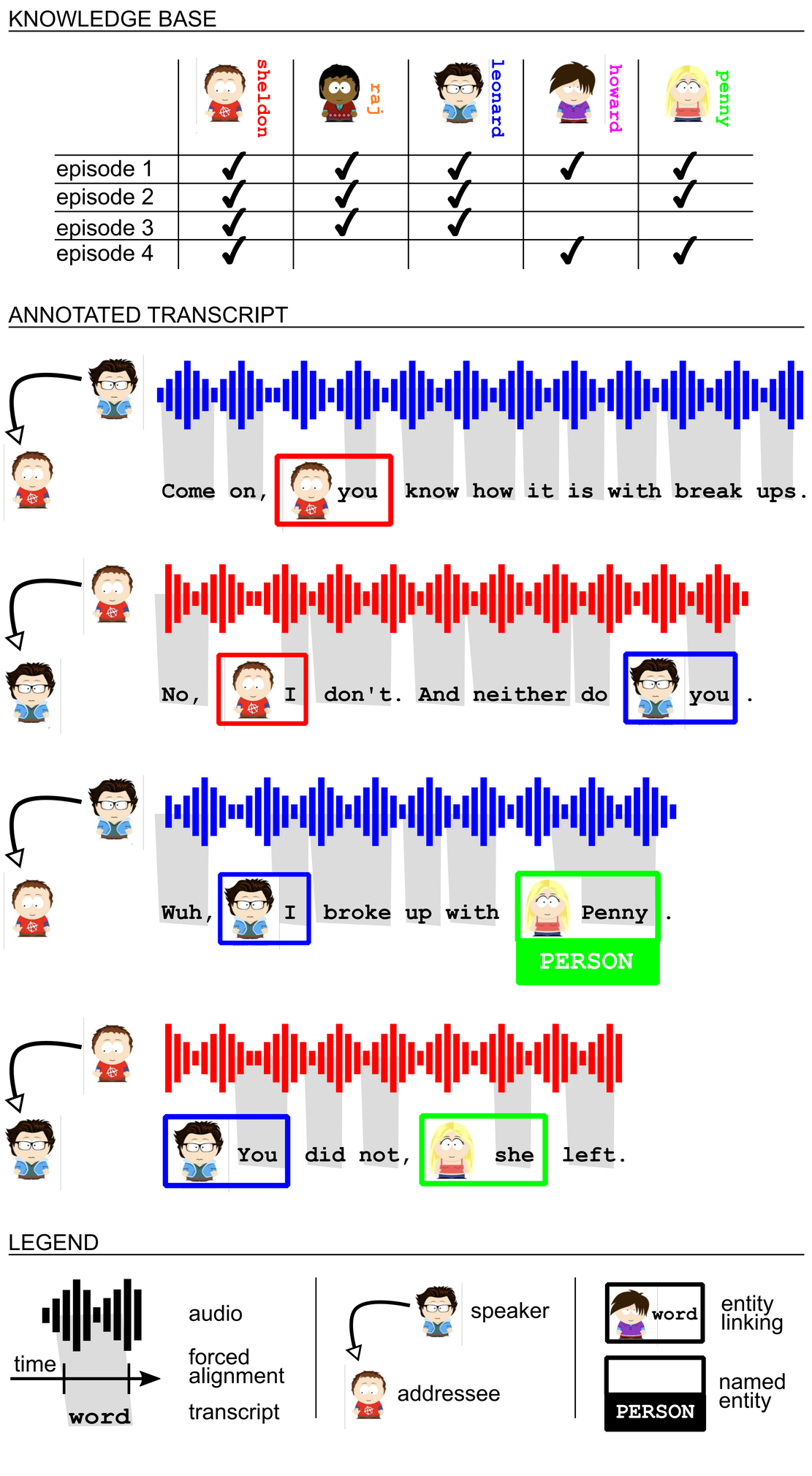
Bazinga! A Dataset for Multi-Party Dialogues Structuring

Paul Lerner[1], Juliette Bergoënd*, Camille Guinaudeau[1], Hervé Bredin[2], Benjamin Maurice*, Sharleyne Lefevre*, Martin Bouteiller*, Aman Berhe*, Léo Galmant*, Ruiqing Yin*, Claude Barras[3]

[1]Université Paris-Saclay, CNRS, LISN, [2]IRIT, Université de Toulouse, CNRS, [3]Vocapia Research, *Work done while at Université Paris-Saclay, CNRS, LISN
[1,3]91400, Orsay, France, [2]Toulouse, France

Overview of the annotations



Tasks

- Speaker Diarization: overall 57.9% diarization error rate (20.3% of false alarm, 17.8% missed detection, and 19.8% speaker confusion) using off-the-shelf `pyannote.audio` pipeline [3] trained on DIHARD [5].
- Speaker Identification: either supervised or unsupervised [2].
- Automatic Speech Recognition: expect drop in performance because of the multi-party dialogue setting [6].
- Punctuation Restoration: overall 43.3% precision and 20.1% recall using off-the-shelf model from [7] trained on Europarl corpus [4]. 59.2% precision and 50.9% recall when trained on *Bazinga!*.
- Person Entity Recognition: 78% precision and 63% recall with Flair [1], 82% precision and 66% recall with spaCy[a].
- Entity Linking: nontraditional setting, unsupervised entity linking could be explored.
- Addressee Detection: *to whom is a speaker talking?*
- Continual Learning: TV series are chronological, characters appear in (or disappear from) the storyline.

Gold-standard annotations

	Episodes	Tokens	Speakers	Speech
Battlestar Galactica	13	56k	119	3.0h
Breaking Bad	7	29k	58	1.6h
Buffy the Vampire Slayer	12	68k	99	3.4h
Friends	24	82k	110	4.0h
Game of Thrones	10	54k	126	3.0h
Lost	25	101k	131	4.9h
The Big Bang Theory	17	58k	41	3.1h
The Office	6	23k	25	1.1h
The Walking Dead	6	25k	38	1.3h
Star Wars	7	72k	281	3.9h
Total	127	569k	1,028	29.4h

Silver-standard annotations

	Episodes	Tokens	Speakers	Speech
24	168	886k	–	46.0h
Battlestar Galactica	58	262k	100+	14.3h
Breaking Bad	54	236k	20+	13.5h
Buffy the Vampire Slayer	131	666k	200+	35.1h
ER	305	2,043k	–	100.4h
Friends	209	702k	190+	38.2h
Game of Thrones	50	247k	200+	14.0h
Homeland	58	278k	–	14.7h
Lost	79	300k	200+	14.3h
Six Feet Under	50	314k	30+	17.2h
The Big Bang Theory	190	581k	70+	33.2h
The Office	182	700k	140+	38.4h
The Walking Dead	93	260k	50+	13.7h
Harry Potter	8	81k	50+	4.6h
The Lord of the Rings	3	28k	30+	1.7h
Total	1,638	7,584k	1,300+	399.4h

Freely available

- hf.co/bazinga

References

[1] A. Akbik, D. Blythe, and R. Vollgraf. Contextual String Embeddings For Sequence Labeling. In *Proceedings of the International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[2] H. Bredin, A. Laurent, A. Sarkar, V.-B. Le, S. Rosset, and C. Barras. Person Instance Graphs for Named Speaker Identification in TV Broadcast. In *Odyssey 2014, The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014.

[3] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill. pyannote.audio: neural building blocks for speaker diarization. In *Proc. ICASSP 2020*, 2020.

[4] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit Conference*, pages 79–86. International Association for Machine Translation, 2005.

[5] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman. The Third DIHARD Diarization Challenge. *arXiv preprint arXiv:2012.01477*, 2020.

[6] P. Szymański, P. Żelasko, M. Morzy, A. Szymczak, M. Ży\la Hoppe, J. Banaszczak, L. Augustyniak, J. Mizgajski, and Y. Carmiel. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.295.

[7] O. Tilk and T. Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, 2016.

[a] https://spaCy.io/