



# DiaBiz – an Annotated Corpus of Polish Call Center Dialogs

Piotr Pezik , 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

## Abstract

This paper introduces DiaBiz – a large, annotated, multimodal corpus of Polish telephone conversations conducted in varied business settings, comprising 4,036 call centre interactions from 9 different domains, i.e. banking, energy services, telecommunications, insurance, medical care, debt collection, tourism, car rental and retail. The corpus was developed to boost the development of third-party speech recognition engines, dialog systems and conversational intelligence tools for Polish. Its current size amounts to nearly 410 hours of recordings and over 3 million words of transcribed speech. We present the structure of the corpus, data collection and transcription procedures, challenges of punctuating and truecasing speech transcripts, dialog structure annotation and discuss some of the ecological validity considerations involved in the development of such resources.

**Keywords:** dialog corpora, speech databases, DiaBiz, Polish

## 1. Motivation

DiaBiz is a newly released corpus of Polish call center dialogs comprising phone-based interactions in several business domains which have recently seen commercial demand for conversational analytics and automation solutions. The corpus was developed to address the problem of extremely low and uneven accessibility of such resources and 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. Although in this section we describe the motivation for compiling DiaBiz from the perspective of Polish language resource developers, we believe that the steps required to increase the availability of commercially relevant dialog corpus data are generally similar for any language.

There are a number of factors restricting access to logs, transcripts and recordings of business communication with customers. Most obviously, customer support interactions recorded by operators of call centers contain sensitive information, which is subject to strict privacy regulations. Although in rare cases anonymized speech databases representing less sensitive domains may be distributed under a public license<sup>1</sup>, in general, interactions recorded by banks, insurance companies and other commercial service providers or government institutions are highly unlikely to be widely released in any useful form. Current EU and worldwide legislation requires that such recordings must not be distributed or even exploited without explicit consent and that they can only be stored for a limited period of time. On the one hand, external subcontractors (e.g. established NLP solution providers) of original data holders (such as banks and insurance corporations) can gain temporary access to customer data under a non-disclosure agree-

ment. Also, depending on privacy policies and internal practices, providers of cloud-based call center services may make use of their clients' data, which also places them in a privileged position. At the other end of this data accessibility continuum, 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. For example, the performance of an intent recognition algorithm trained and evaluated on an air traffic control corpus may decrease significantly when it is used to automate debt collection calls.

DiaBiz substantially improves the availability of spoken Polish resources. A manually transcribed, time-aligned and annotated corpus of 4,036 conversations covering 9 domains and over 250 interaction scenarios can serve as a source of training and evaluation data for a wide range of intrinsic and downstream tasks, such as (8kHz) speech recognition and transcript formatting (see the description of punctuation and truecasing below), 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.

## 2. The Corpus

### 2.1. Acquisition Procedure

The DiaBiz corpus comprises over 3 million words in nearly 410 hours of conversations between 5 agents with professional experience in call center settings and 191 participants assuming the role of customers<sup>2</sup>. The phone-call interactions, recorded via the Genesys PureCloud platform, were based on 120 distinct customer service call scripts from 9 business domains. Inspired

<sup>1</sup>See the Polish component of the LUNA corpus (Raymond et al., 2007).

<sup>2</sup>The participants were rewarded with online shopping vouchers for each recording session.

by samples of real-life data, individual scripts were created in different versions, making up to a total of 251, to showcase some of the divergent paths one conversation may take. For example, a script revolving around unlocking the customer’s online account, may be enacted in two ways. The agent may directly proceed to reset the password over the phone, or the customer may be informed that their ID must be updated in person, resulting in arranging an appointment at the branch. Each version of the script was then prepared in two variants separate for agents and participants. To maximise the authenticity and naturalness of the dialogs, the callers were given only a general context of their situation as a customer, whereas the agent’s variant provided a more detailed view into the company policy and its procedures in a given domain. Participants were encouraged to improvise, incorporate follow-up questions and even go off-topic as long as the main customer issue from the script was retained. Conversations were recorded in one-hour sessions per participant, which on average included fifteen unique inbound and/or outbound interactions from varied domains. Callers used online personal data generators for customer identification purposes so that no real data was used. Any interactions in which participants broke character or ended the call prematurely, e.g. due to technical issues, were removed from the corpus.

## 2.2. Transcription, Punctuation and Truecasing

Recordings from the Genesys PureCloud platform were exported as 16-bit 8 kHz stereo WAV files. Speakers were recorded in separate channels, which made it technically possible to maintain very little speech overlap (occasional overlaps occurred when a participant talked hands free).

The separated agent/client channels of the recordings were first transcribed automatically using the VoiceLab ASR engine<sup>3</sup> with an average word error rate (WER) of 14%.<sup>4</sup> The transcriptions were then manually corrected using a dedicated web application. Transcribers fixed occasional diarization errors which resulted from cross-channel interference.

In the process of transcription correction, we also decided to develop and apply detailed guidelines for spoken dialog punctuation and truecasing. As a result, DiaBiz can serve as a large training set for the development and evaluation of punctuation and truecasing models for spoken Polish. The DiaBiz transcripts were first punctuated automatically using a transformer-based sequential classification model trained on generally available transcripts of spoken Polish and subsequently manually corrected by a team of annotators. The punctuation guidelines had to be developed to reflect the characteris-

<sup>3</sup>See <https://voicelab.ai>.

<sup>4</sup>Benchmarks based on a test subset of DiaBiz showed that the average WER was 14% for VoiceLab’s, 18% for Microsoft’s and 28% for Google’s Polish ASRs.

tics of spoken language, as the conventional rules governing punctuation in Polish were drawn up mostly for written text. Consequently, some of the highly frequent phenomena observable in spoken language are virtually omitted in traditional Polish punctuation guides, and to tackle this problem several new punctuation rules had to be included in the guidelines. After several iterations of inter-annotator agreement tests, we decided to limit the number of punctuation marks to the six symbols shown in Table 1. Commas and full stops are most widely employed. The ellipsis is the third most common punctuation mark used in transcriptions, which is mostly due to the fact that it is used to indicate common instances of hesitation. Exclamation marks and hyphens are more scarcely distributed across the material, the former being used to indicate emotional utterances. Detailed punctuation annotation guidelines are available on the DiaBiz website specified in the Availability section below.

Symbol	Name	Count
.	Full stop	330,695
,	Comma	345,628
?	Question mark	42,786
!	Excl. mark	790
...	Ellipsis	83,975
-	Hyphen	3,337
<b>Total</b>		<b>807,211</b>

Table 1: Counts of punctuation marks added to DiaBiz transcriptions

The impact of punctuation on the legibility of ASR transcripts is illustrated in the example below, which shows a single turn of an agent before (1) and after (2) applying both automatic and manual punctuation and truecasing corrections. The English translation of this passage is given in (3).

- (1) RawPL: rozumiem dobrze no to w takim razie ja sobie te wszystkie informacje zapisałam i jutro ktoś od nas wiem czy to będę ja czy inny konsultant zadzwoni do pana i będziemy ustalać indywidualną strategię działania dla pana dobrze
- (2) FormattedPL: Rozumiem. Dobrze, no to w takim razie ja sobie te wszystkie informacje zapisałam, i jutro ktoś od nas, nie wiem, czy to będę ja czy inny konsultant, zadzwoni do pana i będziemy ustalać indywidualną strategię dla pana, dobrze?
- (3) EN: I see. Okay, in this case I’ve written down all the information, and our consultant, either myself or somebody else, will call you tomorrow and we’ll work out an individual solution for you, okay?

The benefits of such formatting go far beyond mere legibility of ASR transcripts. It has been claimed that a number of spoken language processing tasks such as

segmentation, dialog modelling or entity recognition and linking depend on transcript punctuation and formatting (Pappagari et al., 2021). This is partly due to the fact that so-called “universal” language models are pre-trained mostly on written language and they work more reliably on certain tasks with explicit segmentation and conventional character casing.

### 2.3. Current Contents

Tables 2, 3, 4 and 5 describe the current contents of DiaBiz for 9 business domain represented in the corpus. At the time of writing the transcripts are still undergoing a final review, hence the expected total number of words in Table 2 and punctuation marks in Table 1 may change slightly.

Table 2 shows the total number of interactions and the estimated number of words for each domain.

Domain	# interactions	# words
Banking	907	773,858
Car rental	246	189,741
Debt collection	300	245,031
Energy services	390	248,295
Insurance	401	307,760
Medical care	371	236,057
Telecommunications	700	416,333
Tourism	451	674,066
Retail	270	133,702
<b>Total</b>	<b>4,036</b>	<b>3,224,843</b>

Table 2: The number of interactions and the estimated number of words (excluding punctuation) per domain

The total amount of recorded speech shown in Table 3 ranges from almost 93 hours in the banking domain to 24 for car rental. The distribution of talk time between domains (corresponding to the total number of words presented in Table 2) shows that banking, tourism and telecommunications are the three largest domains in the corpus.

Domain	Length (HH:MM:SS)
Banking	92:56:54
Tourism	86:23:10
Telecommunications	52:21:52
Insurance	40:00:54
Medical care	30:13:57
Energy services	30:05:42
Debt collection	29:23:56
Car rental	24:07:07
Retail	24:24:00
<b>Total</b>	<b>409:57:32</b>

Table 3: The amount of recorded speech per domain

The number of scripts and their versions is given in Table 4. The majority of scripts has either 2 or 3 versions designed to represent different actions taken by

interlocutors in the course of the interaction, hence the number of versions per domain can be two or sometimes three times larger than the total number of scripts. In terms of the number of scripts, banking is the largest domain with 26 distinct scripts and 57 versions. The domains of telecommunications and tourism are also represented. There are 18 scripts and accompanying 44 versions for telecommunications. While there are only 10 scripts for tourism, each of them has at least 3 versions amounting to 31 script versions in this domain. The domains of car rental and retail have the least amount of scripts and their versions.

Domain	# scripts	# versions
Banking	26	57
Car rental	6	13
Debt collection	10	15
Energy services	15	24
Insurance	11	25
Medical care	14	19
Telecommunications	18	44
Tourism	10	31
Retail	10	23
<b>Total</b>	<b>120</b>	<b>251</b>

Table 4: The number of distinct scripts and versions per domain

As shown in Table 5, the majority of participants were females and university graduates. The average age of participants was 31.63 with the median of 30 and almost all of them spoke Polish as their mother tongue.

Total number of speakers		191
Gender	female	124
	male	67
Education	secondary	36
	student	26
	higher	129
Age	<= 19	7
	20-29	83
	30-39	69
	40-49	22
	50-59	6
	>= 60	4

Table 5: Participants contributing to the corpus by gender, education and age

### 3. Ecological Validity

In the context of psychological and linguistic research methodologies the notion of *ecological validity* is defined as “the degree of correspondence between the research conditions and the phenomenon being studied as it occurs naturally or outside of the research setting”

(Frey, 2018). This criterion is also relevant to the assessment of spoken language corpora, which can range from highly controlled transcripts of read speech or narration tasks through scripted interviews to recordings of unscripted in vivo conversations. While the availability of corpora of naturally occurring conversational Polish for both practical applications and theoretical linguistic research has been increased by the release of resources such as Spokes (Pezik, 2018), we consider the ecological validity of DiaBiz as an inevitable trade-off between naturalness and accessibility. The dialogs recorded in the corpus were loosely scripted but nevertheless acted out, which means that both agents and customers had to mentally “transport” themselves into the context of the conversations they were having. The implications of this data collection procedure for the prosodic, pragmatic or lexical realizations of the communicative tasks are sometimes clear. For example, while the speech of agents was highly formulaic and automated, there is occasionally some unnatural hesitation on the part of customers trying to remember their script names or other personal details. However, at the same time, speakers were invariably given maximum freedom. This pertains to the wording of messages but also to the content of conversations since on multiple occasions participants needed to use their knowledge and experience to elaborate on some more general information provided in the instruction. Also, the conversations were recorded over mobile phone connections, which means they contain close approximations of extra-linguistic events such as background noise, interference or inaudible speech.

## 4. Dialog Structure Annotation

### 4.1. Simple Intents

A subset of DiaBiz transcriptions is currently being enriched with dialog structure annotations using a simple business-oriented definition of *intents* as any utterances on the part of both the agent or customer that carry an illocutionary force and thus elicit the interlocutor’s reaction which is arbitrarily relevant to a given business domain. An intent conveys a distinct semantic and/or pragmatic value, hence one sentence may contain multiple intents, as shown in example (4) below, which is an English translation of an original DiaBiz dialog. The first turn of the agent can be interpreted to contain three separate intents: a greeting, an introduction, and a help offer. In general such intents can spread over multiple sentence units.

- (4) **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 verify your identity.

In addition to intent boundaries we also annotate intent parameters, which are also known in dialog systems as “slots”, such as phone numbers, e-mail addresses, customer numbers, dates, amounts, etc. Depending on the intent, such variables can be null, mandatory or optional.

### 4.2. 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 (Bunt, 2019) in order to annotate the DiaBiz data with communicative functions and discourse relations occurring within dialog acts.

## 5. Availability

In accordance with the regulations of its funding scheme, DiaBiz is distributed for research and commercial purposes alike under a commercial license. Up-to-date information about the corpus availability is published at <https://clarin-pl.eu/dspace/handle/11321/887>.

## 6. Acknowledgements

The DiaBiz corpus 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 (<http://voicelab.ai>), Genesys (<http://genesys.com>) and Damovo (<http://damovo.com>) in the data collection and transcription efforts.

## 7. References

- Bunt, H. (2019). *Guidelines for using ISO standard 24617-2*. [s.n.], January. TiCC TR 2019–1.
- Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks, California 91320.
- Pappagari, R., Zelasko, P., Mikolajczyk, A., Pezik, P., and Dehak, N. (2021). Joint prediction of true casing and punctuation for conversational speech in low-resource scenarios. *CoRR*, abs/2109.06103.
- Pezik, P. (2018). Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).
- Raymond, C., Riccardi, G., Rodriguez, K. J., and Wisniewska, J. (2007). The LUNA corpus: an annotation scheme for a multi-domain multi-lingual dialogue corpus. *Proceedings of Decalog 2007*.