Kłosowski *EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:5
DOI 10.1186/s13636-017-0102-8

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based Polish language modelling

Piotr Kłosowski

Abstract

This article presents the original results of Polish language statistical analysis, based on the orthographic and phonemic language corpus. Phonemic language corpus for Polish was developed by using automatic grapheme-to-phoneme conversion of the source orthographic language corpus, obtained from the National Corpus of Polish (NCP). The corpus contains the most frequently used Polish words, written with the use of phonemic notation. Performed statistical analysis of Polish language based on phonemic language corpus, includes frequency of occurrence calculation of the orthographic and phonemic language components, as well as their sequence. Statistical language data, obtained as a result of performed statistical analysis, enable to develop statistical word-based and phoneme-based language models for Polish. Applying these language models can effectively contribute to efficiency improvement of automatic speech recognition for Polish.

Keywords: Automatic grapheme-to-phoneme conversion, Automatic speech recognition, Language corpus, Language modelling, Language statistical analysis

Introduction

The main goal of automatic speech recognition (ASR) is translation of spoken words into a text [1]. Modern speech recognition systems require implementation of the acoustic and language modelling [2]. Both acoustic and language modelling are important parts of modern statistical speech recognition approach [3, 4]. Statistical language modelling enables to develop large vocabulary and effective speech recognition systems [5]. Language modelling can be used not only in speech recognition application, but also in other areas of speech and language processing, e.g., language recognition, machine translation, part-of-speech tagging, parsing, handwriting recognition, information retrieval and other applications.

The main motivation of the research on speech recognition area, is to improve automatic speech recognition process, especially for Polish language [6, 7]. Additionally, research studies have been conducted in the field of properties of Polish phonemes [8, 9], speech recognition based on it [10], speaker recognition [11, 12], speaker verification [13–15], and new applications of speech recognition, e.g., automatic speech translation [16].

Particularly, a good performance of automatic speech recognition is achieved with use of speech recognition by statistical methods [17]. Therefore, the main objective of the research presented in this paper, was to perform statistical analysis of Polish language based on the orthographic and phonemic language corpus, for development of statistical word-based and phoneme-based language models, as well as applying them to improve speech recognition for Polish. The development of statistical language models helps to predict a sequence of recognized spoken words and phonemes. The use of developed language models can effectively contribute to the improvement of the automatic speech recognition effectiveness, based on statistical methods. The development of word-based and phoneme-based language models for speech recognition, built on statistical language data, requires

Correspondence: pklosowski@polsl.pl
Department of Electronics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Springer Open

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

the access to large orthographic and phonemic language corpora [18, 19].

Orthographic language corpus

One of the biggest orthographic Polish language corpus is the National Corpus of Polish (NCP) [20]. The NCP corpus is available for the scientific community and offers great flexibility, as well as it is extremely important in terms of scientific value. The NCP corpus provides crucial reference material reflecting the state of contemporary Polish language which meets all the requirements of modern science [21]. It can be used particularly by linguists, but also by computer scientists interested in natural language processing.

The NCP corpus contains over 1500 million of words. The corpus is searchable by means of advanced tools, developed by the Institute of Computer Science at the Polish Academy of Sciences, which analyse Polish inflection and Polish sentence structure. The list of sources for the NCP corpus, presented in Table 1, contains classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts [22].

The results of the statistical analyses, presented in this paper, can be considered as representative for Polish language as a whole which is justified to a certain extent, considering the corpus size. However, it is worth remembering that the NCP corpus is still primarily based on written texts. Spoken language transcripts constitute a smaller percentage of the corpus contents which might be still significant when it comes to certain specialized continuous or conversational speech recognition tasks. Table 2 presents the details of the orthographic language corpus content, obtained from the NCP corpus resources.

Table 1 Structure of the NCP coprus [20]

Type of a text source	Percentage of the NCP corpus size
Daily newspapers	50.0%
Classic literature	16.0%
Non-fiction literature	5.5%
Specialized periodicals and journals	5.5%
Scientific and educational texts	2.0%
Other written texts	3.0%
Other books	1.0%
Transcripts of conversations	10.0%
Internet texts	7.0%

Table 2 Details of the orthographic language corpus content

No.	Component type	No. of unique components	No. of components in the corpus
1	single words	1,943,462	230,301,313
2	2-word sequences	75,395,184	246,110,034
3	3-word sequences	170,180,746	246,066,692
4	4-word sequences	217,586,930	246,023,356
5	5-word sequences	232,439,967	245,980,021

Phonemic language corpus

Grapheme-to-phoneme conversion

The phonemic Polish language corpus contains words written with the use of phonemic notation, obtained on the basis of automatic grapheme-to-phoneme conversion of an orthographic text. Automatic processing of a natural language, very often requires the implementation of automatic grapheme-to-phoneme conversion. Grapheme-to-phoneme conversion determines phonemic transcriptions directly from orthographic representations [23].

Phonemes are usually written with specially designed alphabets. The most commonly used alphabet for this purpose is the International Phonetic Alphabet (IPA) [24]. It was created on the basis of phonetics and phonology of West-European languages, and it is not satisfactorily adapted into Polish. For Polish, like other Slavic languages, a special transcriptional system, called the Slavistic Phonetic Alphabet (SPA), is most frequently used [25]. The second very often used phonetic alphabet is the Speech Assessment Methods Phonetic Alphabet (SAMPA) [26]. SAMPA is a machine-readable phonetic alphabet, using 7-bit printable ASCII characters, based on the IPA alphabet. Table 3 presents a set of Polish phonemes and the examples of their occurrence in Polish, written with the use of the SPA, IPA, and SAMPA phonetic alphabets.

Knowledge-based grapheme-to-phoneme approaches, unlike data-driven G2P approaches, exploit rules, created by humans or deriving from linguistic studies to convert the sequence of graphemes in a word to a sequence of phonemes [27]. Rule-based grapheme-to-phoneme approaches are typically formulated in the framework of finite state automata, and require the formulation of grapheme-to-phoneme conversion rules [28]. The largest contribution to solve the problem of automatic grapheme-to-phoneme conversion for Polish, were the publications of Maria Steffen-Batóg [29, 30].

Automatic grapheme-to-phoneme conversion process can be described as an F function, defined by the following formula:

$$F(\alpha) = \beta \tag{1}$$

where:

$$\alpha = \alpha_1 \dots \alpha_k \dots \alpha_a \ \wedge \ \alpha_k \in X \ \forall \ (1 \leq k \leq a) \tag{2}$$

$$\beta = \beta_1 \dots \beta_k \dots \beta_b \ \wedge \ \beta_k \in Y \ \forall \ (1 \leq k \leq b) \tag{3}$$

and where a is the length of orthographic character sequence, b is the length of phonemic character sequence,

Table 3 A set of Polish phonemes and examples of their occurrence

No.	Phonetic alphabet symbols [SPA]	[IPA]	[SAMPA]	Example of occurrence in Polish
1	[e]	[ɛ]	[e]	serce
2	[a]	[ɑ]	[a]	baba
3	[o]	[ɔ]	[o]	oko
4	[t]	[t]	[t]	trawa
5	[n]	[n]	[n]	noc
6	[y]	[ɨ]	[I]	syty
7	[j̯]	[j]	[j]	jajo
8	[i]	[i]	[i]	wici
9	[r]	[r]	[r]	rok
10	[s]	[s]	[s]	sok
11	[v]	[v]	[v]	wada
12	[p]	[p]	[p]	praca
13	[u]	[u]	[u]	buk
14	[m]	[m]	[m]	mama
15	[k]	[k]	[k]	kot
16	[ń]	[ɲ]	[n’]	koń
17	[d]	[d]	[d]	dudek
18	[l]	[l]	[l]	lato
19	[ł]	[w]	[w]	łysy
20	[š]	[ʃ]	[S]	szyszka
21	[f]	[f]	[f]	fala
22	[z]	[z]	[z]	koza
23	[c]	[t͡s]	[ts]	cacko
24	[b]	[b]	[b]	baba
25	[g]	[g]	[g]	godło
26	[ś]	[ɕ]	[s’]	siano
27	[ć]	[t͡ɕ]	[ts’]	ciasto
28	[ɣ́]	[j̯]	[x]	higiena
29	[č]	[t͡ʃ]	[tS]	czarny
30	[ž]	[ʒ]	[Z]	każdy
31	[ŋ]	[ŋ]	[e~]	ręka
32	[ḱ]	[c]	[k’]	kino
33	[ʒ́]	[d͡ʑ]	[dz’]	dziedzic
34	[ʒ]	[d͡z]	[dz]	nadzy
35	[ź]	[ʑ]	[z’]	ziarno
36	[ǵ]	[ɟ]	[g’]	magiczny
37	[ǯ]	[d͡ʒ]	[dZ]	drożdże

X is the set of the orthographical alphabet characters in Polish, additionally with special characters, and Y is the set of the phonemic characters alphabet in Polish, described by the Slavistic Phonetic Alphabet:

$$X = \{\text{a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, r, s, ś, t, u, w, y, z, ź, ż, q, v, x, . , ? , ! , , , : , ; , - , (,) , \# , /}\} \quad (4)$$

$$Y = \{\text{i, y, e, a, o, u, j̯, u̯, r, l, m, n, ń, ŋ, f, v, s, z, š, ž, ś, ź, χ, p, b, t, d, k, g, ḱ, ǵ, c, ʒ, č, ǯ, ć, ʒ́}\} \quad (5)$$

Grapheme-to-phoneme conversion of correctly written orthographic texts in Polish is transformation of words written in the orthographic X alphabet to form written in the phonemic alphabet Y. Automatic grapheme-to-phoneme conversion F function can be delineated by a set of formal grapheme-to-phoneme conversion rules defining how each α word, constructed from the orthographic X alphabet, can be transformed into a new β word constructed from the phonemic alphabet defined by the Y set. The rules usually are numerous with varying degrees of complexity. The size and complexity of grapheme-to-phoneme conversion rules depend on the number of letters in the orthographical alphabet and the fact that each letter can be pronounced differently in various contexts.

A set of grapheme-to-phoneme conversion rules for Polish was developed by Maria Steffen-Batóg and it was presented in the monograph dedicated to the automatic grapheme-to-phoneme conversion of texts in Polish [29, 30]. Knowledge included into these monographs was essential in developing implementation of the automatic grapheme-to-phoneme conversion algorithm for Polish. According to Maria Steffen-Batóg, all grapheme-to-phoneme conversion rules, relating to one orthographic letter, can be stored in one table, called grapheme-to-phoneme conversion rules table for one letter.

According to the grapheme-to-phoneme conversion rules for Polish, described in the literature [29–32], the grapheme-to-phoneme conversion for Polish has been implemented in the Python programming language, as automatic grapheme-to-phoneme conversion application named `TransFon` [33]. The implementation includes 975 grapheme-to-phoneme conversion rules for 35 orthographic letters in Polish, additionally conversion rules for special characters and automatic grapheme-to-phoneme conversion algorithm [33]. Block diagram of the grapheme-to-phoneme conversion algorithm for a single orthographic word is presented in Fig. 1. Due to that, many words have multiple variants of the correct pronunciation and the implementation includes only the most common basic variant of the pronunciation. Implementation of additional pronunciation variants is planned in

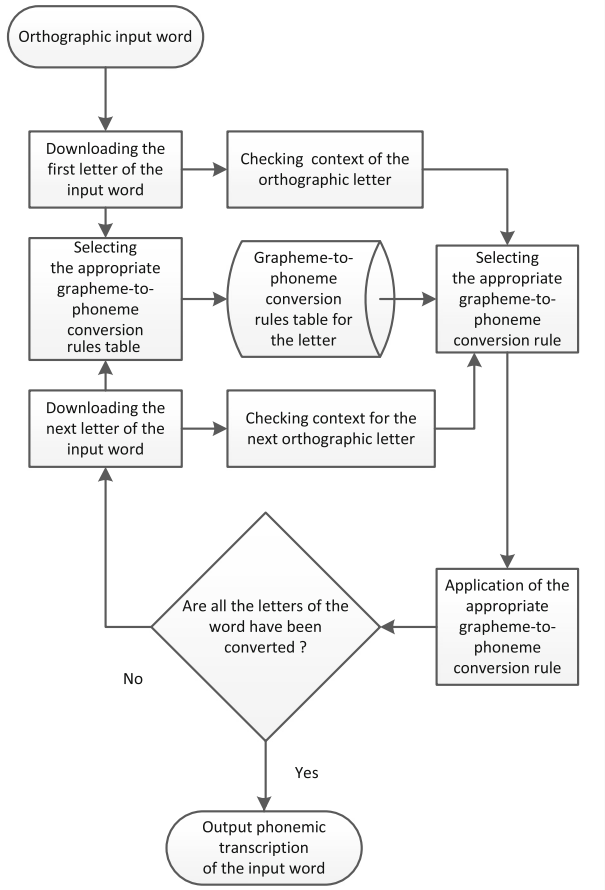

Fig. 1 Block diagram of the grapheme-to-phoneme conversion algorithm for a single orthographic word

the future. The problem of foreign words and acronyms phonemic transcription have been solved by using the dictionary where phonemic transcription of foreign words and acronyms have been defined.

`TransFon` application was developed entirely, without adapting any existing similar tools. The developed grapheme-to-phoneme conversion implementation is not the only one for Polish language [34–38], but only the one of them is available for free use [38]. The implementation of grapheme-to-phoneme conversion allows to apply it to any task (e.g., phonemic language corpus development for Polish).

Table 4 presents the phonemic transcription examples in Polish, written with the use of the SPA, IPA, and SAMPA phonetic alphabets [25, 26].

The `TransFon` application enables to create the phonemic language corpus only on the basis of the orthographic source corpus. After automatic grapheme-to-phoneme conversion of the orthographic corpus with the use `TransFon` application, phonemic language corpus for Polish was obtained, in order to perform statistical analysis of Polish language.

Evaluation of grapheme-to-phoneme conversion implementation

The evaluation of the automatic grapheme-to-phoneme conversion implementation is crucial. During implementation of automatic grapheme-to-phoneme conversion for Polish, it was necessary to check and to prove if it works properly.

The test procedure for automatic grapheme-to-phoneme conversion implementation consisted of:

- Performing the test automatic grapheme-to-phoneme conversion of orthographic text corpus file containing the most frequently used 1,943,462 unique words in Polish, obtained from the National Corpus of Polish resources [20].
- In case of doubt, validation and verification of automatic grapheme-to-phoneme conversion results for words with the use of Polish language dictionary available online, with specifying correct pronunciation of words in Polish [39].
- Registering cases of incorrect automatic grapheme-to-phoneme conversion, conversion errors and other encountered problems.

The automatic phonemic transcription application was implemented in such way, that the conversion algorithm was stopped, if grapheme-to-phoneme conversion problem occurred (e.g., when there was no rule allowing for a correct phonemic transcription). This solution makes it easier to work on improving and developing the automatic grapheme-to-phoneme conversion application. In addition, any doubts about the correct pronunciation was solved with help of wiktionary.org service [39]. This solution obviously has some serious limitations. The dictionary of wiktionary.org service contains only 61,141 Polish words and only in their basic form. The verification was further complicated by other problems such as different variants of the correct pronunciation of words or pronunciation of foreign words in the corpus.

The causes of problems and errors in automatic grapheme-to-phoneme conversion operation were as follows:

- errors in the implementation of the grapheme-to-phoneme conversion algorithm and conversion rules,
- missing grapheme-to-phoneme conversion rules in the tables (i.e., rules not included in the tables) for some orthographic letters contexts,
- grapheme-to-phoneme conversion issue of foreign words, acronyms and words, which are not present in Polish language dictionary.

The above problems were solved in the following way:

Table 4 Phonemic transcription examples in Polish

No.	Orthographic text	Phonemic transcription [SPA]	Phonemic transcription [IPA]	Phonemic transcription [SAMPA]
1	ząb	[zomp]	[zomp]	[zomp]
2	ślub	[ślup]	[ślup]	[s'lup]
3	wkręty	[fkrenty]	[fkrɛnty]	[fkrentI]
4	bieżnia	[bi̯ežńa]	[bjɛʒɲa]	[bjeZn'a]
5	wszystkie	[fšystḱe]	[fʃɨstcɛ]	[fSIstk'e]
6	natomiast	[natomi̯ast]	[natomjast]	[natomjast]
7	przypadku	[pšypatku]	[pʃɨpatku]	[pSIpatku]
8	najbardziej	[nai̯barʒ́ei̯]	[najbard͡ʑej]	[najbardz'ej]
9	oczywiście	[očyviśće]	[ot͡ʃɨviɕt͡ɕɛ]	[otSIvis'ts'e]
10	powiedział	[povi̯eʒ́au̯]	[povjed͡ʑaw]	[povjedz'aw]

- The errors in the implementation of the grapheme-to-phoneme conversion algorithm and in conversion rules tables have been corrected by modifications, made within an application source code in Python programming language.
- The problem of missing grapheme-to-phoneme conversion rules in tables has been solved by adding new conversion rules to the existing tables. In order to complete the missing grapheme-to-phoneme conversion rules, new conversion rules were supplemented for the following orthographic letters "i", "n", "d", "z", "z", "c", "f", "s", in some contexts.
- The problems of foreign words and acronyms, have been solved by using the dictionary, where phonemic transcription of foreign words and acronyms have been defined. As a result, rule-based automatic grapheme-to-phoneme conversion was complemented by dictionary-based automatic grapheme-to-phoneme conversion method.

A number of improvements made it possible to increase effectiveness of the grapheme-to-phoneme conversion implementation. Tables 5 and 6 present the word error rate (WER) values of grapheme-to-phoneme conversion implementation, before and after improvements.

The WER value for 1,943,462 checked unique words, was equal 0.387%. The WER value for corpus contains 230,301,313 words, was equal 0.030%. The changes of WER values, before and after improvements, testify to the fact that implemented modifications have contributed to improving the effectiveness of G2P conversion.

The developed phonemic language corpus for Polish

The phonemic language corpus for Polish was developed by automatic grapheme-to-phoneme conversion of the source orthographic language corpus file obtained from the NCP corpus resources.

Table 7 presents the details of the phonemic language corpus content.

The phonemic language corpus contains the list of 1, 943, 462 Polish words written orthographically, their phonemic transcription written with the SAMPA phonemic alphabet and additionally, the number of word occurrence in the NCP balanced corpus. The measure of the NCP balanced corpus size is the sum of all numbers of the word occurrences, which is equal to 230,301,313 words.

Table 5 WER values of the developed G2P conversion implementation, before improvements

No.	Parameter	Value
1	No. of checked unique words	1,943,462
2	No. of G2P conversion errors for unique words	33,638
3	WER value for unique words in %	1.731
4	No. of words in the corpus	230,301,313
5	No. of G2P conversion errors for words in corpus	3,707,890
6	WER value for corpus in %	1.610

Table 6 WER values of the developed G2P conversion implementation, after improvements

No.	Parameter	Value
1	No. of checked unique words	1,943,462
2	No. of G2P conversion errors for unique words	7525
3	WER value for unique words in %	0.387
4	No. of words in the corpus	230,301,313
5	No. of G2P conversion errors for words in corpus	69,802
6	WER value for words in the corpus in %	0.030

Table 7 Details of the phonemic language corpus content

No.	Component type	No. of unique components	No. of components in the corpus
1	single phonemes	37	1,263,248,497
2	2-phoneme sequences	1096	1,032,922,921
3	3-phoneme sequences	17,340	823,393,519
4	4-phoneme sequences	128,766	644,597,673
5	5-phoneme sequences	402,529	483,987,550

A sample section of the developed phonemic language corpus for Polish is presented in Table 8. It should also be noted that the standard SAMPA for Polish includes several sequences of phonemic transcription labels that may cause ambiguity unless separated by spaces or other characters. To avoid this problem, all phonemes are separated by square brackets.

Analysis of the obtained results and discussion

Statistical analysis of the orthographic and phonemic language corpora

With the use of the orthographic and phonemic language corpora, it was possible to perform statistical analysis of Polish language which includes calculation of the following distributions:

- the frequency of the single orthographic word occurrence,
- the frequency of the n-word sequence occurrence for $n = 2, \ldots, 5$,
- the frequency of the phoneme occurrence,
- the frequency of the n-phoneme sequence occurrence for $n = 2, \ldots, 5$.

The frequency distribution of words in the orthographic language corpus, is presented in Fig. 2.

A sample calculated frequency of word occurrence, is presented in Table 9, where 1% corresponds to about 2303013 occurrences.

A sample calculated frequency of occurrence for the two-word and the three-word sequences, are presented in Tables 10 and 11. The results for the four-word and the five-word sequences, are not presented in this paper, but they can also be helpful to develop advanced word-based language models.

The frequency distribution of the phonemes in the phonemic language corpus, is presented in Fig. 3.

The frequency distributions of the n-phoneme sequences, for $n = 2, \ldots, 5$, are presented in Fig. 4.

Evaluation of the obtained results

The results of the research on statistical analysis of Polish language, performed with the phonemic language corpus, were compared to other results published in the literature [40–46]. Summary comparisons of the obtained statistical language data, to other results, available in the literature, are presented in Tables:

- Table 12 presents the occurrence frequency of Polish phonemes and comparison to the results published in the literature [40, 42, 44, 45],
- Table 13 presents the occurrence frequency of the two-phoneme sequences (diphones) in Polish and comparison to the results published in the literature [45],

Table 8 A sample section of the developed phonemic language corpus for Polish

	Number of occurr. of 230301313	Orthographic word	Phonemic transcription of word
i	$C(w_i)$	w_i	[SAMPA]
1	7,692,997	w	`[f]`
2	5,333,210	i	`[i]`
3	4,235,003	na	`[n] [a]`
4	4,158,902	z	`[s]`
5	3,981,525	się	`[s'] [e]`
6	3,601,719	nie	`[n'] [e]`
7	2,904,114	do	`[d] [o]`
8	2,205,896	że	`[Z] [e]`
9	2,171,877	to	`[t] [o]`
10	1,731,304	o	`[o]`
11	1,728,527	jest	`[j] [e] [s] [t]`
12	1,425,793	a	`[a]`
13	1,003,027	jak	`[j] [a] [k]`
14	983,395	po	`[p] [o]`
15	912,660	od	`[o] [t]`
16	877,522	ale	`[a] [l] [e]`
17	847,373	za	`[z] [a]`
18	775,006	przez	`[p] [S] [e] [s]`
19	754,024	co	`[ts] [o]`
20	663,771	dla	`[d] [l] [a]`
21	645,573	czy	`[tS] [I]`
22	610,035	tym	`[t] [I] [m]`
23	607,673	już	`[j] [u] [S]`
24	544,343	tak	`[t] [a] [k]`
25	534,509	tylko	`[t] [I] [l] [k] [o]`
26	500,801	ma	`[m] [a]`
27	475,172	może	`[m] [o] [Z] [e]`
28	451,225	tego	`[t] [e] [g] [o]`
29	445,705	ze	`[z] [e]`
30	426,201	jego	`[j] [e] [g] [o]`
...	...	...	

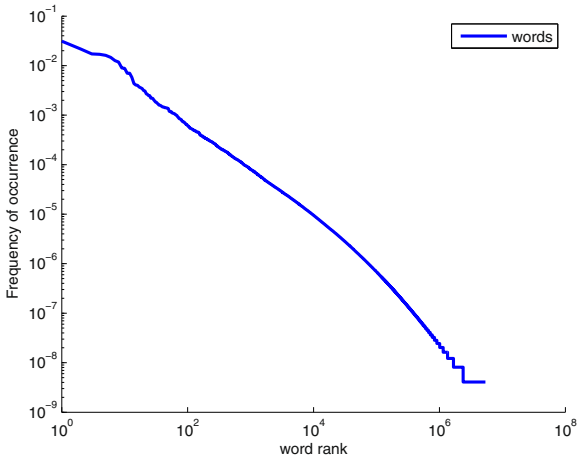

Fig. 2 Frequency distribution of the word occurrence in Polish language

- Table 14 presents the occurrence frequency of the three-phoneme sequences (triphones) in Polish and comparison to the results published in the literature [45].

The reasons of differences among the obtained results of the language statistical analysis performed by other scientists may be: differences in used corpora (e.g., in size, quality, linguistic structure) and development of language and changes over time. Language is constantly changing, evolving, and adapting to the needs of its speakers. All languages change continually, and do so in many and varied ways (e.g., lexical changes, phonetic and phonological changes, spelling changes, semantic and syntactic changes) [47]. Therefore, a results of research performed using different corpora may be very different from each other [48, 49]. The most similar results apply statistical analysis of Polish phonemes occurrence presented in Table 12 [44, 45]. The least accurate results were obtained with much smaller language corpus a few decades ago [40–42]. Taking into account the results, available in the literature, it can be concluded that performed statistical analysis of Polish language, was extensive. No results of a statistical analysis of the n-phoneme sequences occurrence in Polish for $n > 3$ were found in the literature. On the basis of the comparison results, the following conclusion can be drawn: The developed phonemic language corpus in Polish, which was used to perform statistical analysis of Polish language, was very huge, containing 1263248497 phonemes, but not the biggest developed for Polish language [44]. The statistical analysis results obtained based on it, allow to develop statistical models of Polish language.

Table 9 Frequency of the word occurrence in the orthographic corpus file

No.	Frequency of occurrence	Word
i	$f(w_i) \cdot 100$ [%]	w_i
1	3.34041	w
2	2.31575	i
3	1.83890	na
4	1.80585	z
5	1.72883	się
6	1.56392	nie
7	1.26101	do
8	0.95783	że
9	0.94306	to
10	0.75176	o
11	0.75055	jest
12	0.61910	a
13	0.43553	jak
14	0.42700	po
15	0.39629	od
16	0.38103	ale
17	0.36794	za
18	0.33652	przez
19	0.32741	co
20	0.28822	dla
21	0.28032	czy
22	0.26489	tym
23	0.26386	już
24	0.23640	są
25	0.23636	tak
26	0.23209	tylko
27	0.21745	ma
28	0.20633	może
29	0.19593	tego
30	0.19353	ze
...	...	...

Frequency of the word occurrence

The frequency of word occurrence in a language is well described by Zipf's law [50, 51]:

$$Z_r = \frac{a}{r^b} \tag{6}$$

where Z_r is the frequency of the word ranked r, where r is the rank of the word if frequencies are ranked from the most frequent ($r = 1$) to the least frequent ($r = n$), and a and b are parameters to be estimated from obtained statistical data. The usual findings is that b is close to 1 [50]. The fit of Zipf's equation to the ranked frequency distribution of Polish words is presented in Fig. 5.

Table 10 Frequency of the two-word sequence occurrence in the orthographic corpus file

No.	Frequency of occurrence $f(w_{i-1}, w_i) \cdot 100$ [%]	2-word sequence $w_{i-1}w_i$
1	0.12643	się w
2	0.10243	w tym
3	0.08805	się na
4	0.08287	się z
5	0.07604	się do
6	0.06529	nie ma
7	0.05799	nie jest
8	0.05597	się, że
9	0.04932	w tej
10	0.04741	jest to
11	0.04602	że w
12	0.04595	że nie
13	0.04209	i w
14	0.04209	nie tylko
15	0.04118	to nie
16	0.04083	i nie
17	0.03882	to jest
18	0.03698	się nie
19	0.03145	to, że
20	0.02843	że to
21	0.02751	ale nie
22	0.02725	przede wszystkim
23	0.02718	w Polsce
24	0.02675	a w
25	0.02671	a nie
26	0.02668	jest w
27	0.02640	po prostu
28	0.02629	w którym
29	0.02599	jak i
30	0.02543	nie było
...	...	...

Table 11 Frequency of the three-word sequence occurrence in the orthographic corpus file

No.	Frequency of occurrence $f(w_{i-2}, w_{i-1}, w_i) \cdot 100$ [%]	3-word sequence $w_{i-2}w_{i-1}w_i$
1	0.01890	w związku z
2	0.01325	w tym roku
3	0.01238	ze względu na
4	0.01195	w ten sposób
5	0.00921	na to, że
6	0.00920	okazało się, że
7	0.00915	w tej chwili
8	0.00902	o tym, że
9	0.00892	po raz pierwszy
10	0.00891	w stosunku do
11	0.00871	do tej pory
12	0.00791	w tej sprawie
13	0.00786	jeśli chodzi o
14	0.00728	związku z tym
15	0.00665	to nie jest
16	0.00615	nie jest to
17	0.00588	o których mowa
18	0.00582	których mowa w
19	0.00566	że jest to
20	0.00529	w tym czasie
21	0.00522	w tym samym
22	0.00505	nie moze być
23	0.00498	w ogóle nie
24	0.00498	mi się, że
25	0.00473	że nie ma
26	0.00472	nie da się
27	0.00442	w ubiegłym roku
28	0.00440	mam nadzieję, że
29	0.00421	w zależności od
30	0.00410	na tym, że
...	...	...

The ranked frequency distribution of Polish words was estimated by Zipf's equation in the following form:

$$Z_r = \frac{0.041566}{r^{0.9}} \tag{7}$$

The average fit of Zipf's equation to the ranked frequency distribution of Polish words was measured by the coefficient of determination R^2 value. The coefficient of determination for fit of Zipf's equation, presented in Equation (7), to the ranked frequency distribution of Polish words is equal:

$$R^2 = 0.90729 \tag{8}$$

Additionally, root-mean-square error RMSE value was calculated for this case and it is equal:

$$RMSE = 7.6475 \cdot 10^{-6} \tag{9}$$

The R^2 value indicates how well statistical data fit into a statistical model. The R^2 value equals $R^2 = 0.90729$ indicates that the Zipf's equation fits well to the obtained statistical data of the word occurrence frequency in Polish language.

On this basis and on the basis of the results available in the literature [51–53], it can be concluded that the statistical data, obtained as the result of performed statistical

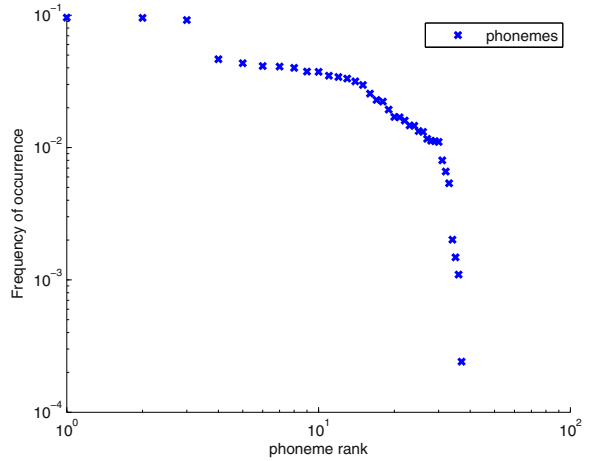

Fig. 3 Frequency distribution of the phoneme occurrence in Polish language

analysis of Polish language, based on the orthographic language corpus, are correct.

Frequency of the phoneme and n-phoneme sequence occurrence

The frequency of word occurrence in a language is well described by Zipf's law [50]. However, Zipf's law does not describe well the distribution of the phonemes and phoneme sequences out of which words are composed. The examination of occurrence frequency in 95 languages, presented in the literature [51], shows that phoneme frequencies are best described by an equation first developed by Yule, that also describes the distribution of DNA codons [54]. The frequency of the phoneme occurrence in a language is described well by Yule's equation formula [51]:

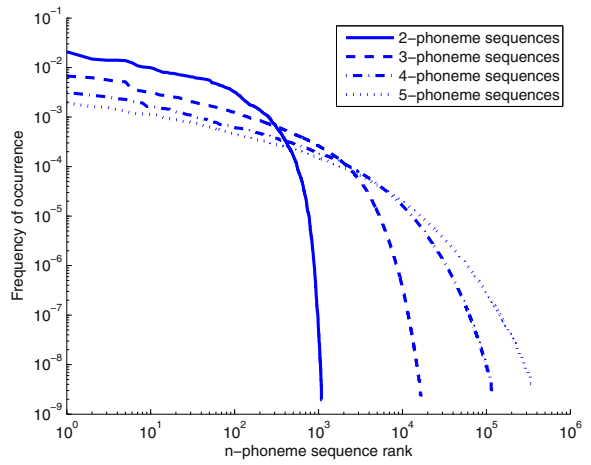

Fig. 4 Frequency distribution of the *n*-phoneme sequence occurrence in Polish language, for $n = 2, \ldots, 5$

$$Y_r = \frac{a}{r^b} \cdot c^r \tag{10}$$

where Y_r is the frequency of the phoneme ranked r, and r is the rank of the phoneme if frequencies are ranked from the most frequent ($r = 1$) to the least frequent ($r = n$), and a, b and c are parameters to be estimated from the obtained statistical data.

The fits of Zipf's and Yule's equations to the ranked frequency distribution of Polish phonemes are presented in Fig. 6.

The evaluation results of the fits of Zipf's and Yule's equations to the ranked frequency distribution of Polish phonemes are presented in Table 15.

Note that the Zipf's equation is a special case of the Yule's equation in which c^r is neglected. It is not always possible to neglect this term. As shown in Fig. 6 and in Table 15, the Yule's equation fits to the distribution of the phoneme frequencies in Polish much better than the Zipf's equation. It is not an isolated case and similar regularity can be observed in other languages [51].

The same regularity was observed for frequency distributions of the *n*-phoneme sequence occurrence for Polish language, for $n = 2, ..., 5$. The Figs. 7 and 8 present the fit of Yule's equation to the ranked frequency distribution of Polish *n*-phoneme sequences for $n = 2$ and $n = 3$.

The summary of evaluation results of the Yule's equation fits to the ranked frequency distribution of Polish phonemes and the *n*-phoneme sequences for $n = 2, \ldots, 5$ are presented in Table 16.

The values of R^2, presented in Table 16, indicate that the Yule's equation fits very well to the obtained statistical data of frequency occurrence of Polish phonemes and the *n*-phoneme sequences for $n = 2, ..., 5$. A similar properties are observed for other languages. On the basis of the obtained results and the results available in the literature [40, 41, 43–46, 51], it can be concluded that statistical data, obtained as the result of performed statistical analysis of Polish language, based on the orthographic and phonemic language corpora, are correct.

Example of practical application of the obtained results for language modelling

This article contains a general statistics of Polish language that can be useful for a variety of language and speech processing applications, including automatic speech recognition with language models [55].

The goal of the word-based language model, is to model the sequence of words in the context of the task, being performed by the speech recognition system. In continuous speech recognition, the incorporation of the language model is crucial to reduce the search speed of recognized words sequence W. The probability $P(W)$ of occurrence W, sequence of n words w_i, can be decomposed as [17]:

Table 12 Frequency of Polish phoneme occurrence—comparison to the results published in the literature [40, 42, 44, 45]

No.	Obtained results	Phoneme [SAMPA]	Results in [44]	Results in [45]	Results in [42]	Results in [40]
i	$f(q_i)$ [%]	$[q_i]$	$f(q_i)$ [%]	$f(q_i)$ [%]	$f(q_i)$ [%]	$f(q_i)$ [%]
1	9.59478	[e]	7.882	9.108	10.6	10.2
2	9.55135	[a]	8.141	9.584	9.7	9.3
3	9.20001	[o]	7.646	8.994	8.0	9.1
4	4.65630	[t]	3.708	4.489	4.8	4.4
5	4.34100	[n]	3.665	4.443	4.0	4.0
6	4.13142	[I]	3.174	3.648	3.8	4.1
7	4.09810	[j]	3.299	3.796	4.4	4.5
8	4.00623	[i]	3.620	4.359	3.4	3.9
9	3.75265	[r]	3.705	4.674	3.2	3.6
10	3.73464	[s]	2.927	3.638	2.8	3.0
11	3.49063	[v]	3.137	3.782	2.9	3.5
12	3.41265	[p]	2.759	3.263	3.0	3.1
13	3.32576	[u]	2.774	3.345	2.8	3.4
14	3.16465	[m]	2.626	2.988	3.2	3.5
15	2.96802	[k]	2.418	2.976	2.5	2.7
16	2.55419	[n']	1.840	2.088	2.4	2.6
17	2.29278	[d]	2.391	2.888	2.1	2.2
18	2.22555	[l]	2.164	2.642	1.9	2.1
19	1.93507	[w]	1.626	1.636	1.8	2.2
20	1.70517	[S]	1.118	1.215	1.9	2.0
21	1.69430	[f]	1.363	1.683	1.3	1.5
22	1.60077	[z]	1.665	1.947	1.5	1.8
23	1.46934	[ts]	1.335	1.692	1.2	1.5
24	1.46097	[b]	1.304	1.497	1.5	1.5
25	1.33050	[g]	1.341	1.547	1.3	1.5
26	1.31409	[s']	0.927	0.965	1.6	1.5
27	1.16326	[ts']	0.643	0.662	1.2	1.3
28	1.12532	[x]	1.153	1.427	1.0	1.1
29	1.11761	[tS]	0.831	0.955	1.2	1.2
30	1.10377	[Z]	0.884	0.944	1.3	1.2
31	0.79984	[e~]	0.582	0.673	0.6	0.7
32	0.65927	[k']	0.570	0.698	0.7	n.a.
33	0.53682	[dz']	0.538	0.554	0.7	0.8
34	0.20125	[dz]	0.227	0.261	0.2	0.2
35	0.14815	[z']	0.183	0.195	0.2	0.2
36	0.10971	[g']	0.198	0.260	0.1	n.a.
37	0.02412	[dZ]	0.037	0.040	0.1	0.0

$$P(W) = P(w_1) \prod_{i=2}^{n} P(w_i|w_1, \ldots, w_{i-1}) \tag{11}$$

where $P(w_i|w_1, \ldots, w_{i-1})$ is the conditional probability that w_i will occur, given the previous word sequence $w_1, \ldots, w_{i-1}$. Unfortunately, it is impossible to compute the conditional word probabilities $P(w_i|w_1, \ldots, w_{i-1})$ for all words and all sequence lengths in a given language. Even though the sequences are limited to moderate values of i, there would not be enough data to estimate reliably all of the conditional probabilities. The conditional probability can be approximated by estimating the

Table 13 Frequency of the two-phoneme sequence occurrence in Polish—comparison to the results published in the literature [45]

No.	Obtained results $f(q_{i-1}, q_i)$ [%]	Diphone [SAMPA] $[q_{i-1}]\,[q_i]$	Results in [45] $f(q_{i-1}, q_i)$ [%]
1	2.09086	[j] [e]	1.7253
2	1.48817	[n] [a]	1.1632
3	1.40880	[n'] [e]	0.8438
4	1.40198	[s] [t]	1.0791
5	1.38280	[p] [o]	1.0479
6	1.25078	[o] [v]	1.1829
7	1.08491	[r] [a]	0.9189
8	1.03023	[o] [n]	0.8756
9	1.01037	[r] [o]	0.9155
10	0.99573	[v] [a]	0.8012
11	0.94847	[t] [a]	0.8035
12	0.88593	[k] [o]	0.7337
13	0.84639	[j] [a]	0.6367
14	0.79998	[o] [e~]	0.506
15	0.79985	[v] [j]	0.442
16	0.79298	[d] [o]	0.6459
17	0.76749	[e] [j]	0.6620
18	0.76340	[a] [w]	0.5595
19	0.75699	[t] [e]	0.6229
20	0.74761	[t] [o]	0.5814
21	0.72197	[z] [a]	0.497
22	0.71902	[e] [m]	0.60411
23	0.71384	[g] [o]	0.515
24	0.69492	[e] [n]	0.6768
25	0.68893	[S] [e]	0.456
26	0.68631	[k] [a]	0.540
27	0.67323	[n] [e]	0.60803
28	0.67050	[v] [I]	0.526
29	0.66791	[l] [i]	0.58227
...	...	...	...

Table 14 Frequency of the three-phoneme sequence occurrence in Polish—comparison to the results published in the literature [45]

No.	Obtained results $f(q_{i-2}, q_{i-1}, q_i)$ [%]	Triphone [SAMPA] $[q_{i-2}]\,[q_{i-1}]\,[q_i]$	Results in [45] $f(q_{i-2}, q_{i-1}, q_i)$ [%]
1	0.67353	[v] [j] [e]	0.3159
2	0.62188	[e] [g] [o]	0.3655
3	0.56670	[o] [v] [a]	0.3801
4	0.52262	[s] [t] [a]	0.3287
5	0.51677	[p] [S] [e]	0.2969
6	0.36109	[m] [j] [e]	0.2503
7	0.34557	[e] [s] [t]	0.1734
8	0.33317	[o] [n] [ts]	0.1749
9	0.32041	[p] [r] [a]	0.1681
10	0.31920	[o] [s'] [ts']	0.1533
11	0.31277	[j] [o] [n]	0.189
12	0.28832	[j] [o] [e~]	0.143
13	0.27851	[p] [S] [I]	0.118
14	0.27638	[k] [t] [u]	0.1311
15	0.27428	[p] [r] [o]	0.1807
16	0.26887	[t] [u] [r]	0.1448
17	0.25608	[n] [I] [x]	0.1673
18	0.25453	[o] [v] [j]	0.1404
19	0.25330	[j] [e] [s]	n.a.
20	0.25049	[o] [s] [t]	0.1785
21	0.24914	[p] [j] [e]	0.120
22	0.24394	[e] [n] [t]	0.1842
23	0.23760	[a] [j] [o]	0.126
24	0.23674	[a] [l] [e]	0.116
25	0.22874	[s'] [ts'] [i]	0.130
26	0.22739	[p] [o] [v]	0.122
27	0.22454	[a] [n'] [e]	0.1586
28	0.21298	[s] [p] [o]	0.1627
29	0.20990	[o] [v] [e]	0.1712
...	...	...	...

probability only on the preceding $N-1$ words defined by the following formula:

$$P(W) = P(w_1) \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1}) \tag{12}$$

This approximation is commonly referred to as N-gram model [17]. The most popular solutions published in the literature, relate to the application of N-gram language models for word-based speech recognition tasks [56–59].

The language modelling may be based on modelling of words, as well as sub-words (e.g. phonemes). Statistical analysis of the phonemic corpus enables to develop statistical language models, based on phonemes.

For sequence of the phonemes $Q = q_1 \ldots q_m$, containing m phonemes q_i, the probability $P(Q)$ is given by a phoneme-based language model and the following formula:

$$P(Q) = P(q_1) \prod_{i=2}^{m} P(q_i | q_1, \ldots, q_{i-1}) \tag{13}$$

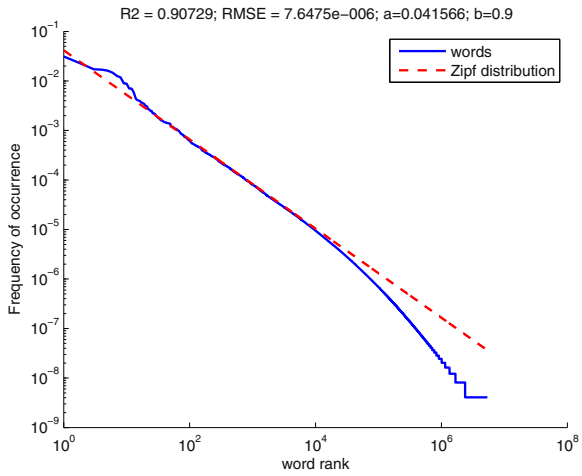

Fig. 5 Fit of Zipf's equation to the ranked frequency distribution of Polish words

where $P(q_i|q_1,\ldots,q_{i-1})$ is the conditional probability that q_i will occur, given the previous phoneme sequence $q_1,\ldots,q_{i-1}$. The $P(Q)$ probability approximation for N-gram phoneme-based language model is defined by the analogous formula:

$$P(Q) = P(q_1)\prod_{i=2}^{m} P(q_i|q_{i-N+1},\ldots,q_{i-1}) \tag{14}$$

On the basis of performed statistical analysis of the orthographic language corpus, there have been developed the N-gram word-based language models for $N = 1,\ldots,3$, intended for Polish language. In a similar way, on the basis of statistical analysis results of the phonemic language corpus, the N-gram phoneme-based language models for $N = 1,\ldots,3$, intended for Polish language, were developed. The details of word-based and phoneme-based language models developing process are presented in the separate publication. This article presents only the example of language statistical analysis application to develop selected language models.

Table 15 Evaluation results of the fits of Zipf's and Yule's equations to the ranked frequency distribution of Polish phonemes

No.	Equation	R^2	*RMSE*
1	$Z_r = \frac{0.095948}{r^{0.5}}$	0.81180	$1.0408 \cdot 10^{-2}$
2	$Y_r = \frac{0.095948}{r^{0.149}} \cdot 0.94^r$	0.92198	$6.7012 \cdot 10^{-3}$

An approach to evaluate a language model is word recognition error rate [60].

However, this approach requires a working speech recognition system. Alternatively, we can measure the average number of possible words that follow any given word sequence in a language. This is the derivative measure of entropy, known as perplexity (PP) [17]. Given a language model $P(W)$, where W is the n-word sequence, the entropy of the language model can be defined as [61]:

$$H(W) = -\frac{1}{n}\log_2(P(W)) \tag{15}$$

For N-gram language model, $H(W)$ entropy can be calculated with the following formula:

$$H(W) = -\frac{1}{n}\sum_{i=1}^{n}\log_2(P(w_i|w_{i-N+1},\ldots,w_{i-1})) \tag{16}$$

Note that as n approaches infinity, the entropy approaches the asymptotic entropy of the source defined by the measure $P(W)$. This means that the typical length of the sequence must approach infinity, which is of course impossible. Thus, entropy $H(W)$ should be estimated on

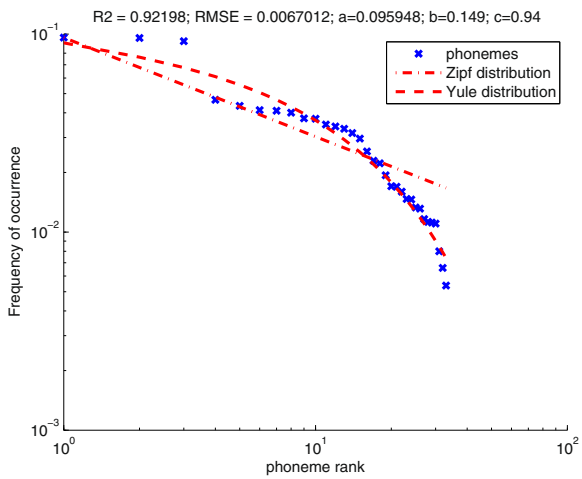

Fig. 6 Fits of Zipf's and Yule's equations to the ranked frequency distribution of Polish phonemes

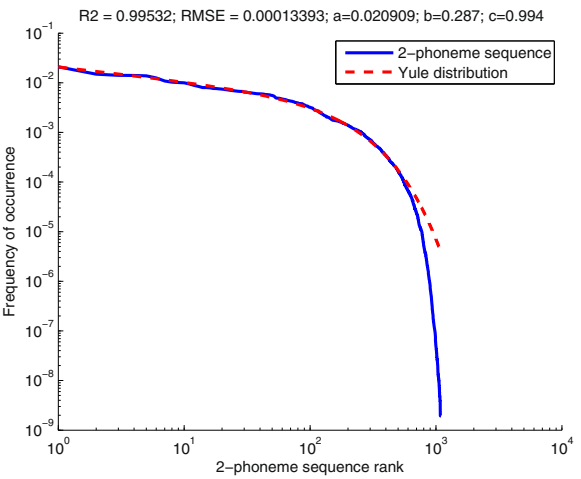

Fig. 7 Fit of Yule's equation to the ranked frequency distribution of Polish two-phoneme sequences

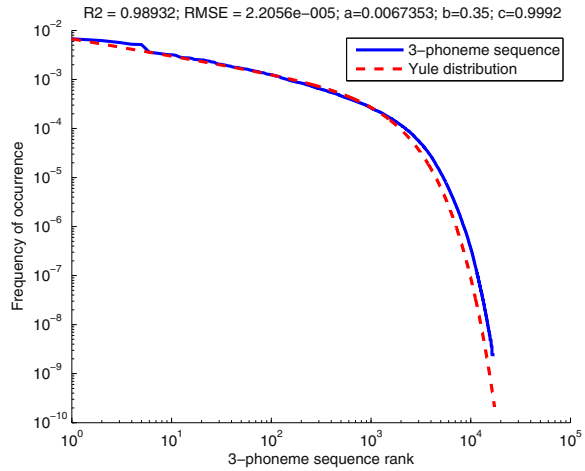

Fig. 8 Fit of Yule's equation to the ranked frequency distribution of Polish three-phoneme sequences

a sufficient large n value. The perplexity $PP(W)$ of the word-based language model is then defined as [17]:

$$PP(W) = 2^{H(W)} \tag{17}$$

The comparison of perplexity $PP_N(W)$ values for the developed word-based N-gram language models for $N = 1,\ldots,3$, is presented in Table 17. The comparison of perplexity $PP_N(Q)$ values for the developed phoneme-based N-gram language models for $N = 1,\ldots,3$, is presented in Table 18.

The PP values, presented in Tables 17 and 18, show that the developed phoneme-based 3-gram language model has the lowest PP value equal to 7.77. The lower perplexity value for language model indicates a greater ability to predict sequence of speech components. A language model is rated as better if the perplexity PP value is less. A language models with low perplexity indicate more predictable language. However, since the perplexity is not related to the complexity of recognizing some acoustic patterns, reducing the language model, perplexity does not guarantee an improvement in automatic speech recognition performance.

Table 16 Evaluation results of the fit of Yule's equation to the ranked frequency distribution of Polish phonemes ($n = 1$) and n-phoneme sequences for $n = 2,\ldots,5$

n	Yule's equation	R^2	$RMSE$
1	$Y_r = \frac{0.095948}{r^{0.149}} \cdot 0.94^r$	0.92198	$6.7012 \cdot 10^{-3}$
2	$Y_r = \frac{0.020909}{r^{0.287}} \cdot 0.994^r$	0.99532	$1.3393 \cdot 10^{-4}$
3	$Y_r = \frac{0.0067353}{r^{0.35}} \cdot 0.9992^r$	0.98932	$2.2056 \cdot 10^{-5}$
4	$Y_r = \frac{0.0031449}{r^{0.35}} \cdot 0.9997^r$	0.98510	$5.5017 \cdot 10^{-6}$
5	$Y_r = \frac{0.0019385}{r^{0.33}} \cdot 0.9998^r$	0.98028	$2.6493 \cdot 10^{-6}$

Table 17 Comparison of perplexity $PP_N(W)$ values for the developed word-based N-gram language model for $N = 1,\ldots,3$

N	Perplexity	Value	Word-based language model
1	$PP_1(W)$	9317.1	1-gram word-based language model
2	$PP_2(W)$	933.0	2-gram word-based language model
3	$PP_3(W)$	278.9	3-gram word-based language model

Potential application of other statistical analysis results

The statistical analysis results for 4 and 5-word sequence occurrence are not presented in this paper. But these results can be helpful to develop advanced (4 and 5-gram) word-based language models for Polish. As previously written, the language modelling may be based on modelling of words, as well as sub-words (e.g., phonemes). Therefore, the statistics of higher than three-phoneme sequence can be used for developing advanced (higher than 3-gram) phoneme-based language models for Polish. The advanced word-based and phoneme-based language modelling, enables to develop a hybrid language models for out-of-vocabulary (OOV) word detection in large vocabulary conversational speech recognition (LVCSR) systems for the language [62, 63]. The language model in most state-of-the-art LVCSR systems is still the N-gram, which assigns probability to the next word based on only the $N-1$ preceding words [64]. But the use of an additional phoneme-based language models improves efficiency of LVCSR systems [65]. Another improvement in an LVCSR system development is the use of higher than 4-gram language models, with particular emphasis on N-gram phoneme-based language models.

Conclusions

This paper presents the original results of statistical analysis of Polish language, performed by means of the orthographic language text corpus, obtained from the NCP corpus and the phonemic language corpus, developed through automatic grapheme-to-phoneme conversion of the orthographic language corpus. The results of statistical analysis of Polish language, enable to develop statistical word-based and phoneme-based language models, in order to be used for automatic speech recognition.

The results of the research on statistical analysis of Polish language were compared and are consistent to other results available in the literature [40–46, 66, 67].

Table 18 Comparison of perplexity $PP_N(Q)$ values for the developed phoneme-based N-gram language model for $N = 1,\ldots,3$

N	Perplexity	Value	Phoneme-based language model
1	$PP_1(Q)$	22.08	1-gram phoneme-based language model
2	$PP_2(Q)$	10.85	2-gram phoneme-based language model
3	$PP_3(Q)$	7.77	3-gram phoneme-based language model

Taking into account the results available in the literature, it can be concluded that performed statistical analysis of the language was extensive. No results of the statistical analysis of n-phoneme sequence occurrence in Polish for $n > 3$ were found in the literature. On the basis of the comparison results, the following conclusion can be drawn: The phonemic language corpus in Polish which used to perform statistical analysis of the language, was very huge (containing 1,263,248,497 phonemes) and the statistical analysis results, obtained and based on it, allows to develop statistical models of Polish language.

Additionally, the validation and evaluation of the obtained statistical data were performed. The frequency of the word occurrence in a language is well described by Zipf's law. The validation of statistical data for words was performed by the fit of Zipf's equation to the ranked frequency distribution of Polish words. Similar regularity was observed for frequency distribution of the phoneme occurrence for Polish language. The examination of frequency occurrence in 95 languages, presented in the literature [51], shows that phoneme frequencies are best described by Yule's equation [54]. The validation of the statistical data for phonemes was performed by the fit of Yule's equations to the ranked frequency distribution of Polish phonemes and n-phoneme sequences. According to the results available in the literature [51], it can be concluded that statistical data obtained as the result of performed statistical analysis of Polish language, based on the orthographic and phonemic language corpora, are correct.

Regularity presented in this paper, it is not an isolated case and similar regularity can be observed in other languages, so also for other language corpora, reflecting the state of contemporary language [51]. It should also be noted, that it seems to be valuable to provide similar fits for existing Polish text corpora for allowing the reader to assess the quality of the created phonemic language corpus. Similarly, it seems to be very valuable to confront word error rate and the perplexity of the language models, created by means of the existing Polish corpora with respect to a common test set. However, it is difficult to perform due to lack of access to other existing Polish text corpora of appropriate size and quality, except NCP corpus. Similarly, the author does not find any available phonemic language corpus for Polish. Therefore, the author attempts to create his own phonemic language corpus with the use of G2P conversion of the existing available orthographic language corpus for Polish (NCP). Since this problem seems to be very important, the author is planning to bring this subject up in the future publications.

The developed word-based and phoneme-based language models were also presented in this paper, as an example of practical applications of the obtained statistical data of Polish language. The obtained statistical data open up further opportunities to continue research on improving automatic speech recognition in Polish. The plan for future research includes the development of statistical word-based and subword-based language models for Polish. The word-based and subword-based language modelling, enables to develop a hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition [64, 68–70].

Acknowledgements
This work was supported by the Polish Ministry of Science and Higher Education funding for statutory activities.

Competing interests
The authors declare that they have no competing interests.

Received: 15 April 2016 Accepted: 8 February 2017
Published online: 28 February 2017

References

1. L Rabiner, B Juang, *Fundamentals of Speech Recognition. Prentice Hall signal processing series*. (PTR Prentice Hall, USA, 1993)
2. JR Bellegarda, C Monz, State of the art in statistical methods for language and speech processing. Comput. Speech Lang. **35**, 163–184 (2016)
3. L Rabiner, B Juang, *Encyclopedia of Language and Linguistics, Statistical methods for the recognition and understanding of speech*. (Elsevier, Amsterdam, 2005)
4. S Sakti, K Markov, S Nakamura, W Minker, in *Incorporating Knowledge Sources into Statistical Speech Recognition, vol 42 of Lecture Notes in Electrical Engineering*. Statistical Speech Recognition (Springer US, USA, 2009), pp. 19–53
5. J Bellegarda, Large vocabulary speech recognition with multispan statistical language models. IEEE Transa. Speech Audio Process. **8**, 76–84 (2000)
6. P Kłosowski, in *Computer Nerworks vol 79 of Communications in Computer and Information Science*, ed. by A Kwiecien, P Gaj, and P Stera. Speech processing application based on phonetics and phonology of the polish language. 17th International Conference Computer Networks, Ustron, Poland, Jun 15-19 (Springer-Verlag, Berlin, 2010), pp. 236–244
7. P Kłosowski, Improving speech processing based on phonetics and phonology of Polish language. Przegląd Elektrotechniczny. **89**, 303–307 (2013)
8. J Izydorczyk, P Kłosowski, Acoustic properties of Polish vowels. Bull. Pol. Acad. Sci. Tech. Sci. **47**(1), 29–37 (1999)
9. J Izydorczyk, P Kłosowski, in *International Conference Programable Devices and Systems PDS2001 IFAC Workshop, Gliwice November 22nd - 23rd*. Base acoustic properties of Polish speech (IFAC, Gliwice, 2001), pp. 61–66
10. P Kłosowski, A Dustor, J Izydorczyk, J Kotas, Slimok J, in *Computer Networks, CN 2014. vol 431 of Communications in Computer and Information Science*, ed. by A Kwiecien, P Gaj, and P Stera. Speech recognition based on open source speech processing software. 21st International Science Conference on Computer Networks (CN), Brunow, Poland, Jun 23-27 (Springer-Verlag, Berlin, 2014), pp. 308–317
11. A Dustor, Kłosowski P, in *Computer Networks, CN 2013. vol 370 of Communications in Computer and Information Science*, ed. by A Kwiecien, P Gaj, and Stera P. Biometric voice identification based on Fuzzy Kernel Classifier. 20th International Conference on Computer Networks (CN), Lwowek Slaski, Poland, Jun 17-21 (Springer-Verlag, Berlin, 2013), pp. 456–465
12. A Dustor, P Kłosowski, J Izydorczyk, in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*. Speaker recognition system with good generalization properties. International Conference on Multimedia Computing and Systems (ICMCS), Marrakech, Morocco, Apr 14-16 (IEEE, USA, 2014), pp. 206–210
13. A Dustor, P Kłosowski, J Izydorczyk, in *Computer Networks, CN 2014. vol 431 of, Communications in Computer and Information Science*, ed. by A Kwiecien, P Gaj, and P Stera. Influence of Feature Dimensionality and

Model Complexity on Speaker Verification Performance. 21st International Science Conference on Computer Networks (CN), Brunow, Poland, Jun 23-27 (Springer-Verlag, Berlin, 2014), pp. 177–186
14. P Kłosowski, A Dustor, J Izydorczyk, in *Computer Networks, CN 2015. vol 522 of Communications in Computer and Information Science*, ed. by P Gaj, A Kwiecien, and P Stera. Speaker verification performance evaluation based on open source speech processing software and timit speech corpus. 22nd International Conference on Computer Networks (CN), Brunow, Poland, Jun 16-19 (Springer-Verlag, Berlin, 2015), pp. 400–409
15. A Dustor, P Kłosowski, J Izydorczyk, R Kopanski, in *Computer Networks, CN 2015. vol 522 of Communications in Computer and Information Science*, ed. by P Gaj, A Kwiecien, and P Stera. Influence of Corpus Size on Speaker Verification. 22nd International Conference on Computer Networks (CN), Brunow, Poland (Springer-Verlag, Berlin, 2015), pp. 242–249
16. P Kłosowski, Dustor A, in *Computer Networks, CN 2013. vol 370 of Communications in Computer and Information Science*, ed. by A Kwiecien, P Gaj, and P Stera. Automatic Speech Segmentation for Automatic Speech Translation. 20th International Conference on Computer Networks (CN), Lwowek Slaski, Poland, Jun 17-21 (Springer-Verlag, Berlin, 2013), pp. 466–475
17. F Jelinek, *Statistical Methods for Speech Recognition. Language, Speech, & Communication: A Bradford Book*. (MIT Press, USA, 1997)
18. S Furui, Recent progress in corpus-based spontaneous speech recognition. IEICE Trans. Inf. Syst. **E88D**, 366–375 (2005)
19. M Adda-Decker, Corpus for automatic speech recognition. Revue Francaise De Linguistique Appliquee. **12**, 71–84 (2007)
20. A Przepiórkowski, M Bańko, RL Górski, B Lewandowska-Tomaszczyk, *The National Corpus of Polish (in Polish: Narodowy Korpus Języka Polskiego)*. (Wydawnictwo Naukowe PWN, Warszawa, 2012)
21. A Przepiórkowski, RL Górski, B Lewandowska-Tomaszczyk, Łaziński M, in *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*. Towards the national corpus of Polish (Marrakech, ELRA, 2008)
22. RL Górski, B Lewandowska-Tomaszczyk, M Bańko, P Pęzik, M Łaziński, A Przepiórkowski, Practical applications of the National Corpus of Polish. Prace Filologiczne. **63**, 231–240 (2012)
23. J Hirschberg, CD Manning, Advances in natural language processing. Science. **349**, 261–266 (2015)
24. Association International Phonetic, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. A Regents publication*. (Cambridge University Press, UK, 1999)
25. R Sussex, P Cubberley, *The Slavic Languages. Cambridge Language Surveys*. (Cambridge University Press, UK, 2006)
26. J Wells, in *Handbook of Standards and Resources for Spoken Language Systems. vol Part IV, section B*, ed. by D Gibbon, R Moore, and R Winski. SAMPA computer readable phonetic alphabet (Mouton de Gruyter, Berlin and New York, 1997)
27. M Razavi, R Rasipuram, MM Doss, Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework. Speech Commun. **80**, 1–21 (2016)
28. RM Kaplan, M Kay, Regular models of phonological rule systems. Comput. Linguist. **20**, 331–378 (1994)
29. M Steffen-Batóg, The problem of automatic phonemic transcription of written Polish. Biuletyn Fonograficzny. **14**, 75–86 (1973)
30. M Steffen-Batóg, in *Polish: Automatyzacja transkrypcji fonematycznej tekstów polskich*. Automatic phonemic transcription of Polish texts (Wydawnictwo Naukowe PWN, Warszawa, 1975)
31. M Steffen-Batóg, Nowakowski P, in *Studia Phonetica Posnaniensia. Vol. 3*, ed. by M Steffen-Batóg, W Awedyk. An algorithm for phonetic transcription of orthographic texts in Polish (Wydawnictwo Naukowe UAM, Poznań, 1993)
32. W Jassem, *A phonemic transcription and syllable division rule engine*. (Onomastica-Copernicus Research Colloquium, Edinburgh, 1996)
33. P Kłosowski, in *Proceedings of 20th IEEE International Conference Signal Processing Algorithms, Architectures, Arrangements, and App.lications, September 21-23*. Algorithm and implementation of automatic phonemic transcription for polish (Poznan University of Technology, Poznań, 2016), pp. 298–303
34. M Wypych, in *Speech and Language Technology. Vol. 3*. Implementation of phonenic transcription alghorithm (in Polish: Implementacja algorytmu transkrypcji fonematycznej) (Polskie Towarzystwo Fonetyczne, Poznań, 1999)
35. G Demenko, M Wypych, E Baranowska, Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis. Speech Lang. Technol. **7**(17) (2003)
36. P Przybysz, W Kasprzak, in *2013 6th International Conferance on Human Systems Interactions (HSI)*, ed. by WA Paja, BM Wilamowski. The generation of letter-to-sound rules for grapheme-to-phoneme conversion. Conference on Human System Interaction. Gdansk Univ Technol; Univ Informat Technol & Management; IEEE Ind Elect Soc (Gdansk University of Techlology, Gdansk, 2013), pp. 292–297
37. D Skurzok, B Ziółko, Ziółko M, in *7th Language & Technology Conference, Poznań*. Ortfon2 - tool for orthographic to phonetic transcription (Adam Mickiewicz University in Poznan, Poznan, 2015)
38. D Koržinek, Ł Brocki, Marasek K, Polish grapheme-to-phoneme tool and service, CLARIN-PL digital repository (2016). http://hdl.handle.net/11321/295, (Online: 2016.08.01)
39. Wiktionary, Polish Language Dictionary (2015). https://pl.wiktionary.org/. Accessed 17 Feb 2017
40. W Jassem, *Podstawy fonetyki akustycznej (eng. Rudiments of acoustic phonetics)*. (PWN, Warszawa, 1973)
41. P Łobacz, W Jassem, Fonotaktyczna analiza mówionego tekstu polskiego (eng. Phonotactic analysis of spoken Polish texts). Biuletyn Polskiego Towarzystwa Ję. **32**, 179–195 (1974)
42. C Basztura, *Rozmawiac z komputerem (Eng. To speak with computers)*, (1992)
43. B Ziółko, J Gałka, S Manandhar, RC Wilson, M Ziółko, in *Human Language Technology: Challenges of the Information Society. Vol 5603 of Lecture Notes in Artificial Intelligence*, ed. by Z Vetulani, H Uszkoreit. Triphone Statistics for Polish Language. 3rd Language and Technology Conference 2007, Poznan, Poland, Oct 05-07, (2009), pp. 63–73
44. B Ziółko, J Gałka, M Ziółko, Polish phoneme statistics obtained on large set of written texts. Comput. Sci. (AGH). **10**, 97–106 (2009)
45. B Ziółko, Gałka J, in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*. Polish phones statistics (AGH Univesity of Science and Technology, Krakow, 2010), pp. 561–565
46. B Ziółko, P Zelasko, Skurzok D, in *2014 XXII Annual Pacific Voice Conference (PVC)*. Statistics of diphones and triphones presence on the word boundaries in the Polish language. App.lications to ASR. Annual Pacific Voice Conference, AGH; Pacific Voice Speech Fdn, 2014. 22nd Annual Pacific Voice Conference (PVC) (Krakow, AGH Univesity of Science and Technology, 2014)
47. D Lightfoot, *The development of language: Acquisition, change, and evolution*. (Wiley-Blackwell, Hoboken, 1999)
48. D Biber, S Conrad, R Repp.en, *Corpus linguistics: Investigating language structure and use*. (Cambridge University Press, Cambridge, 1998)
49. R Facchinetti, M Rissanen, *Corpus-based studies of diachronic English, vol. 31*. (Peter Lang, 2006)
50. GK Zipf, Human behavior and the principle of least effort. J. Clin. Psychol. **6**(3), 306–306 (1950)
51. Y Tambovtsev, C Martindale, Phoneme frequencies follow a yule distribution. SKASE J. Theor. Linguist. **4**(2) (2008)
52. ST Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions. Psychonimic Bull. Rev. **21**, 1112–1130 (2014)
53. A Corral, G Boleda, R Ferrer-i Cancho, Zipf's law for word frequencies: word forms versus lemmas in long texts. Plos ONE. **10**(7), e0129031 (2015). doi:10.1371/journal.pone.0129031
54. GU Yule, A mathematical theory of evolution, based on the conclusions of Dr.J. C. Willis, F.R.S. Phil. Trans. R. Soc. London B Biol Sci. **213**(402-410), 21–87 (1925)
55. S Dziadzio, A Nabożny, A Smywiński-Pohl, B Ziółko, in *Computer Science and Information Systems (FedCSIS) 2015 Federated Conference on*. Comparison of language models trained on written texts and speech transcripts in the context of automatic speech recognition (Lodz University of Technology, Lodz, 2015), pp. 193–197
56. S Takahashi, T Morimoto, in *2012 International Conference on Asian Language Processing (IALP 2012)*, ed. by D Xiong, E Castelli, M Dong, and PTN Yen. N-gram Language Model Based on Multi-Word Expressions in Web Documents for Speech Recognition and Closed-Captioning (Soochow University, China, 2012), pp. 225–228
57. A Hatami, A Akbari, B Nasersharif, in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*. N-gram Adaptation Using Dirichlet Class

Language Model Based on Part-of-Speech for Speech Recognition (Ferdowsi University of Mashhad, Mashhadm, 2013)
58. M Bahrani, H Sameti, N Hafezi, S Momtazi, in *New Frontiers in App.lied Artificial Intelligence, vol 5027 of Lecture Notes in Artificial Intelligence*, ed. by NT Nguyen, L Borzemski, A Grzech, and M Ali. New word clustering method for building n-gram language models in continuous speech recognition systems (Springer, Berlin, 2008), pp. 286–293
59. B Rapp, in *2008 International Multiconference on Computer Science and Information Technology (IMCSIT), Vols 1 and 2*, ed. by M Ganzha, M Paprzycki, and T PelechPilichowski. N-gram language models for Polish language. Basic concepts and applications in automatic speech recognition systems (IEEE Computer Society Press, Los Alamitos, 2008), pp. 295–298
60. D Klakow, P Jochen, Testing the correlation of word error rate and perplexity. Speech Commun. **38**(1–2), 19–28 (2002)
61. T Cover, J Thomas, *Wiley series in telecommunications: Elements of information theory*. (John Wiley and Sons, USA, 1991)
62. P Yu, FTB Seide, in *Interspeech*. A hybrid word/phoneme-based app.roach for improved vocabulary-independent search in spontaneous speech (Citeseer, Jeju Island, 2004)
63. V Chunwijitra, A Chotimongkol, C Wutiwiwatchai, A hybrid input-type recurrent neural network for lvcsr language modeling. EURASIP J. Audio Speech Music Process. **2016**(1), 15 (2016)
64. A Yazgan, M Saraclar, in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. vol 1 (IEEE, 2004), pp. I–745
65. M Larson, Sub-word-based language models for speech recognition: implications for spoken document retrieval. Whorkshop on Language Modeling and Information Retrieval (2001)
66. A Czardybon, O Hellwig, W Petersen, in *Advances in Natural Language Processing. vol 8686 of Lecture Notes in Artificial Intelligence*, ed. by A Przepiorkowski, M Ogrodniczuk. Statistical Analysis of the Interaction between Word Order and Definiteness in Polish. Polish Acad Sci, Inst Comp Sci, 2014. 9th International Conference on Natural Language Processing (NLP), Warsaw, Poland, Sep 17-19 (Polish Academy of Science, Institute of Conputer Scince, Warsaw, 2014), pp. 144–150
67. P Mandera, E Keuleers, Z Wodniecka, M Brysbaert, Subtlex-pl: subtitle-based word frequency estimates for Polish. Behav. Res. Methods. **47**, 471–483 (2015)
68. JR Bellegarda, Large vocabulary speech recognition with multispan statistical language models. IEEE Trans. Speech Audio Process. **8**, 76–84 (2000)
69. H Schwenk, Continuous space language models. Comput. Speech Lang. **21**(3), 492–518 (2007)
70. MAB Shaik, E-D AMousa, R Schlüter, H Ney, in *INTERSPEECH*. Hybrid language models using mixed types of sub-lexical units for open vocabulary German LVCSR (International Speech Communication Association (ISCA), Baixas, 2011), pp. 1441–1444

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com