SpokesBiz

SpokesBiz is a corpus of conversational Polish developed within the CLARIN-BIZ project and currently comprising over 650 hours of recordings. The transcribed recordings have been diarized and manually annotated for punctuation and casing.

A sample of the corpus can be accessed here.

Instructions for downloading the entire corpus (over 100GB of transcriptions with recordings) can be found below.

To access an online index of the corpus see: https://spokes.clarin-pl.eu.

A general overview of the corpus can be found here:

- Piotr Pęzik, Sylwia Karasińska, Anna Cichosz, Łukasz Jałowiecki, Konrad Kaczyński, Małgorzata Krawentek, Karolina Walkusz, Paweł Wilk, Mariusz Kleć, Krzysztof Szklanny, Szymon Marszałkowski 2023 (forthcoming) SpokesBiz – an Open Corpus of Conversational Polish

Subsets

SpokesBiz is made up of several distinct subcorpora.

Subcorpus	Recordings	Words	Utterances	Hours	Speakers
CBIZ_BIO	170	1383646	68429	166	170
CBIZ_INT	10	26006	1575	2	11
CBIZ_LUZ	297	1510024	103571	157	116
CBIZ_POD	178	991464	47221	92	12
CBIZ_PRES	56	256922	18120	38	39
CBIZ_VC	84	655539	63188	71	110
CBIZ_VC2	84	760671	28397	89	86
CBIZ_WYW	46	327148	16778	37	46
Total	925	5911420	347279	652	590

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

The table below summarises the metadata fields used to describe each utterance in the corpus.

Column	Description
conversation_id	unique identifier for each conversation
recording_id	unique identifier for each recording
recording_path	path to download recording file corresponding with the recording_id
subcorpus	subcorpus name
conversation_style	type of communication
recording_year	year of recording
recording_place	city where the recording was created (empty when unknown)

Column	Description
recording_time_ss	recording time in seconds
segment_seq	segment order within the conversation
segment_id	unique segment identifier
segment_text	segment text after manual correction
segment_word_count	number of words within the segment (using SpaceTokenizer from NLTK)
segment_ts_start_ms	segment beginning timestamp in milliseconds
segment_ts_end_ms	segment ending timestamp in milliseconds
segment_words_ts_ms	timestamps for every word in segment
speaker_id	unique identifier for speaker
speaker_sex	speaker sex with levels f and m
speaker_education	speaker education with levels none, primary, secondary, vocational, higher
exact_speaker_age	exact speaker age (NULL when unknown)
speaker_age_range_from	minimum range value with levels 0, 20, 30, 40, 50, 60, 70, 80, 90
speaker_age_range_to	maximum range value with levels 19, 29, 39, 49, 59, 69, 79, 89, 99
speaker_region	Polish voivodeship (empty otherwise)
speaker_first_language	speaker first language or languages

Availability

Please fill out this form to get access to SpokesBiz: https://forms.office.com/e/cpn88mcFC6.

For more information contact piotr.pezik@uni.lodz.pl.

License

The current license of SpokesBiz is CC-BY-NC-ND. This means that:

1. Users must cite the above-mentioned publication announcing SpokesBiz.
2. The corpus must not be used for commercial purposes.
3. “If you remix, transform, or build upon the material, you may not distribute the modified material.” In other words, you can build and distribute tools or models based on the material, but you must not redistribute the corpus data itself or any parts of it. If you need a different license for the corpus, please contact us at pelcra@uni.lodz.pl

Project Team

- Piotr Pęzik
- Michał Adamczyk
- Małgorzata Krawentek
- Paweł Wilk
- Sylwia Karasińska
- Angelika Peljak-Łapińska
- Karolina Adamczyk
- Monika Garnys

- Karolina Walkusz
- Anna Cichosz
- Anna Kwiatkowska
- Mikołaj Deckert
- Paulina Rybińska
- Izabela Grabarczyk
- Maciej Grabski
- Karol Ługowski
- Michał Koźmiński
- Zuzanna Deckert
- Piotr Górniak
- Konrad Kaczyński
- Łukasz Jałowiecki

Acknowledgments

CLARIN-BIZ

SpokesBiz 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 VoiceLab in the data transcription and processing efforts.

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

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

Last update: **2025/05/15 09:14**