PELCRA Spoken Offline Corpora

PELCRA Spoken Offline Corpora of conversational Polish (a.k.a SpokesMix) were collected as part of the CLARIN-PL project.

Each corpus consists of speech recordings (in WAV format) and word-by-word transcriptions, which also include some non-speech events. The transcriptions (in EAF format) are complemented with words and phones annotations (_out.eaf files), and, if available, with video content (MP4 format) and PDF transcripts.

Metadata are provided in XML files listing information about the recordings (titles, topics, dates, and URLs), media available (audio, video, pdf), and annotation details (file, date, annotator, place, duration, with additional information about the speakers whenever such data was available). A Document Type Definition specifying the structure of the elements and attributes of an XML document is included in each of the corpora.

Most of these offline corpora are indexed in Spokes and Spokes2. A subset of them can be obtained by filling out this form. Once the form is submitted you will get a password necessary to download the corpora.

corpus	description	recordings	speakers	word count	voice activity time (hh:mm)	total duration (hh:mm)
PELCRA_EMO	A corpus of focused interviews (people reflecting upon their emotions).	40	80	252,000	26:53	28:12
PELCRA_LUZ	A corpus of open interviews.	21	42	213,000	20:14	19:58
PELCRA_EMI	A corpus of Polish emmigrants to Scotland.	22	44	96,000	09:36	18:07
PELCRA_PARL	Samples of spoken parliamentary data.	48	241	99,000	12:22	14:13
PELCRA_YT_1	Samples of Polish YouTubers' videos.	25	106	49,000	04:56	06:39
PELCRA_YT_2	Second part of Polish YouTubers' videos.	23	45	49,000	05:10	05:46
MMW_1	A corpus of Polish conversations recorded in Wrocław in the 1980s.	14	65	60,000	7:02	8:33
MMW_2	Second part of the conversations recorded in Wrocław in the 1980s.	14	38	70,000	7:31	7:50
MMK	A corpus of Polish conversations recorded in Kraków in the 1980s.	4	11	15,900	1:46	1:49
PELCRA_IDIO	A corpus of open interviews in Polish.	146	148	327,500	:	38:51
	TOTAL	357	820	1,220,400	100:47	149:58

The following paper should be cited to fulfill the CC attribution condition of the license for these resources:

- Pęzik, Piotr. “Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix,” 4297–4300. Miyazaki, Japan, 2018. http://www.lrec-conf.org/proceedings/lrec2018/pdf/888.pdf.