Proceedings of SIGUL2022 @LREC2022, pages 41–47
Marseille, 24-25 June 2022

Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish

Mathilde Hutin, Marc Allassonnière-Tang
Université Paris-Saclay / LISN-CNRS (UMR 9015),
Muséum national d'Histoire naturelle / EA (UMR 7206)
Rue du Belvedère bât 507, 91405 Orsay, France; 17, place du Trocadéro, 75016 Paris, France
mathilde.hutin@lisn.upsaclay.fr, marc.allassonniere-tang@mnhn.fr

Abstract

Oral corpora for linguistic inquiry are frequently built based on the content of news, radio, and/or TV shows, sometimes also of laboratory recordings. Most of these existing corpora are restricted to languages with a large amount of data available. Furthermore, such corpora are not always accessible under a free open-access license. We propose a crowd-sourced alternative to this gap. Lingua Libre is the participatory linguistic media library hosted by Wikimedia France. It includes recordings from more than 140 languages. These recordings have been provided by more than 750 speakers worldwide, who voluntarily recorded word entries of their native language and made them available under a Creative Commons license. In the present study, we take Polish, a less-resourced language in terms of phonetic data, as an example, and compare our phonetic observations built on the data from Lingua Libre with the phonetic observations found by previous linguistic studies. We observe that the data from Lingua Libre partially matches the phonetic inventory of Polish as described in previous studies, but that the acoustic values are less precise, thus showing both the potential and the limitations of Lingua Libre to be used for phonetic research.

Keywords: Crowd-sourcing, open-access, language description, Polish

1. The "Resource Problem"

Languages are said to be "less-resourced" when the amount of data available and language-specific technologies are less developed for them than for other well-resourced languages such as English, Spanish, French or Chinese. At the root of the problem lies the question of the quantity of data available: This data is necessary in massive amounts to train and then test language technologies. Phoneticians and phonologists, i.e., researchers interested in speech, have to overcome an additional challenge: They cannot use written data as a proxy for language production and need audio recordings when working on vocal languages or video recordings when working on sign languages.

To overcome this challenge, researchers developed two strategies. The first one consists in collecting their own large corpora, either field-recorded, such as the PFC project for French (Durand et al., 2002), or recorded in laboratories such as the TIMIT database for English (Garofolo et al., 1993) or NCCFr for French (Torreira et al., 2010). The second strategy consists in gathering audio recordings from other sources such as TV or radio shows, as was done for instance in the framework of the international project OSEO Quaero (`www.quaero.org`), or from audio books, as exemplified by the LibriSpeech corpus for English (Panayotov et al., 2015, `www.openslr.org/12`). Both options have the disadvantage of being overly costly, both in money and human resources, and sometimes not freely accessible to the community. A third path has been recently explored: crowd-sourced data, recorded by volunteers and therefore much less costly in time and money and generally open-source. The project Common Voice (Ardila et al., 2020, `http://commonvoice.mozilla.org`) for instance was launched in 2017 by Mozilla for the intended purpose of creating a free database for the development of speech recognition software. In March 2022, it contains ~18,000h of speech, 14,000 of which have been validated by other speakers, in 87 languages.

In the present paper, we explore a similar project: Lingua Libre, a participatory linguistic media library developed by Wikimedia France (`www.lingualibre.org`). It was launched in 2015, and, in March 2022, it counts ~700,000 recordings in 148 languages across 777 speakers. This database is interesting to explore because it differs from Common Voice in the fact that its aim is not primarily the development of new technologies, or even linguistic inquiry in general, but patrimonial conservation of languages. Lingua Libre was used only once for academic purposes, i.e., to estimate the transparency of graphic systems in 17 languages with an artificial neural network (Marjou, 2021). With this study, we aim to show that such data is also easily processable and useful for language description. In this proof of concept, we use Lingua Libre to describe the phonetics-phonology interface in Polish, a language we claim can be considered as less-resourced.

In the following, we present an overview of Polish corpora available today to show how Polish can be considered a less-resourced language (Section 2) and describe the Polish phonology and why describing associated phonetic characteristics is essential to both computer scientists and linguists (Section 3). In Section 4, we present our corpus and methodology. In Section 5, we provide counts of the consonants and vowels in our Polish data (5.1) as well as acoustic values of vowels (5.2. Finally, in Section 6, we conclude and discuss the results.

2. Oral Corpora for Polish

In this Section, we provide an overview of oral resources available for Polish and advocate for the need to explore new, open-source, less expensive alternatives. Even today, oral corpora for Polish are indeed problematic: Their scarcity, technical characteristics or expensiveness allow us to define Polish as a less-resourced language.

First, most oral corpora for Polish were designed to train language models, and are thus often expensive to produce and to use. One of the oldest databases for this language, the BABEL Polish Database (ELRA-S0307) [1] is a speech database produced under the COPERNICUS program whose objective was to create a database of languages of Central and Eastern Europe. The Polish part consists in ∼16h of read speech (30 males, 30 females) from the 1990s and its license is expensive. Polish is also part of the GlobalPhone corpus (Schultz, 2002), also designed to provide read speech data for the development and evaluation of large continuous speech recognition systems in 22 languages. The Polish part of GlobalPhone was collected from 48 female and 54 male native speakers in Poland aged 18 to 65. Each speaker read ∼100 utterances from newspaper articles, resulting in 10130 utterances of journalistic speech (and their transcriptions). The Polish Speecon database (ELRA-S0179) [2] comprises both adult (286 males, 264 females) and child (25 boys, 25 girls) speech, providing 248h of speech recorded in various environments, but is again extremely costly. Most recently, in 2019, the Polish Speech Database (Szwelnik et al., 2019) was developed by VoiceLab. It consists of ∼280h of speech (and corresponding transcripts), i.e., 263,424 utterances of Polish speech data from 200 speakers (103 male and 97 female ranging 15 to 60), recorded in Poland. Speakers were asked to record themselves reading a text on a website for at least 60 minutes from their home computer using a headset. The text comprised sentences covering most speech sounds in Polish. The corpus is thus rather representative of read Polish, but its usage is free only to LDC members.

Some of these expensive corpora are not even representative of the actual Polish-speaking community, with only one or few speakers. For instance, the Bonn Open Synthesis System (BOSS) synthesizer (Demenko et al., 2009) has a unit selection corpus for Polish of only 115 minutes of speech read by one professional radio speaker. Similarly, Polish entered the Collins Multilingual database (ELRA-S0383) [3] , covering Real Life Daily vocabulary in a variety of topics in 32 languages (the WordBank, see ELRA-T0376) and a multilingual set of sentences in 28 languages (the PhraseBank, see ELRA-T0377). The audio was recorded by only one native speaker of each language, resulting in 2,000 audio files for each language, and the corpus' license is also very expensive and limited to non-commercial use.

Less representative also are corpora dedicated to specific language domains, such as the ONOMASTICA project (ELRA-S0043)[4], a European project aiming to produce a multi-language pronunciation lexicon of proper names in 11 languages, or the JURISDIC project (Demenko et al., 2008), which aims to create a database to help develop technologies for the dictation of legal texts and includes ∼1200h of both semi-spontaneous and read domain-specific speech from ∼1000 judges, lawyers, police officers or university staff.

Other corpora can be problematic from a technical point of view. For instance, Polish is represented in the CSLU corpus of telephone speech (Lander, 2005), which contains ∼84h of fixed vocabulary and fluent continuous telephone speech (and orthographic transcriptions for a subset of the utterances). Polish is also part of the Multi-Language Conversational Telephone Speech 2011 - Slavic Group (Jones et al., 2016), comprising ∼60h of telephone speech in Polish, Russian and Ukrainian. Portions of these telephone calls were also used in the NIST 2011 Language Recognition Evaluation (LRE) (Greenberg et al., 2018), containing 204h of conversational telephone speech and broadcast audio in 24 languages. Yet telephone speech can be challenging to process, since it is usually recorded on reduced bandwidth (4 kHz), which is enough for some usages but may induce an inadequacy with models trained on larger bandwidth (8 kHz).

Finally, the most easily usable oral corpus for Polish is the National Corpus of Polish (NKJP) (Przepiórkowski et al., 2012, `www.nkjp.pl/`). It is mainly a corpus of written Polish, comprising over 1.5 billion words from classical literature, daily newspapers, specialist periodicals and journals, a variety of Internet texts, and transcripts of conversations by both male and female speakers, in various age groups, coming from various regions of Poland. However, the NKJP also comprises a sample of spoken, conversational Polish of ∼2 million tokens.

As can be seen from this overview of Polish oral corpora, most were created with the intended purpose

[1] `http://catalog.elra.info/en-us/repository/browse/ELRA-S0307/`

[2] `http://catalog.elra.info/en-us/repository/browse/ELRA-S0179/`

[3] `http://catalog.elra.info/en-us/repository/browse/ELRA-S0383/`

[4] `http://catalog.elra.info/en-us/repository/browse/ELRA-S0043/`

of developing tools or training language technologies, sometimes for specific sociolects. Several are not representative of a large portion of the population or suffer from technical defects, and most of them are expensive to use. In the present paper, we are interested in how everyday vocabulary gathered for free for other purposes than software development can be used to investigate linguistic questions.

3. Polish Phonology

Polish (ISO 639-3) is a Slavic language currently spoken by 36.5 million speakers, mainly in Poland, Europe (`www.ethnologue.com`). In terms of number of speakers, Polish is the largest language in the West Slavic group and the second largest of all Slavic languages after Russian (Lewis et al., 2013). It is a highly inflected language, with a much richer inflection of nouns, adjectives, verbs, pronouns, and numerals than most Germanic languages.

Describing the phonetic characteristics of Polish is important, from a linguistic point of view, for the understanding of its sound system, its variability and its possible evolution. From an applicable perspective, understanding these linguistic characteristics is helpful for Automatic Speech Recognition (ASR) systems, especially for such an inflected language (Demenko et al., 2012).

In terms of phonetic inventory, grammars describe Polish as displaying 31 consonants and 6 vowels (Jassem, 2003).

Consonants are displayed in Table 1. They are divided across 6 modes of articulation: stops, fricatives, and affricates, that have a two-fold distinction between voiceless and voiced, as well as nasals, one lateral, one flap and two approximants, and across 5 places of articulation: labial(-dental), dental, alveolar, (alveo-)palatal, and velar.

	Lab	**L-d**	**(P-)d**	**Al**	**Al-p**	**P**	**V**
Plos	p b		t d			c ɟ	k g
Nas	m		n		ɲ		ŋ
Fri		f v	s z	ʃ ʒ	ɕ ʑ		x
Aff			ts dz	ʧʤ	tɕ dʑ		
Lat			l				
F/t				r			
App						j	w

Table 1: The consonants of Polish. The abbreviations are read as follows. Lab = Labial, L-d = Labiodental, (P-)d = (Post-)dental, Al = Alveolar, Al-p = Alveo-palatal, P = Palatal, V = Velar, Plos = Plosive, Nas = Nasal, Fri = Fricative, Aff = Affricates, Lat = Lateral, F/t = Flap/trill, App = Approximant.

Vowels on the other hand, are displayed in Table 2. They are distributed across three aperture levels, i.e., high, mid and low vowels, and across three antero-posteriority positions (front, central and back vowels). The vowels /i/ and /ɨ/ are debatably positionally-conditioned allophones, at least in non-initial position (Jassem, 1958). Therefore, in the current paper, we only consider the [i] sounds and we do not include /ɨ/.

	Front	Central	Back
High	i	ɨ	u
Mid	e		o
Low		a	

Table 2: The vowels of Polish.

4. Materials and Method

In this paper, we use the data from Wikimedia's participatory linguistic library: Lingua Libre. As a crowdsourcing tool, any speaker can log in, fill in a profile with basic metadata for themselves or for other speakers, and record themselves or their guests reading lists of words in their native language. The device detects pauses, which allows for the recording to end when the word has been read and the next recording to start automatically after, therefore effortlessly generating relatively short audio files for each word. Each audio file is supposed to be titled on the same template of 'Language - Speaker name - Item name'. For example, for the recording 'pol.-KaMan-dokumentalny.wav', the language of the recording is Polish ('pol'), the speaker ID is 'KaMan', and the recorded item is 'dokumentalny', which means 'documentary'. The speaker then checks the validity of their aufio files and uploads them in Creative Commons, meaning that all files are open-source.

We chose to investigate Polish because it is the second most represented language in Lingua Libre, with 81,071 recordings across 15 speakers. The most represented language in Lingua Libre is French, with thrice as much recordings (241,825) across 283 speakers, but since this language can be considered as well-resourced and well-documented, it was less interesting to test our methodology.

The workflow for data extraction is as follows. First, the recordings are scrapped from the Lingua Libre database. In the present study, we extract all the +80,000 recordings available in Lingua Libre. Second, the recordings are segmented and aligned using WebMAUS (Kisler et al., 2017), the online open-access version of the MAUS software (Schiel, 2004), which is used to automatically time-align a recording based on its orthographic transcription. MAUS creates a pronunciation hypothesis graph based on the orthographic transcript of the recording (extracted from the name of the audio file) using a grapheme-to-phoneme converter. During this process, the orthographic transcription is converted to the Speech Assessment Methods Phonetic Alphabet (SAMPA). The signal is then aligned with the hypothesis graph and the alignment with the highest probability is chosen. As an overview of its accuracy, experiments have shown that the MAUS alignments match human alignments 95% of the time (Kipp et al., 1997). At this point, the extracted data allow us to have a frequency count of each phoneme that is found within the data. Third, the recordings of the selected vowels are extracted and analyzed in terms of formants. For each recording of each vowel, the mean F1 and F2 of the entire sound are extracted. The mean formants are considered to attenuate the influence of context-induced noise in the recordings. During this process of data extraction and analysis, the following R packages are used: `emuR` (Winkelmann et al., 2021), `PraatR` (Albin, 2014), and `tidyverse` (Wickham, 2017).

5. Results

Investigating the frequency of phonemes, and of sequences of two or three phonemes (especially across word boundaries compared to word-internally), has been proposed in past research mainly to improve speech recognition system with statistical language modelling (Jassem, 1973; Basztura, 1992; Ziółko et al., 2009; Ziółko and Gałka, 2010; Kłosowski, 2017). However, such explorations are also useful to theorists investigating language variation in synchrony and language evolution through the lens of frequency-based exemplar models (Bybee, 2002).

For this preliminary proof-of-concept, we propose to first investigate the frequency of single phonemes. We will compare the ratio of each phoneme found in the data from Lingua Libre with the ratio of phonemes found in previous studies using controlled linguistic materials. Second, we will focus on Polish vowels and compare the formant values found in previous studies with the formant values of the vowels found in Lingua Libre. We focus on F1 and F2 since it has been shown that the most important acoustic property of vowels are positions and shapes of the first two formants (Izydorczy and Kłosowski, 1999).

5.1. Phoneme Frequency

With regard to the frequency of phonemes, Table 3 displays the results from 5 previous studies, all using written text (converted grapheme-to-phoneme) as data [5], as well as the ratio found from the Lingua Libre data.

	1973	**1992**	**2009**	**2010**	**2017**	LiLi
e	10.2	10.6	9.1	7.8	9.5	8.0
a	9.3	9.7	9.5	8.1	9.5	11.1
o	9.1	8.0	8.9	7.6	9.2	8.7
t	4.4	4.8	4.4	3.7	4.6	3.9
n	4.0	4.0	4.4	3.6	4.3	4.7
ɨ	4.1	3.8	3.6	3.1	4.1	3.1
j	4.5	4.4	3.7	3.2	4.0	3.3
i	3.9	3.4	4.3	3.6	4.0	4.3
r	3.6	3.2	4.6	3.7	3.7	2.5
s	3.0	2.8	3.6	2.9	3.7	3.4
v	3.5	2.9	3.7	3.1	3.4	4.0
p	3.1	3.0	3.2	2.7	3.4	2.8
u	3.4	2.8	3.3	2.7	3.3	2.7
m	3.5	3.2	2.9	2.6	3.1	2.6
k	2.7	2.5	2.9	2.4	2.9	4.8
ŋ	2.6	2.4	2.0	1.8	2.5	3.3
d	2.2	2.1	2.8	2.3	2.2	2.0
l	2.1	1.9	2.6	2.1	2.2	3.1
w	2.2	1.8	1.6	1.6	1.9	1.6
ʃ	2.0	1.9	1.2	1.1	1.7	1.6
f	1.5	1.3	1.6	1.3	1.6	1.4
z	1.8	1.5	1.9	1.6	1.6	1.7
ts	1.5	1.2	1.6	1.3	1.4	1.4
b	1.5	1.5	1.4	1.3	1.4	1.7
g	1.5	1.3	1.5	1.3	1.3	1.2
ɕ	1.5	1.6	0.9	0.9	1.3	1.3
tɕ	1.3	1.2	0.6	0.6	1.1	2.4
x	1.1	1.0	1.4	1.1	1.1	0.9
ʧ	1.2	1.2	0.9	0.8	1.1	1.4
ʒ	1.2	1.3	0.9	0.8	1.1	1.0
ɛ̃	0.7	0.6	0.6	0.5	0.7	0.1
c	n.a.	0.7	0.6	0.5	0.6	n.a.
dʑ	0.8	0.7	0.5	0.5	0.5	0.1
dz	0.2	0.2	0.2	0.2	0.2	0.1
ʑ	0.2	0.2	0.1	0.1	0.1	n.a.
ɟ	n.a.	0.1	0.2	0.1	0.1	n.a.
ʤ	0.0	0.1	0.1	0.1	0.1	0.1

Table 3: Rates (%) of each phoneme in 5 past corpora (Jassem, 1973; Basztura, 1992; Ziółko et al., 2009; Ziółko and Gałka, 2010; Kłosowski, 2017) and in Lingua Libre (abbreviated as LiLi). The frequencies from Lingua Libre are extracted based on all the recorded words available in Lingua Libre. The cells with 'n.a.' indicate that a phoneme was not found in the sample.

We can see in Table 3 that the ratio found in Lingua Libre generally matches the ratio found in previous studies. Taking vowels as an example, /a/, /e/, and /o/ are nearly twice more frequent than the vowels /i/ and /u/. In terms of consonants, we also see that the consonants that have a low ratio in previous studies (e.g., dʑ, dz, ʑ, and ʤ) are also rare in the Lingua Libre data. As another example, the voiceless stops /p, k/ are regularly more frequent than their nasal counterparts /m, ŋ/, and both voiceless stops and nasals /t, n/, /p, m/ and /k, ŋ/ are more frequent than their voiced oral counterparts /d/, /b/ and /g/. Finally, alveolar obstruants are generally more frequent than labials and labials than velars, and voiceless obstruants than their voiced counterparts.

[5] Other studies, such as (Ziółko et al., 2014), have explored the frequency of diphones and triphones in oral corpora, but not that of single phonemes.

In Lingua Libre, however, compared to the lowest rate in the past five research papers, there are less /r/ (δ=0.7%) and, to a lesser extent, less /ɛ̃/ (δ= 0.4%), /dʑ/ (δ=0.4%), /d/ (δ=0.1%), /g/ (δ=0.1%) and /dz/ (δ=0.1%). On the other hand, compared to the highest rate from the past five analyses, there are much more /a/ (δ=1.4%) and /k/ (δ=1.9%), more /v/ (δ=0.3%), /ŋ/ (δ=0.7%) and /l/ (δ=0.5%), and, to a lesser extent, more /ʧ/ (δ=0.2%). This may be due to the fact that we investigate isolated words, i.e., mostly lexical words, whereas previous studies analyzed (written) connected speech, i.e., mixing lexical and functional words. It may also be due to the fact that contemporary vocabulary has evolved to some extent.

5.2. Vowel Qualities

In this subsection, we analyze the first and second formants of vowels. As a reference point, consider the values from 10 speakers analyzed with a Sona-Gram (Jassem, 1968) reproduced in Table 4, and the values from 10 other speakers analyzed spectrographically (Krzyśko et al., 1999) in Table 5.

	F1 (S-G)	F2 (S-G)	F1 (LiLi)	F2 (LiLi)
a	630-900	1100-1600	500-990	1300-2500
e	520-630	1600-2200	320-830	1670-2520
i	190-270	2100-2200	210-410	2220-2670
o	490-680	790-1100	420-810	1050-2650
u	240-340	560-780	300-650	950-2670

Table 4: Ranges of F1 and F2 values (in Hertz) for the 5 cardinal vowels of Polish according to the Sona-Gram analysis (S-G) of 8 male and 2 female speakers (Jassem, 1968) on the left and to our own analysis of Lingua Libre (LiLi) on the right.

As one can see from Table 4 , the values for F1 and F2 in Lingua Libre are much less precise, expanding on a larger range than in Jassem (1968)'s data. This effect is especially obvious for the F2 values of /a/, /o/ and /u/, which display, between their lowest and their highest values, a 1200 Hz delta for /a/, a 1600 Hz delta for /o/ and a 1720 Hz delta for /u/ in Lingua Libre, *vs* a 500 Hz delta for /a/, a 310 Hz delta for /o/ and a 220 Hz delta for /u/. The values for /e/ are more precise, as they span across 110 Hz for F1 and 600 Hz for F2 according to Jassem (1968) and across 510 Hz for F1 and 850 Hz for F2 according to our Lingua Libre data. The acoustic analysis is the most precise for /i/, which spans across 80 Hz for F1 and 100 Hz for F2 according to Jassem (1968), and across 200 Hz for F1 and 450 Hz for F2 according to the data from Lingua Libre. This may be due to the fact that our data come from 15 speakers with various sociolinguistic markers (e.g., 5 male, 3 female and 7 unknown), which is a known source of phonetic variation (Adda-Decker and Lamel, 2005). Another factor that could add noise in the data is the segmentation process, which might have included co-articulatory effects for the vowels, which could result in a larger variation of formants as well.

	F1 (spec)	F2 (spec)	F1 (LiLi)	F2 (LiLi)
a	724	1473	769	1891
e	538	1941	566	2126
i	322	2424	331	2446
o	556	1110	618	1850
u	386	940	470	1960

Table 5: Mean F1 and F2 values (in Hertz) for the 5 cardinal vowels of Polish according to the spectrographic analysis of 5 male and 5 female speakers (Krzyśko et al., 1999) on the left and to our own analysis of Lingua Libre (LiLi) on the right.

The means are also different in Lingua Libre and in Krzyśko et al. (1999)'s data, as can be seen in Table 5, with F1 and F2 being generally higher, especially for F2 with /u/ (δ=1020 Hz), /o/ (δ=764 Hz), /a/ (δ=418 Hz) and, to a lesser extent, /e/ (δ=185 Hz). This could be due, however, to the distribution of pre-palatal consonants in each dataset (Cavar et al., 2017), which advocates for more in depth analyses, in particular regarding immediate left and right contexts. An exception is /i/, for which our results match previous results, with only a 9 Hz difference between Krzyśko et al. (1999)'s and Lingua Libre's F1 and a 22 Hz difference between Krzyśko et al. (1999) and Lingua Libre's F2. These results are encouraging for future research.

6. Conclusion and Discussion

The main goal of this paper was to compare the phoneme inventory and the vowel formants extracted from Lingua Libre with similar data from previous studies on Polish phonetics, and show that such crowd-sourced data can be useful for linguistic investigations.

For the phoneme inventory, the distribution generally matches the existing knowledge. However, for formants, we observe a partial divergence with the formants' ranges and mean values identified in previous research. This divergence in formant values is, in a way, expected, since the recording environment of Lingua Libre is much less controlled than published phonetic experiments.

This divergence could be interpreted in two ways. On the one hand, it shows the limitation of the Lingua Libre data. On the other, it also shows that there is a considerable variation between crowd-made recordings and controlled recordings, while both data sources reflect a different facet of natural production of Polish. This divergence in absolute values thus does not negate the potential of Lingua Libre data, as the recordings could still be used to investigate the relative variation of formants across vowels of the same language. As an example, the data of Lingua Libre could still be used to measure the intra-speaker variation of Polish vowels.

The use of MAUS is also to be further analyzed, as the model could have induced noise in the data by including the surrounding context of different vowels during the segmentation process.

Finally, the issue of metadata is problematic in Lingua Libre. While each contributor can provide profile information such as gender or geographical location, not all contributors do so (as shown within the Polish contributors). Therefore, it is hard to control for such variables during our analysis based on data from Lingua Libre, although they would affect phonetic realization.

As a summary, while the Lingua Libre data is not as controlled as are materials in phonetic studies, we show that it still partially matches the output of existing studies. The variation of formants also hints toward the possibility that formants observed in daily recorded speech differ from those observed in controlled environments. Both environments are relevant not only for technological purposes such as speech recognition, but also for scientific aims such as typological comparisons. Therefore, they should both be considered in future studies. In the short-term, we hope to use our methodology to investigate diphones and triphones as well as more precise acoustic measures on vowels (i.e., F3, F4 and F5) and on consonants, especially /r/ and the fricatives, while controlling for gender and regional variation as much as possible.

7. Acknowledgements

This research was partially supported by Institut DATAIA and the MSH Paris-Saclay in the framework of the Excellency Award for the project OTELO - OnTologies pour l'Enrichissement de l'analyse Linguistique de l'Oral (PI Ioana Vasilescu and Fabian Suchanek), and by the French National Research Agency in the framework of the grant EVOGRAM: The role of linguistic and non-linguistic factors in the evolution of nominal classification systems, ANR-20-CE27-0021 (PI Marc Allassonnière-Tang). The authors would also like to thank the Wikimedia community for their interest in the project, and in particular Lucas Lévêque for his help on the Lingua Libre tool.

8. Bibliographical References

Adda-Decker, M. and Lamel, L. (2005). Do speech recognizers prefer female speakers? In *INTERSPEECH*.

Albin, A. (2014). Praatr: An architecture for controlling the phonetics software "praat" with the r programming language. *Journal of the Acoustical Society of America*, 135(4):2198.

Basztura, C. (1992). *Rozmawiac z komputerem*. Wydaw.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14(3):261–290.

Cavar, M. E., Lulich, S. M., and Nelson, M. (2017). Allophonic variation of polish vowels in the context of prepalatal consonants. *The Journal of the Acoustical Society of America*, 141(5):3820–3820.

Demenko, G., Möbius, B., and Klessa, K. (2009). The design of polish speech corpus for unit selection speech synthesis. In *Language Technology*.

Izydorczy, J. and Kłosowski, P. (1999). Acoustic properties of polish vowels. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, Vol. 47, nr 1:29–37.

Jassem, W. (1958). A phonologic and acoustic classification of polish vowels. *STUF - Language Typology and Universals*, 11(1-4):299–319.

Jassem, W. (1968). Vowel formant frequencies as cues to speaker discrimination. *Speech Analysis and Synthesis*, 1:9–41.

Jassem, W. (1973). *Podstawy fonetyki akustycznej*. Panstwowe Wydaw Naukowen.

Jassem, W. (2003). Polish. *Journal of the international Phonetic Association*, 33(1):103–107.

Kipp, A., WesenickM, M.-B., and Schiel, F. (1997). 2004): Maus goes iterative. In *Proceedings of the Fifth European Conference on Speech Communication and Technology EUROSPEECH 1997*.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, September.

Kłosowski, P. (2017). Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based polish language modelling. *EURASIP Journal on Audio, Speech, and Music Processing*, 1.

Krzyśko, M., Jassem, W., and Czajka, S. (1999). The formants of polish vowels:a multivariate analysis of variance with two factors. *Speech and Language Technology*, 3(3):173–189.

Lewis, M. P., Simons, G., and Fennig, C. D. (2013). Ethnologue: Languages of the world. *SIL International*.

Marjou, X. (2021). Oteann: Estimating the transparency of orthographies with an artificial neural network. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 1–9. Association for Computational Linguistics.

Schiel, F. (2004). 2004): Maus goes iterative. In *Proceedings of the LREC 2004*, pages 1015–1018.

Wickham, H. (2017). tidyverse: Easily install and load the Tidyverse. *R package version*, 1.2.1.

Winkelmann, R., Jaensch, K., Cassidy, S., and Harrington, J., (2021). *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.3.0.

Ziółko, B. and Gałka, J. (2010). Polish phones statistics. In *Proceedings of the International Multiconfer-*

ence on Computer Science and Information Technology, pages 561–565. IEEE.

Ziółko, B., Gałka, J., Manandhar, S., Wilson, R. C., and Ziółko, M. (2009). Triphone statistics for polish language. In Zygmunt Vetulani et al., editors, *Human Language Technology. Challenges of the Information Society*, pages 63–73, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ziółko, B., Żelasko, P., and Skurzok, D. (2014). Statistics of diphones and triphones presence on the word boundaries in the polish language. applications to asr. In *XXII Annual Pacific Voice Conference (PVC)*, pages 1–6.

9. Language Resource References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *LREC*.

Demenko, G., Grocholewski, S., Klessa, K., Ogórkiewicz, J., Wagner, A., Lange, M., Śledziński, D., and Cylwik, N. (2008). Jurisdic: Polish speech database for taking dictation of legal texts. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Demenko, G., Szymański, M., Cecko, R., Kuśmierek, E., Lange, M., Wegner, K., Klessa, K., and Owsianny, M. (2012). Development of large vocabulary continuous speech recognition for polish. *Acta Physica Polonica A*, 121(1A1A):A–086–A–091.

Durand, J., Laks, B., and Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. *Romanistische Korpuslinguistik-Korpora und gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language*, 1.2.1:93–106.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Greenberg, C., Martin, A., Graff, D., Walker, K., Jones, K., and Strassel, S. (2018). 2011 nist language recognition evaluation test set. *Linguistic Data Consortium*.

Jones, K., Graff, D., Walker, K., and Strassel, S. (2016). Multi-language conversational telephone speech 2011 – slavic group. *Linguistic Data Consortium*.

Lander, T. (2005). Cslu: 22 languages corpus. *Linguistic Data Consortium*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B. (2012). Narodowy korpus języka polskiego. *Wydawnictwo Naukowe PWN*.

Schultz, T. (2002). Globalphone: A multilingual speech and text database developed at karlsruhe university. In *Proceedings of the ICSLP*, pages 345–348.

Szwelnik, T., Kawalec, J., and Gutowska, D. (2019). Polish speech database. *Linguistic Data Consortium*.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The nijmegen corpus of casual french. *Speech Communication*, 52:201–212.