



**HAL**  
open science

## Toward a scary comparative corpus: The Werewolf Spoken Corpus

Laurent Prevot, Yao Yao, Arnaud Gingold, Bernard Bel, Kam Yiu Joe Chan

► **To cite this version:**

Laurent Prevot, Yao Yao, Arnaud Gingold, Bernard Bel, Kam Yiu Joe Chan. Toward a scary comparative corpus: The Werewolf Spoken Corpus. THE 19TH WORKSHOP ON THE SEMANTICS AND PRAGMATICS OF DIALOGUE, 2015, Gothenburg, Sweden. pp.204, 2015. hal-01231889

**HAL Id: hal-01231889**

**<https://hal.science/hal-01231889>**

Submitted on 15 Nov 2018

**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.

# Toward a scary comparative corpus

## The Werewolf Spoken Corpus

Laurent Prévot<sup>1</sup>, Yao Yao<sup>2</sup>, Arnaud Gingold<sup>1</sup>, Bernard Bel<sup>1</sup>, Kam Yiu Joe Chan<sup>2</sup>

<sup>1</sup> Aix Marseille Université, Laboratoire Parole et Langage, France

<sup>2</sup> The Hong Kong Polytechnic University, Chinese and Bilingual Studies, Hong Kong



### Context

- dialogue and multilogue differ significantly in terms of dialogue structure [3]
- Not much multilogue corpora available
- No comparable multilogue corpus available, see however [4] for English
- Existing multilogues [1, 5, 2] have been created within cooperative communicative situation
- Not much quantitative / systematic study of cross-linguistic / intercultural variation of this kind of natural data
- More theoretically: how different sources of variation blends within actual performances

→ Record and analyse a comparable corpus of Werewolf social game in France and in Hong Kong.

### Elicitation Protocol

- Role: 2 Werewolf, 6 villagers, 1 game master
- Alternation of night / day phases
  - night: werewolves decide to make disappear a villager (silent)
  - day: everyone vote to kill someone (discussion phase)
- Game ends when either all villagers or werewolves have been killed

### Recording Setting



French side

- Usually difficult to record clean speech data in 'one-location' multilogues
- Here a circle of people (French), 2 rows of 4 people facing each other (Mandarin)

### Objectives

- Explore linguistic and cultural variation in non-cooperative multilogue situation (deceptive speech, persuasive speech, competitive speech)
- First step: gather data and check first variables (overlap, turn-taking, amount of speech) variations across speaker situation

### Data curation

The WEREWOLF pilot corpus and its annotations were stored and described so as to allow a wide dissemination

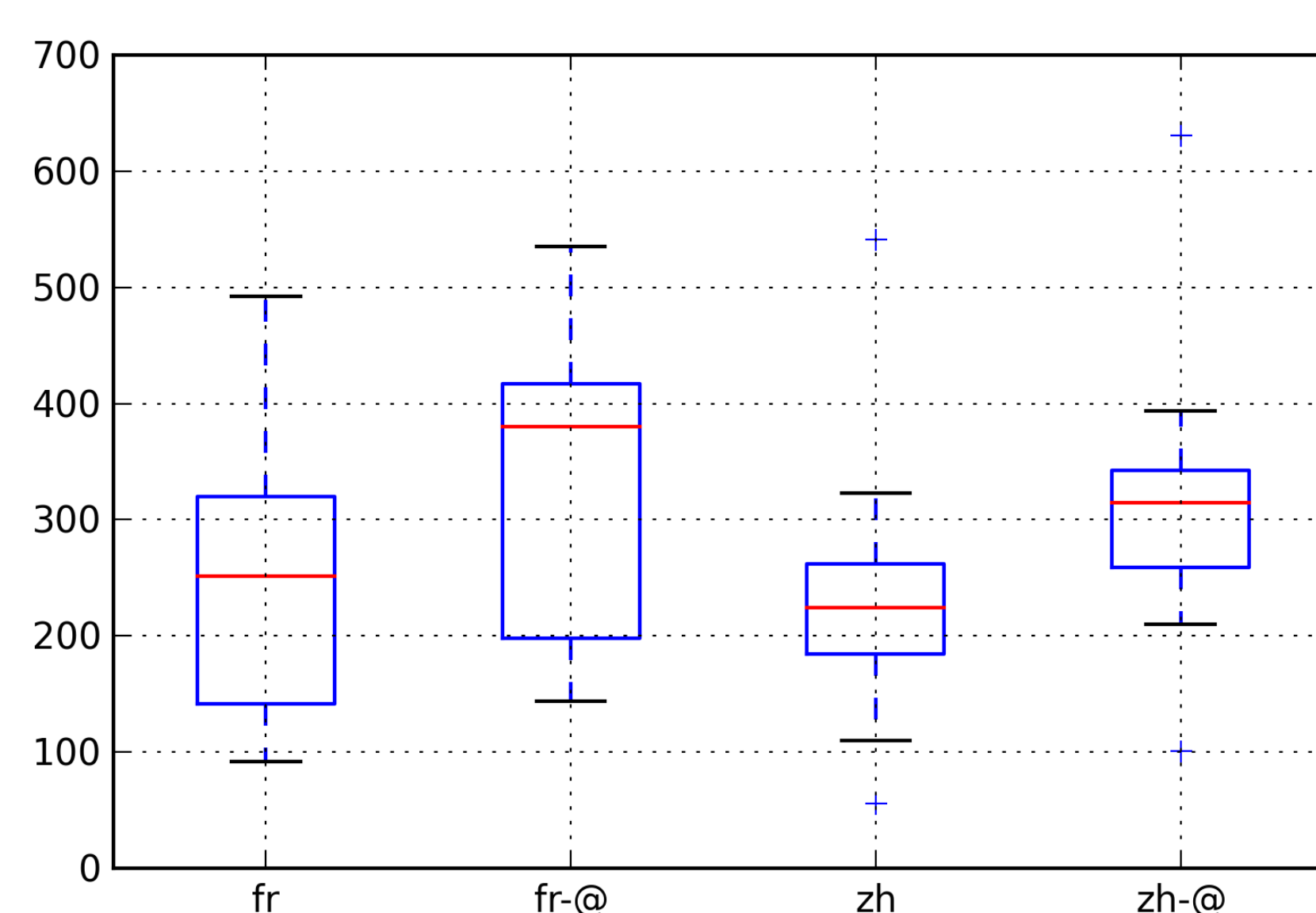
- <http://hdl.handle.net/11041/ortolang-000900>
- <http://hdl.handle.net/11041/ortolang-000908>

### References

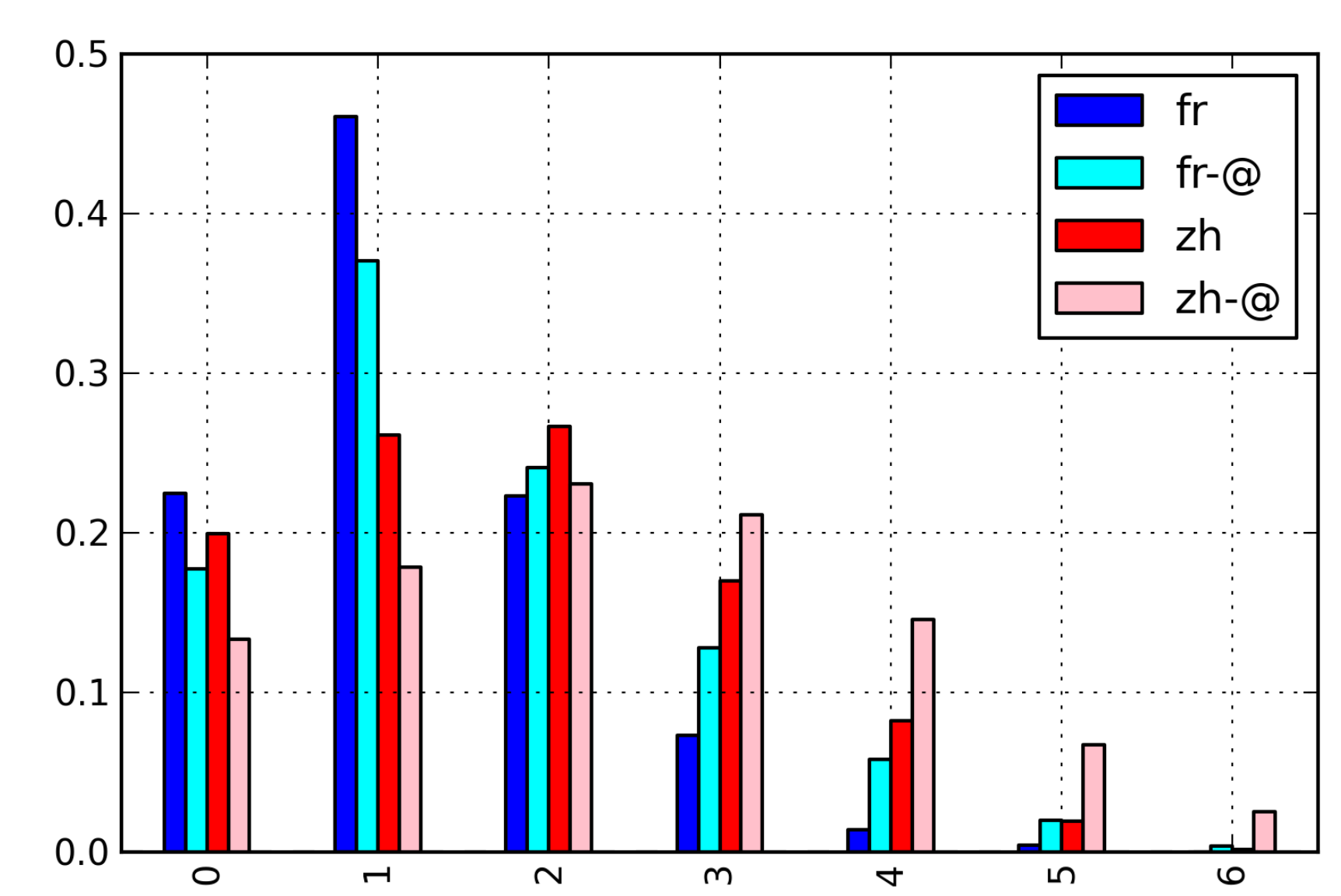
- [1] Susanne Burger, Victoria MacLaren, and Hua Yu. The isl meeting corpus: the impact of meeting type on speech style. In *INTERSPEECH*, 2002.
- [2] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *Machine learning for multimodal interaction*, pages 28–39. Springer, 2006.
- [3] Jonathan Ginzburg and Raquel Fernández. Action at a distance: the difference between dialogue and multilogue. In *Proceedings of DIALOR*, volume 5, 2005.
- [4] Hayley Hung and Gokul Chittaranjan. The IDIAP wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the international conference on Multimedia*, pages 879–882. ACM, 2010.
- [5] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Piskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I-364. IEEE, 2003.

### Basic info about a game

- So far 4 games for each language, 7-27 minute per game



Actual Speaking Duration (s)



# of simultaneous speakers (%), 1Hz

### Illustration

	toi a	#	@	#	@	#	@	#	@	#	@	#	et	#	@	#										
#	bah	to	aussi	hein	#	comment	elle	s'e	#	ah	si	c'est	le	genre	com	#	@	ah	merde	j	e	me	s-	je	me	
#			@	#	@				#																	
#	hm	@	#	@	#	@	c'étai	@	#																	
#					#																					
#					#																					
	uo	que	non	tu	es	végétar	#	a	@	#	n	#	non	non	mai	#	s'	j'	essaie	de	#					
#							@				@															

### Acknowledgements

This work has been carried out thanks to the support of the A\*MIDEX Variamu project (ANR-11-IDEX-0001-02), the Ortolang (ANR-11-EQPX-0032) French Government program and of the Faculty of Humanities of The Hong Kong Polytechnic University. We would like to thank all the students in Aix and Hong-Kong that participated in the recordings and in the transcription.