DiaBiz – an Annotated Corpus of Polish Call Center Dialogs

Piotr Pęzik*, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert and Michał Adamczyk

University of Łódź
piotr.pezik@uni.lodz.pl

Rationale

DiaBiz is a large, annotated, multimodal corpus of Polish telephone conversations conducted in varied business settings. It was developed to increase the accessibility of speech resources and boost the development of third-party dialog systems, conversational intelligence tools and speech recognition engines for Polish. The data set is enriched with annotation which can be used to develop other downstream applications including: tools for spoken transcript formatting, speaker diarization, conversational intent and named entity recognition, spoken dialog segmentation, labelling and classification.

Dialog Structure

Intents A subset of DiaBiz transcripts is annotated with customer and agent intents relevant to a given business domain. Three intents of this kind are illustrated in the example below: a greeting, an introduction and a help offer:

1. **Agent**: Good morning, Joanna Kwiatkowska, (uh) Everyday Bank, how may I help you?
   Customer: Good morning, madam. (uh) I have blocked my access to, what do you call it, to my online account.
   Agent: Uhm, I understand. I will check your access status in a moment. But first, I need to ask a couple of questions to verify your identity.

Dialog acts Simultaneous with simple intent labelling, a more ambitious dialog act annotation effort is currently underway at Wrocław University of Science and Technology, where a team of annotators is adapting the 24617-2 ISO standard [1] in order to annotate the DiaBiz data with communicative functions and discourse relations occurring within dialog acts.

The Corpus

Phone-call interactions recorded via the Genesys PureCloud platform were exported as 16-bit 8 kHz stereo WAV files with separate speaker/agent channels. Recordings were first transcribed and punctuated automatically and subsequently manually corrected using punctuation and truecasing guidelines customized for spoken Polish.

- 3 million words of transcribed speech
- 4,036 call center interactions
- 410 hours of recordings
- 120 call scripts with 251 variants
- 5 agents, 191 customers
- 9 different business domains

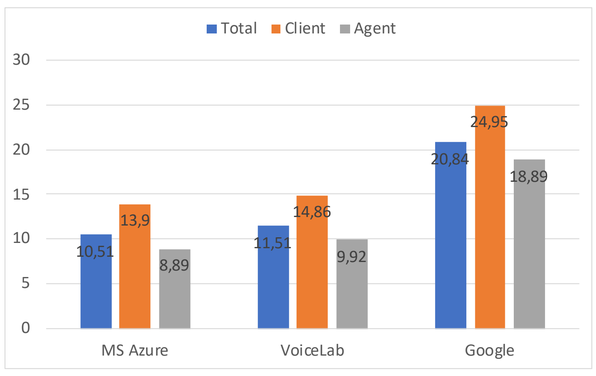
Coverage

Domain	# interactions	# words	Length	# scripts	# versions
Banking	907	773,858	92:56:54	26	57
Car rental	246	189,741	24:07:07	6	13
Debt collection	300	245,031	29:23:56	10	15
Energy services	390	248,295	30:05:42	15	24
Insurance	401	307,760	40:00:54	11	25
Medical care	371	236,057	30:13:57	14	19
Telecommunications	700	416,333	52:21:52	18	44
Tourism	451	674,066	86:23:10	10	31
Retail	270	133,702	24:24:00	10	23
Total	**4,036**	**3,224,843**	**409:57:32**	**120**	**251**

Use case

ASR evaluation

DiaBiz can has been used to evaluate out-of-the box accuracy of automatic speech recognition engines.[2]



Vendor	Total WER	Client WER	Agent WER
Microsoft Azure	10.51	13.9	8.89
Voicelab	11.51	14.86	9.92
Google	20.84	24.95	18.89

Availability

- **http://hdl.handle.net/11321/887**
- Distributed under a commercial license
- Funded by "CLARIN - Common Language Resources and Technology Infrastructure" (2014-2020 Smart Growth Operational Programme, POIR.04.02.00-00C002/19
- The data collection process was substantially supported by VoiceLab (**voicelab.ai**), Damovo (**damovo.com**) and Genesys (**genesys.com**)

References

[1] Harry Bunt. *Guidelines for using ISO standard 24617-2*. [s.n.], January 2019. TiCC TR 2019–1.

[2] Piotr Pęzik and Michał Adamczyk. An evaluation report with accompanying datasets benchmarking the performance of commercially available ASR services of Polish on the DiaBiz corpus. https://clarin-pl.eu/dspace/handle/11321/894, 2022.