DiaBiz

DiaBiz corpus is a dialog corpus comprising **recordings** and annotated **transcriptions** of **phone-based customer-agent interactions** in several key business domains.

A general overview of the corpus can be found in this paper:

- Pęzik, Piotr, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert, and Michał Adamczyk. 'DiaBiz – an Annotated Corpus of Polish Call Center Dialogs'. In Proceedings of the Language Resources and Evaluation Conference, 723–26. Marseille, France: European Language Resources Association, 2022. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.76.pdf

Also see the accompanying poster here:

- https://drive.google.com/file/d/1f1PNXa98TdjnzVqaml16VCp5Z3myxt0i/view?usp=sharing

The corpus comprises:

- 4,010 conversations amounting to nearly 410 hours and over 3.2 million words
- dialogues between 5 call-center agents and 191 participants as customers
- data from 9 business domains with high commercial demand for conversational analytics and automation solutions
- dialogues based on 251 real-life interaction scenarios

The domains covered:

Domain	Dialogs	Words	Duration (HH:MM:SS)
Banking	907	773,858	92:56:54
Car rental	246	189,741	24:07:07
Debt collection	300	245,031	29:23:56
Energy services	390	248,295	30:05:42
Insurance	401	307,760	40:00:54
Medical care	371	236,057	30:13:57
Telecommunications	700	416,333	52:21:52
Tourism	451	674,066	86:23:10
Retail	270	133,702	24:24:00
Total	**4,010**	**3,224,843**	**409:57:32**

The data was automatically automatically **transcribed** and **time-aligned** and subsequently manually **corrected** and **annotated**.

Applications

Customer support interactions recorded by operators of call centers are highly unlikely to be widely released in any useful form as they contain sensitive information which is subject to strict privacy regulations. NLP start-ups and academic research groups have to develop their own datasets or rely on limited resources which cannot be directly adapted to commercially viable domains. The **DiaBiz corpus** can serve as a **source of training and evaluation data** for a wide range of intrinsic and downstream tasks, such as:

- speech recognition and transcript formatting
- speaker diarization
- conversational intent and named entity recognition
- spoken dialog segmentation, labelling and classification
- conversational analytics as well as more sophisticated modelling of dialog systems.

The **DiaBiz corpus** is therefore a major new resource for spoken Polish, offering research potential and making it possible to bootstrap the development of language processing tools for automating linguistic interactions with high volumes of customers, such as voice bots and other dialog systems.

Availability

All the samples and supplementary materials available on this webpage are copyrighted. They are only included to illustrate the content of the DiaBiz database and should not be used for any other purposes without explicit permission from the University of Lodz representatives.

Click HERE to download sample recordings.

The current version of the recording catalog is available HERE.

For more information about the DiaBiz license for both commercial and scientific use, please contact piotr.pezik@uni.lodz.pl.

DiaBiz EN

A representative sample of the DiaBiz corpus has been localised into English:

Domain	Dialogs	Word count	DiaBiz	Percentage
Banking	127	69 184	773 858	9%
Telecommunications	117	64 805	416 333	16%
Tourism	71	58 626	674 066	9%
Insurance	57	31 009	307 760	10%
Energy services	55	29 740	248 295	12%
Retail	46	25 316	133 702	19%
Medical care	45	22 044	236 057	9%
Debt collection	34	17 776	245 031	7%
Car rental	31	17 199	189 741	9%

Domain	Dialogs	Word count	DiaBiz	Percentage
Total	583	335 699	3 224 843	10%

A sample od the localised corpus can be downloaded here

Acknowledgments

CLARIN-BIZ

DiaBiz was developed in the project titled “CLARIN - Common Language Resources and Technology Infrastructure”, which is financed under the 2014-2020 Smart Growth Operational Programme, POIR.04.02.00-00C002/19. We would also like to acknowledge the support of three companies: VoiceLab, Genesys and Damovo in the data collection and transcription efforts.

From:
http://docs.pelcra.pl/ - **Group**

Permanent link:
http://docs.pelcra.pl/doku.php?id=diabiz

Last update: **2025/05/18 10:57**