Publisher: Oxford Text Archive.

Chui, K., Lai, H.-l., et al. (2008). The nccu corpus of spoken chinese: Mandarin, hakka, and southern min.

Durand, J., Laks, B., and Lyche, C. (2009). Le projet pfc (phonologie du français contemporain): une source de données primaires structurées. *Phonologie, variation et accents du français*, pages 19–61.

Ginzburg, J., Fernández, R., and Schlangen, D. (2014a). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):1–64, June.

Ginzburg, J., Tian, Y., Amsili, P., Beyssade, C., Hemforth, B., Mathieu, Y., Saillard, C., Hough, J., Kousidis, S., and Schlangen, D. (2014b). The Disfluency, Exclamation and Laughter in Dialogue (DUEL) Project. In *Proceedings of the 18th SemDial Workshop (DialWatt)*, pages 176–178, Herriot Watt University, Edinburgh.

Gu, Y. (2000). Compiling a spoken chinese corpus of situated discourse. In *Keynote speech given at the 8th national conference on contemporary linguistics. Guangzhou*.

Hough, J. and Purver, M. (2013). Modelling expectation in the self-repair processing of annotat-, um, listeners. In *Proceedings of the 17th SemDial Workshop (DialDam)*, pages 92–101, Amsterdam, December.

Hough, J. and Purver, M. (2014). Strongly incremental repair detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 78–89, Doha, Qatar, October. Association for Computational Linguistics.

Hough, J. and Schlangen, D. (2015). Recurrent neural networks for incremental disfluency detection. In *Proceedings of Interspeech 2015*, pages 849–853.

Hough, J., de Ruiter, L., Betz, S., and Schlangen, D. (2015). Disfluency and laughter annotation in a light-weight dialogue mark-up protocol. In *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*.

Kennington, C., Kousidis, S., and Schlangen, D. (2014). Multimodal dialogue systems with inprotks and venice. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt). Posters*.

Kohler, K. J. (1996). Labelled data bank of spoken standard german: the kiel corpus of read/spontaneous speech. In *ICSLP 96*, volume 3, pages 1938–1941. IEEE.

Kousidis, S., Pfeiffer, T., and Schlangen, D. (2013). Mint. tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. *Interspeech 2013*.

Kousidis, S., Hough, J., and Schlangen, D. (2015). Exploring the body and head kinematics of laughter, filled pauses and breaths. In *Proceedings of The 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech*, pages 23–25.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., Tchobanov, A., et al. (2014). Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Language Resources and Evaluation Conference*.

Meteer, M. W., Taylor, A. A., MacIntyre, R., and Iyer, R. (1995). *Disfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania.

Peters, B. (2005). The database-the kiel corpus of spontaneous speech. *Prosodic Structures in German Spontaneous Speech, AIPUK 35a*, pages 1–6.

Schiel, F., Heinrich, C., and Barfüsser, S. (2012). Alcohol language corpus: the first public corpus of alcoholized german speech. *Language resources and evaluation*, 46(3):503–521.

Schmidt, T., Hedeland, H., Lehmberg, T., and Wörner, K. (2010). Hamatac–the hamburg maptask corpus.

Tian, Y., Beyssade, C., Mathieu, Y., and Ginzburg, J. (2015). Editing phrases. Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL), pages 149–156.

Truong, K. P. and Trouvain, J. (2012). Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis. *Proceedings of the 4th International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 20–24.

Xiao, R. and Tao, H. (2007). The lancaster los angeles spoken chinese corpus.

Zwarts, S., Johnson, M., and Dale, R. (2010). Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1371–1378, Stroudsburg, PA, USA. Association for Computational Linguistics.