Similar place avoidance in Slavic and other languages: Appendix

Aleš Bičan

Last modified: 9 September 2025

This is one part of the appendix for the paper published in the *Journal of Slavic Linguistics* 32(2).

The other paper is statistical calculations of the distribution of CVC sequences in 100 major languages and alternative dataset for nine Slavic languages listed below. They are found in a separate file here:
https://ojs.ung.si/index.php/JSL/article/view/469/344

Additional resources and updates can be found here: https://osf.io/8kdqb/

A general description of the procedure behind the collection and processing of the languages is found in Section 2: Data and methodology of the paper.

The lexical sources are of three kinds:

a) *Phonetic/phonological databases*
   Most of the sources are lexical databases providing phonetic or phonological transcription. In general, I simply extracted the transcriptions and modified the symbols to fit the software used for frequency calculations (digraphs were converted to single characters). If the database provides information on the origin of words, the word lists were limited to native-origin vocabulary. This is specified in the notes below for individual languages.
b) *Dictionaries with pronunciation*
   Some dictionaries provide pronunciation, which was extracted therefrom. Sometimes, pronunciation is provided only for some words. The others were converted (see next). This fact is specified in the list below for the relevant languages.
c) *Dictionaries, word lists, and dictionary indices without pronunciation*
   When a phonetic/phonological database was not available for a selected language, I relied on dictionaries and other similar vocabulary lists. Such sources are *asterisked below. Preference was given to languages with a reasonably reliable phonological spelling from which pronunciation could be determined. I used simple grapheme-to-phoneme substitutions.

Variants
Sometimes the sources provide several variants (pronunciations) for some word. Only the first listed variant was retained; the others were deleted.

Errors
Some sources contain obvious errors such as nonsensical or clearly wrong pronunciations / phonological forms or false words. When spotted, these errors were deleted or ignored in statical calculations. Otherwise, I regarded the sources as reliable without checking their content.

A) Slavic languages – Primary data

Proto-Slavic (PSl), Old Church Slavonic (OCS), Bulgarian (Bul), Macedonian (Mac), Serbo-Croatian (SCr), Slovenian (Slo), Slovak (Slk), Old Czech (OCz), Modern Czech (MCz), Upper Sorbian (USo), Lower Sorbian (LSo), Polabian (Plb), Pomeranian (Pom), Polish (Pol), Belarusian (Bel), Ukrainian (Ukr), Old Russian (ORu), Modern Russian (MRu)

*Havlová, Eva, Adolf Erhart & Ilona Janyšková (eds.). (1989–2022) *Etymologický slovník jazyka staroslověnského 1–21* [Etymological dictionary of the Old Church Slavonic language]. Prague: Karolinum, Brno: Tribun.

The word lists were from the indices for the particular Slavic languages from the 20th volume of the dictionary. Pronunciation was generated from the spelling of these languages, following standard pronunciation rules. The words for Modern Czech and Slovak were checked by native speakers.

B) Slavic languages – Additional data

Belarusian

WikiPron dictionaries. https://github.com/CUNY-CL/wikipron/tree/master/data/scrape (accessed 28 March 2024). https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/bel_cyrl_narrow.tsv

Lee, Jackson L., Lucas F. E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy and Kyle Gorman (2020). "Massively multilingual pronunciation mining with WikiPron". In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the 12th Language Resources and Evaluation Conference*, 4223–4228. European Language Resource Association. https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/bel_cyrl_narrow.tsv

An IPA-transcribed word list.

Bulgarian

WikiPron dictionaries. See Belarusian for the reference. https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/bul_cyrl_narrow.tsv

IPA-transcribed words.

Czech, Modern

Bičan, Aleš. 2020. *Phonological Corpus of Czech*. Version 2020, subcorpora: Database of native words and Database of Loanwords. https://www.phil.muni.cz/phoncorp (accessed 23 February 2024).

The Corpus uses its own transcription, convertible to IPA.

Macedonian

WikiPron dictionaries. See Belarusian for the reference.

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/mkd_cyrl_narrow.tsv

An IPA-transcribed word list.

Polish

WikiPron dictionaries. See Belarusian for the reference.

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/pol_latn_broad.tsv

An IPA-transcribed word list.

Russian, Modern

*Vasmer, Max. 1953–1958. *Russisches etymologisches Wörterbuch 1–3*. Heidelberg: Winter. XML version available here: https://github.com/tamila-krashtan/vasmer (accessed 27 September 2023).

The words were extracted from the XML file. Pronunciation generated by grapheme-to-phoneme conversion.

Serbo-Croatian

WikiPron dictionaries. See Belarusian for the reference.

The broad filtered database was used:

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/hbs_latn_broad_filtered.tsv

An IPA-transcribed word list.

Slovenian

*Furlan, Metka & Miha Sušnik. 2017–2023. *Novi etimološki slovar slovenskega jezika* [New etymological dictionary of the Slovenian languages]. Inštitut za slovenski jezik Frana Ramovša. https://fran.si (accessed 20 April 2021).

*Snoj, Marko. 2015. *Slovenski etimološki slovar* [Slovenian etymological dictionary]. 3rd edition. Inštitut za slovenski jezik Frana Ramovša. https://fran.si (accessed 20 April 2021).

The words are headwords from the online version of these two dictionaries. Pronunciation generated by grapheme-to-phoneme conversion.

Ukrainian

WikiPron dictionaries. See Belarusian for the reference.

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/ukr_cyrl_narrow.tsv

An IPA-transcribed word list.

C) Non-Slavic languages

Afrikaans

Lwazi Afrikaans pronunciation dictionary. https://taalmaterialen.ivdnt.org/download/tstc-lwazi-afrikaans-pronunciation-dictionary/ (accessed 23 November 2023).

Davel, Marelie and Olga Martirosian. (2009) "Pronunciation dictionary development in resource-scarce environments". In *10th annual conference of the International Speech Communication Association (INTERSPEECH)*, 2851–2854. Brighton.

Available for free after registration. SAMPA-transcribed pronunciation.

Albanian

*Newmark, Leonard. (2005) *Albanian-English Dictionary.* http://www.seelrc.org:8080/albdict/ (accessed 16 November 2023).

Dictionary headwords were used. Pronunciation generated by grapheme-to-phoneme conversion.

Azerbaijani

WikiPron dictionaries. See Belarusian for the reference.

The narrow filtered database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/aze_latn_narrow_filtered.tsv (accessed 5 April 2024).

IPA-transcribed words.

Armenian

Armenian CV dictionary v2.0.0.
https://mfa-models.readthedocs.io/en/latest/dictionary/Armenian/Armenian%20CV%20dictionary%20v2_0_0.html#Armenian%20CV%20dictionary%20v2_0_0 (accessed 7 December 2023).

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

IPA-transcribed words.

Basque

Basque CV dictionary v2.0.0.
https://mfa-models.readthedocs.io/en/latest/dictionary/Basque/Basque%20CV%20dictionary%20v2_0_0.html#Basque%20CV%20dictionary%20v2_0_0 (accessed 10 April 2024)

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

IPA-transcribed words.

Bengali

Bhattacarya, Subhasha. (2003) *Samsada Bangala uccarana abhidhana*. Kalikata: Sahitya Samsad. Electronic version from Digital dictionaries of South Asia, https://dsal.uchicago.edu/dictionaries/bhattacharya/ (accessed 6 December 2023).

The dictionary provides IPA-transcribed pronunciation.

Breton

NorthEuraLex. http://northeuralex.org (accessed 24 July 2024).
http://northeuralex.org/languages/bre

Dellert, Johannes, Daneyko, Thora, Münch, Alla et al. (2019) *Lang Resources & Evaluation*.
https://doi.org/10.1007/s10579-019-09480-6 (version 0.9).

An IPA-transcribed word list.

Burmese

WikiPron dictionaries. See Belarusian for the reference.

The broad filtered database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/mya_mymr_broad_filtered.tsv

An IPA-transcribed word list.

Catalan

WikiPron dictionaries. See Belarusian for the reference.

The narrow database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/cat_latn_narrow.tsv

An IPA-transcribed word list.

Cebuano

WikiPron dictionaries. See Belarusian for the reference.

The narrow database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/ceb_latn_narrow.tsv

An IPA-transcribed word list.

Chinese, Mandarin

Chinese lexical database. https://www.chineselexicaldatabase.com (accessed 15 April 2024).

Sun, Ching. C., Hendrix, Peter, Ma, Jianqiang and Rolf H. Baayen (2018). “Chinese Lexical Database (CLD): A large-scale lexical database for simplified Mandarin Chinese”. *Behavior Research Methods* 50: 2606–2629.

An IPA-transcribed word list.

Chong

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 23 July 2024).

Baradat, R. (1941). *Les dialectes des tribus sâmrê*. Paris: Manuscrit de l’Ecole Française d’Extrême-Orient. Unpublished manuscript.

The dictionary provides IPA-transcribed pronunciation.

Chukchi

NorthEuraLex. See Breton for the reference.
http://northeuralex.org/languages/ckt

An IPA-transcribed word list.

Danish

NST pronunciation lexicon for Danish. (2003) Nordic language technology AS. https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-26/#resource-common-info (accessed 7 November 2023).

The database uses its own transcription, convertible to IPA.

Dutch

Baayen, R. Harard, Richard Piepenbrock and Leon Gulikers. (1995–1996) *CELEX2*. https://catalog.ldc.upenn.edu/LDC96L14.

The data taken from the WebCelex interface: http://celex.mpi.nl (accessed 19 October 2023).

SAMPA-transcribed pronunciation.

English, Old

WikiPron dictionaries. See Belarusian for the reference.

The broad database was used:

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/ang_latn_broad.tsv

An IPA-transcribed word list.

English, Middle

WikiPron dictionaries. See Belarusian for the reference.

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/enm_latn_broad.tsv

An IPA-transcribed word list.

English, Modern

CELEX2. See Dutch for the reference.

A SAMPA-transcribed word list.

Faroese

Ravrur BLARK 1.0. https://www.maltokni.fo/en/resource/ravnur-blark-1-0/ (accessed 24 November 2023).

Simonsen, Annika, Sandra Saxov Lamhauge, Iben Nyholm Debess and Peter Juel Henrichsen. (2022) “Creating a basic language resource kit for Faroese”. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk and Stelios Piperidis, eds., *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, 4637–4643. European Language Resource Association.

A SAMPA-transcribed word list.

Finnish

*Every Finnish Word. https://github.com/hugovk/everyfinnishword (accessed 13 March 2024). Originally from KOTUS wordlist (Institute for the Languages of Finland / Kotimaisten kielten keskus, https://www.kotus.fi/aineistot/sana-aineistot/nykysuomen_sanalista).

Pronunciation generated by grapheme-to-phoneme conversion.

French

Lexique. http://www.lexique.org (accessed 23 October 2023).

New, Boris, Pallier, Christophe, Brysbaert, Marc and Ludovic Ferrand. (2004) “Lexique 2: A New French Lexical Database”. *Behavior Research Methods, Instruments, & Computers* 36 (3): 516–524.

A SAMPA-transcribed word list.

Frisian, Old

*Boutkan, Dirk & Sjoerd Michiel Siebinga. (2005) *Old Frisian Etymological Dictionary*. Leiden, Boston: Brill.

The words were extracted from the online index (restricted access): https://dictionaries.brillonline.com/search#dictionary=frisian&id=fr0001

Pronunciation generated by grapheme-to-phoneme conversion.

Galician

WikiPron dictionaries. See Belarusian for the reference.

The narrow database was used: https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/glg_latn_narrow.tsv

An IPA-transcribed word list.

Georgian

Georgian CV dictionary v2.0.0. https://mfa-models.readthedocs.io/en/latest/dictionary/Georgian/Georgian%20CV%20dictionary%20v2_0_0.html (accessed 9 April 2024).

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

An IPA-transcribed word list.

German

CELEX2. See Dutch for the reference.

A SAMPA-transcribed word list.

Gothic

*Glossary from Joseph Wright’s *Grammar of the Gothic language*. http://lexicon.ff.cuni.cz/tmp/goth_wright_glossary.html (accessed 9 November 2023).

Wright, Joseph. 1910. *Grammar of the Gothic language*. Oxford: Clarendon Press.

Pronunciation generated by grapheme-to-phoneme conversion.

Greek, Ancient

*Liddel, Henry G. and Robert Scott. (1887) *An intermediate Greek-English lexicon*. New York, Cincinnati, Chicago: American book company. XML version. https://github.com/blinskey/middle-liddell/tree/master (accessed 30 November 2023).

The words were extracted from the XML file. Pronunciation generated by grapheme-to-phoneme conversion.

Greek, Modern

GreekLex 2. https://psychology.nottingham.ac.uk/greeklex/ (accessed 31 October 2023).

Kyparissiadis, Antonios, Walter J. B. van Heuven, Nicola J. Pitchford and Timothy Ledgeway. (2017) "GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological and stress information". *PLoS ONE* 12(2). e0172493.

A SAMPA-transcribed word list.

Greenlandic

*Oqaasileriffik / The Language Secretariat of Greenland. (2008) *Oqaasersiorfik Dictionary.* https://ordbog.gl/2008-kal/ (accessed 25 July 2024).

Pronunciation generated by grapheme-to-phoneme conversion.

Hindi

Hindi CV dictionary v2.0.0. https://github.com/MontrealCorpusTools/mfa-models/releases/tag/dictionary-hindi_cv-v2.0.0#dictionary-details (accessed 8 December 2023).

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

An IPA-transcribed word list.

Hittite

*HFR-Team. (2022) *Das Corpus der hethitischen Festrituale*. https://www.hethport.uni-wuerzburg.de/HFR/glossar_liste.php?gl=hit (accessed 12 December 2023).

Pronunciation generated by grapheme-to-phoneme conversion.

Hungarian

WikiPron dictionaries. See Belarusian for the reference.

The narrow filtered database was used: https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/hun_latn_narrow_filtered.tsv

An IPA-transcribed word list.

Icelandic

Nikulásdóttir, Anna Björk, Bjarki Ármannsson, Bryndís Bergþórsdóttir and Eiríkur Rögnvaldsson. (2023) *Icelandic pronunciation dictionary for language technology*, https://github.com/grammatek/iceprondict/tree/master (accessed 7 October 2023).

Nikulásdóttir, Anna Björk, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson and Steinþór Steingrímsson. (2020) "Language technology programme for Icelandic 2019–2023". In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the 12th international conference on language resources and evaluation (LREC 2020)*, 3414–3422. European Language Resource Association.

ice_pron_dict_standard_clear_IPA.csv file was used. IPA-transcribed pronunciation.

Irish

WikiPron dictionaries. See Belarusian for the reference.

The broad database was used:

https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/gle_latn_broad.tsv

An IPA-transcribed word list.

Italian

PhonItalia, http://phonitalia.org/ (currently unavailable; accessed 9 December 2021).

Goslin, Jeremy, Claudia Galluzzi and Cristina Romani. (2013) PhonItalia: A phonological lexicon for Italian. *Behavior research methods* 46. 872–886.

A SAMPA-transcribed word list. Word forms were used.

Japhug

Buret, Céline, Guillaume, Séverine, Jacques, Guillaume, Lahaussois, Aimée and Alexis Michaud. (not dated) *HimalCo project*. http://himalco.huma-num.fr (accessed 27 March 2024).

IPA-transcribed dictionary headwords were used.

Kalmyk

NorthEuraLex. See Breton for the reference.

http://northeuralex.org/languages/xal

An IPA-transcribed word list.

Khanty, Surgut

Kiss, Katalin É. et al. (not dated). *Languages under the Influence. Uralic syntax changing in an asymmetrical contact situation*. https://archive.nytud.hu/depts/tlp/uralic/dbases.html (accessed 27 March 2024).

IPA-transcribed texts were used.

Khasi

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 23 July 2024).

Singh, U. Nissor. (1906) *Khasi / English Dictionary.* Shillong: Eastern Bengal and Assam Secretariat Press.

The dictionary provides IPA-transcribed pronunciation.

Khmer, Surin

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 26 September 2024).

Chantrupanth, Dhanan. (1978) *Khmer (Surin) – Thai – English Dictionary.* Bangkok: Chulalongkorn University.

The dictionary provides IPA-transcribed pronunciation.

Khmu

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 24 July 2024).

Suwilai Premsrirat. (2002) *Thesaurus of Khmu Dialects in Southeast Asia*. Special Publication Num. 1, Vol. 1. Bangkok: Mon-Khmer Studies and Mahidol University.

The dictionary provides IPA-transcribed pronunciation.

Kukatja

Chirila database. https://github.com/YaleDHLab/chirila (accessed 10 April 2024).

Bowern, Claire. (2016) "Chirila: Contemporary and historical resources for the indigenous languages of Australia". *Language documentation & conservation* 10: 1–44.

The words were extracted (filtered) from the database. An IPA-transcribed word list.

Korean

Holliday, Jeffrey J., Rory Turnbull and Julien Eychenne (2016) *K-SPAN (Korean Surface Phones and Neighborhoods*, DataverseNO, V2, https://doi.org/10.18710/TWM79F (accessed 15 April 2024).

The database uses its own transcription, convertible to IPA.

Koryak

Koryak-Chukchi topical dictionary. In Kurebito, Tokusu and Yukari Nagayama. (not dated) *The languages and cultures of Northeast Eurasia*, https://hokuto-asia.aa-ken.jp/r/kydic/show/1.html (accessed 9 July 2024).

The IPA-transcribed words were extracted from particular entries.

Kurmanji

Kurmanji CV dictionary v2.0.0, https://mfa-models.readthedocs.io/en/latest/dictionary/Kurmanji/Kurmanji%20CV%20dictionary%20v2_0_0.html#Kurmanji%20CV%20dictionary%20v2_0_0 (accessed 10 April 2024).

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

An IPA-transcribed word list.

Lao

WikiPron dictionaries. See Belarusian for the reference. https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/lao_laoo_narrow.tsv

An IPA-transcribed word list.

Latin

Nelson, Lynn H. 1997. ORB Latin word list. https://the-orb.arlima.net/latwords.html (accessed 3 November 2023).

Pronunciation generated by grapheme-to-phoneme conversion.

Latvian

Jērāne, Santa, Ieva Kuplā, Gunta Lejniece, Ilga Migla, Laimdota Oldere, Ārija Ozola, Vija Požarnova, Anitra Roze, Imants Šmidebergs, Dorina Šnē, Māra Šnē, Ieva Zuicena, Lauma Pretkalniņa, Ieva Auziņa, Santa Briede, Imants Šmidebergs & Agris Timuška. 2023. *Dictionary of contemporary Latvian language* (MLVV) (2023-09-21), CLARIN-LV digital library at IMCS. University of Latvia, http://hdl.handle.net/20.500.12574/88 (accessed 7 November 2023).

The words were extracted from the XML version of the dictionary. Only the words marked as "lemma" were only, i.e., derivates were excluded. The dictionary provides pronunciation for many words. The pronunciation of the words with no explicit pronunciation was derived from spelling.

Lezgian

NorthEuraLex. See Breton for the reference.
http://northeuralex.org/languages/lez

An IPA-transcribed word list.

Lithuanian

*Lithuanian-Latvian dictionary (2007) based on *Lithuanian-Latvian dictionary* (1995) by Jons Balkevičs, Laimute Balode, Apolonija Bojate, Valters Subatnieks, ed. by Alberts Sarkanis. https://www.letonika.lv/groups/default.aspx?g=2&r=10631062 (accessed 13 November 2023).

The words were extracted from the headwords. Pronunciation generated by grapheme-to-phoneme conversion.

Malayalam

WikiPron dictionaries. See Belarusian for the reference.

The broad database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/mal_mlym_broad.tsv

An IPA-transcribed word list.

Maltese

WikiPron dictionaries. See Belarusian for the reference.

The broad filtered database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/mlt_latn_broad_filtered.tsv

An IPA-transcribed word list.

Mazahua

*Reyna, Rufino Benítez. (2017) *Vocabulario práctico bilingüe Mazahua-Español*. México: Departamento de Medios Digitalas. https://www.gob.mx/cms/uploads/attachment/file/192866/cdi-vocabulario-mazahua-rufino-benitez-reyna-web.pdf (accessed 25 July 2024).

Headwords were extracted from the dictionary. Pronunciation generated by grapheme-to-phoneme conversion.

Nancowry

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 22 July 2024).

Man, Edward Horace. (1889) *A dictionary of the central Nicobarese language (English-Nicobarese and Nicobarese-English)*. London: Witt. Allan and Co.

The dictionary provides IPA-transcribed pronunciation.

Norwegian

NST pronunciation lexicon for Norwegian Bokmål (2023). Nordic language technology AS. https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-26/#resource-common-info (no longer available; accessed 7 November 2023).

The database uses its own transcription, convertible to IPA.

Nyah Kur

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 24 July 2024).

Thongkum, Theraphan L. (1984) *Nyah Kur (Chao bon)-Thai-English Dictionary.* Monic Language Studies, Vol. 2. Bangkok: Chulalongkorn University Printing House, Bangkok.

The dictionary provides IPA-transcribed pronunciation.

Ossetian

NorthEuraLex. See Breton for the reference. http://northeuralex.org/languages/oss

An IPA-transcribed word list.

Otomi

Lexibank. https://lexibank.clld.org (accessed 21 April 2024). https://lexibank.clld.org/languages/wold-Otomi

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch and Russell D. Gray. (2022) “Lexibank, a public repository of standardized wordlists with computed phonological and lexical features”. *Scientific data* 9(316): 1–16.

An IPA-transcribed word list.

Palaung

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 23 July 2024).

Milne, Leslie. (1931) “Palaung and Palê dialects”. In *A dictionary of English-Palaung and Palaung-English*. Rangoon, Superintendent, Government Printing and Stationary.

The dictionary provides IPA-transcribed pronunciation.

Passamaquoddy-Maliseet

*Francis, David A. and Robert M. Leavitt. (2008) *A Passamaquoddy-Maliseet Dictionary.* Orono, Maine: The University of Maine Press. https://pmportal.org/browse-dictionary (accessed 14 April 2024).

Headwords were extracted from the dictionary. Pronunciation generated by grapheme-to-phoneme conversion.

Pashto, Northern

NorthEuraLex. See Breton for the reference.
http://northeuralex.org/languages/pbu

An IPA-transcribed word list.

Persian

*Steingass, Francis J. (1892) *A Comprehensive Persian-English Dictionary, including the Arabic Words and Phrases to be Met with in Persian Literature*. London: Routledge & K. Paul. Electronic version from Digital dictionaries of South Asia, https://dsal.uchicago.edu/dictionaries/steingass/ (accessed 19 November 2023).

The words were extracted from the dictionary headwords (in Latin alphabet). The words classified as loanwords were excluded. Pronunciation generated by grapheme-to-phoneme conversion.

Plautdietsch

Koehler, Loren. *Plautdietsch lexicon*. http://plautdietsch.22web.org/home/index.htm?i=1 (accessed 24 November 2023).

The dictionary provides IPA-transcribed pronunciation.

Portuguese, European

PorLex, https://hdl.handle.net/21.11129/0000-000D-F933-1 (accessed 10 October 2023).

Gomes, Inês and São Luís Castro. (2003) “Porlex, a lexical database in European Portuguese”. *Psychologica* 32: 91–108.

The database uses its own transcription, convertible to IPA.

Proto-Celtic

*Matasović, Ranka. (2009) *Etymological Dictionary of Proto-Celtic*. Leiden, Boston: Brill.

The words were extracted from the online index (restricted access):
https://dictionaries.brillonline.com/search#dictionary=proto_celtic&id=pc0001

Pronunciation generated by grapheme-to-phoneme conversion.

Proto-Germanic

*Kroonen, Guus. (2013) *Etymological Dictionary of Proto-Germanic*. Leiden, Boston: Brill.

The words were extracted from the online index (restricted access):
https://dictionaries.brillonline.com/search#dictionary=proto_germanic&id=pg0001

Pronunciation generated by grapheme-to-phoneme conversion.

Proto-Indo-European

*Havlová, Eva, Adolf Erhart & Ilona Janyšková (eds.). (1989–2022) *Etymologický slovník jazyka staroslověnského 1–21* [Etymological dictionary of the Old Church Slavonic language]. Prague: Karolinum, Brno: Tribun.

PIE roots listed in the indices were used. Pronunciation generated by grapheme-to-phoneme conversion.

Punjabi

Punjabi CV dictionary v2.0.0. https://github.com/MontrealCorpusTools/mfa-models/releases/tag/dictionary-punjabi_cv-v2.0.0#dictionary-details (accessed 8 December 2023).

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

An IPA-transcribed word list.

Romanian

MaRePhoR, An open access machine-readable phonetic dictionary for Romanian. https://speech.utcluj.ro/marephor/ (accessed 7 December 2023).

Toma, Ştefan-Adrian, Adriana Stan, Mihai-Lică Pura and Traian Bârsan. (2017) “MaRePhoR – An open access machine-readable phonetic dictionary for Romanian”. In *Proceedings of the 9th conference on speech technology and human-computer dialogue*. Bucharest. https://adrianastan.com/papers/2017_SPED_Marephor.pdf.

A SAMPA-transcribed word list.

Semai

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 16 July 2024).

Means, Natalie and Paul Means. (1987) *Senoi-English, English-Senoi Dictionary.* The Joint Center on Modern East Asia, University of Toronto and York University.

The dictionary provides IPA-transcribed pronunciation.

Sanskrit

*Monier-Williams, Monier. (1899) *A Sanskrit-English dictionary: Etymologically and philologically arranged with special reference to Cognate Indo-European languages*. Oxford: The Clarendon Press. XML digital version. https://www.sanskrit-lexicon.uni-koeln.de/scans/MWScan/2020/web/index.php (accessed 9 November 2023).

The words were extracted from the XML file. Pronunciation generated by grapheme-to-phoneme conversion.

Spanish

CLEARPOND database. https://clearpond.northwestern.edu (accessed 23 October 2023).

Marian, Viorica, James Bartolotti, Sarah Chabal and Anthony Shook. (2012) “CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities”. *PLoS ONE* 7(8). e43230.

A SAMPA-transcribed word list.

Swedish

NST pronunciation lexicon for Swedish. (2003) Nordic language technology AS. https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-26/#resource-common-info (accessed 7 November 2023).

The database uses its own transcription, convertible to IPA. Only the items marked as SWE (native words) and LEX (lexemes) were used, that is, loanwords and derivates were excluded.

Tagalog

WikiPron dictionaries. See Belarusian for the reference.

The narrow database was used:
https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/tgl_latn_narrow.tsv

An IPA-transcribed word list.

Tampuan

SEAlang Mon-Khmer Languages Project. http://www.sealang.net/monkhmer/database/ (accessed 23 July 2024).

Crowley, James D. (2004) *Tampuan Dictionary.* Unpublished manuscript (since published).

The dictionary provides IPA-transcribed pronunciation.

Tatar

Tatar CV dictionary v2.0.0.
https://mfa-models.readthedocs.io/en/latest/dictionary/Tatar/Tatar%20CV%20dictionary%20v2_0_0.html#Tatar%20CV%20dictionary%20v2_0_0 (accessed 10 April 2024).

Ahn, Emily and Eleanor Chodroff. (2022) *VoxCommunis corpus*. https://osf.io/t957v.

An IPA-transcribed word list.

Tocharian B

*Adams, Douglas Q. (2013) *A Dictionary of Tocharian B*. Amsterdam, New York: Rodopi.

The words were extracted from the online index (restricted access):
https://dictionaries.brillonline.com/search#dictionary=tocharian_b&id=tchb00337

Pronunciation generated by grapheme-to-phoneme conversion.

Turkish

McAuliffe, Michael and Morgan Sonderegger. (2024) *Turkish MFA dictionary v3.0.0*.
https://github.com/MontrealCorpusTools/mfa-models/releases/tag/dictionary-turkish_mfa-v3.0.0 (accessed 15 April 2024).

An IPA-transcribed word list.

Udmurt

Kiss, Katalin É. et al. (not dated). *Languages under the Influence. Uralic syntax changing in an asymmetrical contact situation*. https://archive.nytud.hu/depts/tlp/uralic/dbases.html (accessed 27 March 2024).

IPA-transcribed texts were used.

Vietnamese

McAuliffe, Michael and Morgan Sonderegger. (2024) *Vietnamese MFA dictionary v3.0.0*. https://github.com/MontrealCorpusTools/mfa-models/releases/tag/dictionary-vietnamese_mfa-v3.0.0 (accessed 10 April 2024)

An IPA-transcribed word list.

Welsh

Geiriadur Ynganu Prifysgol Bangor / Bangor University pronunciation dictionary. https://github.com/techiaith/geiriadur-ynganu-bangor (accessed 7 November 2023).

An IPA-transcribed word list.

Xitsonga

Lwazi Xitsonga Pronunciation Dictionary. https://taalmaterialen.ivdnt.org/download/tstc-lwazi-xitsonga-pronunciation-dictionary/ (accessed 23 November 2023).

Davel, Marelie and Olga Martirosian. (2009) "Pronunciation dictionary development in resource-scarce environments". In *10th annual conference of the International Speech Communication Association (INTERSPEECH)*, 2851–2854. Brighton.

Available for free after registration. SAMPA-transcribed pronunciation.