A Rule-Based Grapheme-to-Phoneme Conversion System

Subjects: Computer Science, Interdisciplinary Applications

Contributor: Piotr Kłosowski

Natural language processing often requires grapheme-to-phoneme (G2P) conversion of an orthographic text. G2P converts strings of graphemes to corresponding sequences of phonetic transcription characters, directly from orthographic representations and it is crucial for many applications in various areas of speech and language processing.

Keywords: grapheme-to-phoneme conversion ; speech recognition ; language corpus

1. Problem Formulation

The process of converting graphemes to phonemes in orthographic text involves converting a string of orthographic characters into a corresponding string of phonetic transcription characters (representing phonemes or allophones) [1]. A 'grapheme' is any of the units of any writing system for any language, a term coined by analogy with the 'phoneme' of a spoken language [2]. Graphemes include alphabetic letters, typographic ligatures, numerical digits, punctuation marks, and other individual symbols of writing systems. Since the orthographic text is the only source of pronunciation information in the process of converting graphemes into phonemes, this process must be based on appropriate formal rules, depicting the correct pronunciation of orthographic strings in a given language [3].

Phonemes are usually written in specially designed alphabets. The most widely used alphabet is the International Phonetic Alphabet (IPA) [4]. For the Polish language, as with other Slavic languages, a special transcription system, called the Slavistic Phonetic Alphabet (SPA), is most frequently used [5]. The other very commonly used phonetic alphabet is the Speech Assessment Methods Phonetic Alphabet (SAMPA) [6]. SAMPA is a machine-readable phonetic alphabet, using 7-bit printable ASCII characters, based on the IPA. **Table 1** presents the phonemic inventory of Polish with examples, in the SPA, IPA, and SAMPA phonetic alphabets and corresponds to the set of phonemes.

Table 1. The set of Polish phonemes with examples, written in the SPA, IPA, and SAMPA phonetic alphabets.

	Phonetic Alphabet			**Example of**
No.	**Symbols**			**Occurrence**
	[SPA]	**[IPA]**	**[SAMPA]**	**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	**[i̯]**	**[j]**	**[j]**	**jajo**
8	**[i]**	**[i]**	**[i]**	**wici**
9	**[r]**	**[r]**	**[r]**	**rok**
10	**[s]**	**[s]**	**[s]**	**sok**
11	**[v]**	**[v]**	**[v]**	**wada**

	Phonetic Alphabet			Example of
No.	Symbols			Occurrence
	[SPA]	[IPA]	[SAMPA]	in Polish
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	[u̯]	[ɫ]	[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	[]	[ʝ]	[x]	higiena
29	[č]	[t͡ʃ]	[tS]	czarny
30	[ž]	[ʒ]	[Z]	każdy
31	[]	[]	[e]	ręka
32	[ḱ]	[c]	[k′]	kino
33	[]	[d͡ʑ]	[dz′]	dziedzic
34	[ʒ]	[d͡z]	[dz]	nadzy
35	[ź]	[ʑ]	[z′]	ziarno
36	[ǵ]	[ɟ]	[g′]	magiczny
37	[]	[d͡ʒ]	[dZ]	drożdże

Automatic grapheme-to-phoneme conversion is not a new problem. The first linguist who noted it, and tried to provide a solution for a particular language (Czech), was H. Kučera [7]. Research on solutions to the automatic grapheme-to-phoneme conversion problem have also been initiated for other languages [8][9][10].

In Poland, the first linguist who wrote about the possibility of phonetic interpretation of text by machines was W. Doroszewski in 1969 [11]. The largest contributions to solving the problem of automatic grapheme-to-phoneme conversion for Polish, were the publications of Maria Steffen-Batóg [12][13]. The first implementation of a grapheme-to-phoneme conversion algorithm for Polish, designed for the machine ODRA 1204, was made in 1971 by M. Warmus [14].

2. Grapheme-to-phoneme Conversion

- Automatic conversion of graphemes into phonemes in orthographic texts is not only a technical issue, consisting in developing appropriate algorithms for converting graphemes into phonemes, but also a serious linguistic problem. Only specialists in linguistics and phonetics of a given language are able to formulate appropriate rules for converting graphemes into phonemes for speech [15];

- An additional complication is that automatic conversion of graphemes to phonemes is a language-specific problem with different spelling and pronunciation conventions within the same language [16][17][18][19];

- Effective solutions for automatic grapheme-to-phoneme conversion in one language may not help solve the same problems for a different language. There is not only one language and technical problem of automatic conversion of graphemes to phonemes to be solved, but many different problems with different levels of difficulty that should be solved for each language separately [15];

- Automatic grapheme-to-phoneme conversion is widely used not only in speech synthesis, but also in speech recognition [20][21];

- A separate, but very important problem is the evaluation of grapheme-to-phoneme conversion processes [21][22]. Evaluation and validation of grapheme-to-phoneme conversion implementations is a laborious and time-consuming process. All problems registered for the G2P implementation discussed were positively resolved;

- The G2P implementation developed for this research is not the only one for Polish [3][23][24][25][26], however only one of the others is available for free use [24];

- The author of the paper analysed for comparison the only available application for the Polish language, named Transcriber [24]. The application was implemented in the C++ programming language. The implemented method uses a dictionary of 5018 words and 767 defined conversion rules. For comparison, the software was implemented in Python programming language, 975 conversion rules were implemented and the dictionary is very limited and plays only a supporting role. This means that TransFon has implemented 208 more transcription rules, which is over 27% more. The application failed to compile due to the lack of inclusion in the source code of the appropriate libraries that were used by the programmer to create the application. This made it impossible to evaluate the correctness of the application and seriously hindered the comparison with the software created by the researcher of the paper; However, based on the analysis of the application's source code, you can see that the principle of the application is also rule-based, but the author of the Transcriber application tried to refine and improve the application's performance by adding new words to the dictionary (exceptions). The author of the TransFon application, on the other hand, tried to add and supplement transcription rules in a similar way as is known in the literature. This is evidenced by the dictionary size used in both applications;

- The G2P system presented here could be used for Polish corpus development;

- The G2P implementation presented here did not exploit any similar pre-existing tools [27];

- It is worth noting that the solutions presented here for the development of language and speech corpora in Polish are not the only ones and publications on this subject are available [28][29];

- Of particular interest are the results presented in publications by Grażyna Demenko et al. [23][30][31][32][33][34][35].

The grapheme-to-phoneme conversion system developed and its ability to create phonemic language corpora for Polish open up further opportunities for research on improving automatic speech recognition in Polish. The plan for further research towards achieving this goal, using the phonemic language corpus developed, includes:

- Performing a better and more detailed statistical analysis of the Polish language based on the phonemic language corpus developed [36][37];

- Developing more efficient word-based and phoneme-based statistical language models for speech recognition applications in Polish [38][37];

- Application of deep learning methods to language modelling and speech recognition [39][40].

References

1. Lee, F. Automatic grapheme-to-phone translation of english. J. Acoust. Soc. Am. 1967, 41, 1594A.
2. Coulmas, F. The Blackwell's Encyclopedia of Writing Systems; Blackwells: Oxford, UK, 1996.

3. Przybysz, P.; Kasprzak, W. The generation of letter-to-sound rules for grapheme-to-phoneme conversion. In Proceedings of the 2013 6th International Conference on Human System Interactions (HSI), Sopot, Poland, 6–8 June 2013; pp. 292–297.

4. International Phonetic Association. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet; A Regents Publication; Cambridge University Press: Cambridge, UK, 1999.

5. Sussex, R.; Cubberley, P. The Slavic Languages. In Cambridge Language Surveys; Cambridge University Press: Cambridge, UK, 2006.

6. Wells, J. SAMPA computer readable phonetic alphabet. In Handbook of Standards and Resources for Spoken Language Systems, Vol. Part IV, Section B; Gibbon, D., Moore, R., Winski, R., Eds.; Mouton de Gruyter: Berlin, Germany; New York, NY, USA, 1997.

7. Kučera, H. Mechanical phonemic transcription and phoneme frequency count of czech. Int. J. Slav. Lingguistic Phon. 1963, 6, 36–50.

8. Bhimani, B.; Dolby, J. Acoustic phonetic transcription of written English. In Annual Report: Automatic Indexing and Abstracting; AIP Publishing: Palo Alto, CA, USA, 1966.

9. Pratt, B.; Silva, G. Phontrns: A Procedure which Uses a Computer for Transcribing French Text info Phonetic Symbols; Monash University: Melbourne, Australia, 1967.

10. Ungerhruer, G.; Kästner, W. Untersuchungen zur Transformation Deutcher Schirifttexte in Entsprechende Phonemtexte mit Hilfe Elektronischer Rechenmaschinen; Forschungsbericht; Institut für Phonetic und Kommunikationsforshung der Universität Bonn: Bonn, Germany, 1966.

11. Doroszewski, W. Speech and writing (in Polish: Mowa a pismo). Porad. Jęz. 1969, 4, 181–188.

12. Steffen-Batóg, M. The problem of automatic phonemic transcription of written Polish. Biul. Fonogr. 1973, XIV, 75–86.

13. Steffen-Batóg, M. Automatic Phonemic Transcription of Polish Texts (In Polish: Automatyzacja Transkrypcji Fonematycznej Tekstów Polskich); Wydawnictwo Naukowe PWN: Warszawa, Poland, 1975.

14. Warmus, M. Software implementation for ODRA 1204 of automatic phonemic transctiption of polish texts (in Polish: Program na maszynę ODRA 1204 dla automatycznej transkrypcji fonematycznej tekstów języka polskiego). In Zastosowanie Maszyn Matematycznych do Badań nad Językiem Naturalnym; Bolc, L., Ed.; Wydawnictwo Uniwersytetu Warszawskiego: Warszawa, Poland, 1973.

15. Auzina, I.; Pinnis, M.; Dargis, R. Comparison of Rule-based and Statistical Methods for Grapheme to Phoneme Modelling. In Frontiers in Artificial Intelligence and Applications, Proceedings of the Human Language Technologies—The Baltic Perspective, Baltic HLT 2014, Kaunas, Lithuania, 26–27 September 2014; Utka, A., Grigonyte, G., Kapociute Dzikiene, J., Vaicenoniene, J., Eds.; Vytautas Magnus University ViaConventus: Vilnius, Lithuania, 2014; Volume 268, pp. 57–60.

16. Schlippe, T.; Ochs, S.; Schultz, T. Grapheme-to-phoneme model generation for indo-european languages. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4801–4804.

17. Kosaner, O.; Birant, C.C.; Aktas, O. Improving Turkish language training materials: Grapheme-to-phoneme conversion for adding phonemic transcription into dictionary entries and course books. In Procedia Social and Behavioral Sciences, Proceedings of the 13th International Educational Technology Conference, Lisbon, Portugal, 30 October–1 November 2014; Isman, A., Siraj, S., Kiyici, M., Eds.; Volume 103, pp. 473–484.

18. Lee, J.; Kim, B.; Lee, G.G. Hybrid Approach to Grapheme to Phoneme Conversion for Korean. In Proceedings of the InterSpeech 2009: 10th Annual Conference of the International Speech Communication Association 2009, Brighton, UK, 6–10 September 2009; Volume 1–5, pp. 1299–1302.

19. de Jesus Aguiar Pontes, J.; Furui, S. Predicting the phonetic realizations of word-final consonants in context—A challenge for French grapheme-to-phoneme converters. Speech Commun. 2010, 52, 847–862.

20. Bagshaw, P. Phonemic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression. Comput. Speech Lang. 1998, 12, 119–142.

21. Jouvet, D.; Fohr, D.; Illina, I. Evaluating grapheme-to-phoneme converters in automatic speech recognition context. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4821–4824.

22. Schraagen, M.; Bloothooft, G. A qualitative evaluation of phoneme-to-phoneme technology. In Proceedings of the 12th Annual Conference of the International-Speech-Communication-Association 2011 (Interspeech 2011), Florence, Italy, 27–31 August 2011; Volume 1–5, pp. 2332–2335.

23. Demenko, G.; Wypych, M.; Baranowska, E. Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis. Speech Lang. Technol. 2003, 7, 79–97.

24. Koržinek, D.; Brocki, Ł.; Marasek, K. Polish Grapheme-to-Phoneme Tool and Service, CLARIN-PL Digital Repository (2016). Available online: https://clarin-pl.eu/dspace/handle/11321/295 (accessed on 10 January 2022).

25. Skurzok, D.; Ziółko, B.; Ziółko, M. Ortfon2—Tool for orthographic to phonetic transcription. In Proceedings of the 7th Language & Technology Conference, Poznań, Poland, 27–29 November 2015.

26. Wypych, M. Implementation of phonenic transcription alghorithm (in Polish: Implementacja algorytmu transkrypcji fonematycznej). In Speech and Language Technology; Polskie Towarzystwo Fonetyczne: Poznań, Poland, 1999; Volume 3.

27. Kłosowski, P. Algorithm and implementation of automatic phonemic transcription for Polish. Proceedings of 20th IEEE International Conference Signal Processing Algorithms, Architectures, Arrangements, and Applications, Poznań, Poland, 21–23 September 2016; pp. 298–303.

28. Żelasko, P.; Ziółko, B.; Jadczyk, T.; Skurzok, D. AGH corpus of Polish speech. Lang. Resour. Eval. 2016, 50, 585–601.

29. Ziółko, B.; Jadczyk, T.; Skurzok, D.; Żelasko, P.; Gałka, J.; Pędzimąż, T.; Gawlik, I.; Pałka, S. SARMATA 2.0 automatic Polish language speech recognition system. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; Interspeech: Dresden, Germany, 2015.

30. Cylwik, N.; Wagner, A.; Demenko, G. The euronounce corpus of non-native polish for asr-based pronunciation tutoring system. In Proceedings of the 2nd ISCA Workshop of Speech and Language Technology in Education, Warwickshire, UK, 3–5 September 2009.

31. Demenko, G. Korpusowe Badania JęZyka MóWionego; Akademicka Oficyna Wydawnicza EXIT: Warszawa, Polish, 2015; ISBN 9788378370437.

32. Demenko, G.; Bachan, J.; Wagner, A.; Wyroślak, P. Speech corpus creation for automatic analysis of phonetic convergence. In Studientexte zur Sprachkommunikation, Proceedings of 27th Conference on Electronic Speech Signal Processing (ESSV), Leipzig, Germany, 2–4 March 2016; Oliver, J., Ed.; Hochschule für Telekommunikation Leipzig (HfTL): Leipzig, Germany, 2016; pp. 183–190.

33. Demenko, G.; Grocholewski, S.; Klessa, K.; Rau, Z. Polish language resources for speech technology: Jurisdic lvcsr corpora. In Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of the 4th Language & Technology Conference, Poznań, Poland, 6–8 November 2009; Zygmunt, V., Ed.; Adam Mickiewicz University: Poznań, Poland, 2009; pp. 165–169.

34. Demenko, G.; Klessa, K.; Szymański, M.; Breuer, S.; Hess, W. Polish unit selection speech synthesis with boss: Extensions and speech corpora. Int. J. Speech Technol. 2010, 13, 85–99.

35. Demenko, G.; Szymański, M.; Cecko, R.; Lange, M.; Klessa, K.; Owsianny, M. Development of large vocabulary continuous speech recognition using phonetically structured speech corpus. In Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, China, 17–21 August 2011; pp. 568–571.

36. Kłosowski, P. Statistical analysis of Polish language corpus for speech recognition application. In Proceedings of the 20th IEEE International Conference Signal Processing Algorithms, Architectures, Arrangements, and Applications, Poznań, Poland, 21–23 September 2016; pp. 304–309.

37. Kłosowski, P. Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based Polish language modelling. EURASIP J. Audio Speech Music Process. 2017, 2017, 5.

38. Kłosowski, P. Polish language modelling for speech recognition application. In Proceedings of the 21th IEEE International Conference Signal Processing Algorithms, Architectures, Arrangements, and Applications, Poznan, Poland, 20–22 September 2017; pp. 313–318.

39. Kłosowski, P. Deep learning for natural language processing and language modelling. In Proceedings of the 22th IEEE International Conference Signal Processing Algorithms, Architectures, Arrangements, and Applications, Poznan, Poland, 19–21 September 2018; pp. 223–228.

40. Kłosowski, P. Polish language modelling based on deep learning methods and techniques. In Proceedings of the 23th IEEE International Conference Signal Processing Algorithms, Architectures, Arrangements, and Applications, Poznan, Poland, 18–20 September 2019; pp. 223–228.