Introducing a phonotactic probability calculator for Czech

Petra Čechová | Department of Linguistics, Faculty of Arts, Charles University
Luca Cilibrasi | Department of Social, Political and Cognitive Sciences, University of Siena – Department of Linguistics and Department of English Language and ELT Methodology, Faculty of Arts, Charles University
Jan Henyš | Institute of the Czech National Corpus and Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University
Jaroslav Čecho

Phonotactic probability refers to the frequency with which phonological segments and sequences of phonological segments occur in words in a given language (Vitevitch – Luce, 2004). It has been shown that phonotactic probabilities of words are important in language processing and language acquisition (Jusczyk et al., 1994; Mattys – Jusczyk, 2001; Pitt – McQueen, 1998). For example, words with high phonotactic probability are processed faster by native speakers in same-different tasks (Luce – Large, 2001), and pseudowords with high phonotactic probability are judged as more word-like by adults (Vitevitch et al., 1997). In this paper we present a phonotactic calculator for Czech implemented as a Python script. The script relies on frequency data from three freely available corpora of Czech: SYN2015 and SYN2020, corpora of written Czech (Křen et al., 2015; 2020), and ORAL v1, a corpus of spoken Czech (Kopřivová et al., 2017). The steps of the calculation mirror those developed by Vitevitch and Luce (2004) for English, and the script can provide phonotactic (and additionally orthotactic) probability for any Czech word or pseudoword. The script can be downloaded at <https://phonocalc.github.io>.

Key words: Czech, language processing, phonotactic probability, phonotactics, pseudoword

Klíčová slova: čeština, fonotaktická probabilita, fonotaktika, pseudoslovo, zpracování jazyka

1 Phonotactic probabilities

Phonotactic probabilities are a measure of the frequency of phonological patterns in a given language (Vitevitch – Luce, 2004). Typically, phonotactic probability calculators can provide two types of information: the first one is a segmental measure, and it indicates how probable it is for a certain phoneme to appear in a particular position. The second goes beyond the probability of a certain segment, and it indicates the probability of two (or more) segments appearing next to each other.

The first measure may be explained by referring to phonological rules in a more traditional sense. For example, in Czech, a phonological rule establishes that an obstruent in the final positions of an isolated word is devoiced (Šimáčková et al., 2012).

Due to this rule, the probability of voiced obstruents appearing in the final position of isolated words equals zero. In Italian, the quasi-totality of words end in a vowel, as established by one of its phonological rules (Rogers – d'Arcangeli, 2004). Consequently, the probability of any consonant appearing in a word final position is zero (or approximately zero) in Italian. These principles can be applied to any phoneme, and the probability can be calculated of each phoneme appearing in a given position. For instance, one may want to know what is the probability that a word will start with /p/ in English. Phonotactic probability calculators analyse large corpora and can provide a numerical estimate that answers this question. Since the frequencies of phonemes tend to be skewed (some phonemes are extremely frequent in certain positions while others are extremely rare), calculations of phonotactic probabilities are often conducted using the logarithmic scale. The logarithmic scale "reduces" the distances between the frequencies of individual phonemes, while preserving the appropriate order. For instance, if we imagine a phoneme A appearing 8% of the time in the word-initial position, a phoneme B appearing 0.5% of the time, and a phoneme C appearing 0.008% of the time, these are the log-transformed values (base 10): 0.9, −0.3, and −2.1 This approach collapses the differences and makes it easier to use the phonotactic value for further comparisons.

The most widely used calculator for English (Vitevitch – Luce, 2004) relies on this procedure, thus indicating the logarithmic phonotactic probability for each phoneme (in each position). For instance, the log-transformed probability of /p/ appearing in the word-initial position is 0.08, while for /t/ the log-transformed probability is 0.04. These values are obtained based on a sample of language "in use," and are the result of dividing the logarithm of the frequencies of all words starting with /p/ or /t/ by the logarithm of frequencies of all words with any segment in the same position (i.e., all words that are at least one segment long). Note that alternative systems have been proposed; see for instance Hayes (2012) and Dziubalska-Kołaczyk and Pietrala (2020).

In substance, this is a bottom-up approach to linguistic analysis, where large collections of language in use (corpora) are used to extract phonological patterns and thus to understand phonological phenomena (Gries, 2010).

As already mentioned, the second type of information that a phonotactic probability calculator typically provides goes beyond the probability of a certain segment in a given position, since it indicates the probability of two (or more) segments appearing next to each other. For instance, the sequence /rpa/ is composed of two biphones: /rp/ and /pa/. When we insert this sequence in the English phonotactic probability calculator, the output given is zero for the first biphone, and 0.001 for the second one. In plain words, this indicates that the sequence /rp/ is never attested in English in the word-initial position, while the sequence /pa/ in the second position is marginally attested.

Phonotactic probabilities have been shown to be important in language processing and language acquisition in a wide number of studies: From the very early stages of linguistic development, infants have been found to be sensitive to the phonotactic probabilities of their linguistic environment (Jusczyk et al., 1994). A complex challenge for preverbal infants is speech segmentation, namely the ability to recognise where words start and end in the stream of speech. Experimental studies with 9-month-olds demonstrate that phonotactic probabilities in child-directed speech predict what items will be recognised by children as words (Mattys – Jusczyk, 2001). As for adults, phonotactic probabilities have been shown to be relied on by speakers in several contexts: first, studies on speech perception show that phonotactic probabilities are used for speech segmentation in adults as well, in both first (Pitt – McQueen, 1998) and second language processing (Weber – Cutler, 2004). Second, studies on lexical processing show that words with higher phonotactic probabilities are processed more quickly in same-different tasks (Luce – Large, 2001), indicating a role of phonotactic probabilities in lexical entrenchment. Finally, studies with pseudowords demonstrate that the speed of processing of novel items is also predicted by their phonotactic probability. Pseudowords consisting of unlikely sequences of phonemes are processed more slowly (Janse – Newman, 2013), suggesting that not only our previously stored words, but also our expectations about new lexical items are related to phonotactic probabilities. Altogether, these findings indicate that the phonological representations in our mind are highly affected by probabilistic parameters, and that our sensitivity to phonological sequences in speech is related to our experience with the given language, since phonotactic probabilities are an emergent property of corpora (or, more accurately, of what corpora represent: language use).

2 Beyond English

As is often the case in psycholinguistic research, the first language ever investigated when it comes to phonotactic probabilities was English. The first widely available phonotactic probability calculator for English was developed by Vitevitch and Luce (2004) and can be accessed online.[1] Vitevitch and Luce's calculator has been used in hundreds of studies since its publication, and it has become a work of reference for the field. Vitevitch et al. (2014) suggest that English may be seen as a model language for the evaluation of linguistic phenomena, similarly to what happens in biology, where certain species are seen as model species. In biology, some species have received an uneven amount of attention in research, and they are often used to

[1] See <http://www.people.ku.edu/~mvitevit/PhonoProbHome.html>.

investigate phenomena that are then extended to various related species. For example, rats have been widely used in clinical research, for very pragmatic reasons: they are easily available, considerably more numerous than any other mammal, and their biology is quite close to ours (Bryda, 2013). Vitevitch et al. (2014) propose that a parallel approach may be found in psycholinguistics, where English speakers are by far the largest group available for testing, and many phenomena studied in English extend to other languages. However, just as with the limits of model species in biology, the concept of model language in linguistics bears the obvious limit that not all traits of English are relevant for other languages. Psycholinguistic research is thus gradually addressing more languages today (Norcliffe et al., 2015). Phonotactic research is no exception, and due to these considerations, a number of other phonotactic probability calculators have been created over the years. Calculators are now available for Spanish (Davis – Perea, 2005), Dutch (Messer et al., 2010) and Arabic (Aljasser – Vitevitch, 2018), and tools have been developed for extracting phonotactic probabilities from additional languages, given a corpus (Hall et al., 2019). This article introduces a freely available Python script that computes phonotactic probability for Czech.

Czech is a language that belongs to the Indo-European family, stemming from the Proto-Slavic branch (Schenker, 1993). Slavic languages display some interesting phonotactic phenomena that are not mirrored in other branches of the Indo-European family and thus call for their own phonotactic probability calculators. One relevant aspect observed in the phonology of Slavic languages is that they allow considerably longer and more varied consonant clusters than other Indo-European languages; in fact, Slavic languages appear to be on the extreme end of the spectrum when it comes to licit consonant sequences, even if one includes a wider sample of languages from different families (Easterday, 2019). This simple yet crucial property makes Czech an ideal candidate for further research on phonotactic probabilities.

In addition to these relevant properties of consonants, Czech vowels are also very interesting because Czech displays pure lengthening. English is often described as a language containing long and short vowels, but a phonetic analysis of minimal pairs that are supposedly differentiated by this property shows that differences in length are often additionally associated with differences in quality. For example, the distinction between *pull* and *pool* is not uniquely based on length, but also on quality (at least in received pronunciation): /pʊl/ vs /puːl/. In Czech, instead, words that form minimal pairs related to vowel lengthening actually display exactly the same vowels, but with different length (this holds for the vowels /a/ – /aː/, /ɛ/ – /ɛː/, /o/ – /oː/ and at least partially /u/ – /uː/ (see for example the distinction between the word *past* ‘a trap’ and the word *pást* ‘to graze’: /past/ vs /paːst/), while the difference between /ɪ/ and /iː/ is also qualitative (Skarnitzl – Volín, 2012)).

In summary, due to typological considerations regarding already available calculators and due to some inherent phonetic properties of Czech, we suggest that a calculator for this language could offer an important contribution to our overall understanding of phonotactic probabilities in human language.

3 The Czech calculator

The Czech calculator presented here is implemented as a script in Python. It can be used via the command line. It follows the original calculator for phonotactic probability in English by Vitevitch and Luce (2004) in that it uses the same method of calculation. Unlike the original calculator, it also computes orthotactic probability based on grapheme positions in words (note that unlike in English, the correspondence between orthographic and phonemic representation of words in Czech is fairly straightforward). To calculate phonotactic probability based on the pronunciation of words, it relies on automatic IPA transcription (note that Vitevitch and Luce (2004) used a manually transcribed frequency list). See <https://phonocalc.github.io> for instructions and download.

3.1 Source data

To estimate phonotactic or orthotactic probabilities, one needs to know the frequency of phonemes or graphemes and sequences of phonemes or graphemes in given positions within words. For this purpose, the present script draws on case-insensitive token frequency lists from three freely available Czech corpora: SYN2015, a synchronic reference corpus of contemporary written Czech mostly from the years 2010–2014 (Křen et al., 2015), SYN2020, a newer version of the same type of corpus from the years 2015–2019 (Křen et al., 2020), and ORAL v1, a spoken corpus of informal conversations (Kopřivová et al., 2017). The user of the script can select which frequency list to use; SYN2015 is currently set as the default. All three lists are adjusted versions of the list provided by the corpus interface. To avoid word forms with typographical or tokenizing errors, only tokens with frequency higher than 10 are used. This procedure also removes all hapax legomena (i.e., all tokens that occur only once in a given corpus), which famously account for roughly 50% of tokens in a corpus (Baayen, 2001), hence rendering the execution of the script much faster. Next, word forms containing non-Czech characters or an apostrophe were removed, as these are usually foreign words (most often names or toponyms). We also attempted to remove initialisms, e.g., *OSN* /oːʔɛsʔɛn/ ‘UN.’ For this purpose, we manually examined a list of abbreviations found in SYN2020 (Křen et al., 2020), excluded all entries that are orthographically homonymous with other word forms, such as *SRN* /ɛsʔɛrʔɛn/ (abbr. for the Federal Republic of Germany) and *srn* /sr̩n/

(the genitive plural of 'roe deer'), and acronyms, e.g., *NATO* /nato/ 'NATO,' and removed all the remaining abbreviations from the frequency lists. The last step was the deletion of word forms ending with a full stop, most frequently abbreviations as well, e.g., *dr.* 'Dr.' (usually pronounced as a whole word, /*doktor*/ or /*doktorka*/).

In order to calculate phonotactic probability, the script automatically transcribes the chosen frequency list into IPA, using the Python package CorPy (Lukeš, 2021). The transcription consists in a rule-based grapheme-to-phoneme conversion with an exception system for the highest frequency irregular words. The accuracy of the transcription was measured based on the manual inspection of a sample of 1% of words in the adjusted frequency lists of SYN2015 and ORAL v1. The accuracy in the frequency list of the spoken corpus, ORAL v1, equalled 98.1% (meaning that 98.1% of words were transcribed correctly, the remaining words contained one or more errors). The accuracy in the written corpus, SYN2015, equalled 90.7%. The accuracy in SYN2015 was lower mainly due to a higher number of foreign words, especially names and toponyms that were not removed because they did not contain any non-Czech characters, and exceptional Czech words of foreign origin that do not respect the typical grapheme-to-phoneme correspondence. Nevertheless, we regard the accuracy as sufficient since it is similar to the accuracy of automatic rule-based phonetic transcription of Romanian (Toma – Munteanu, 2009) and of an English automatic transcriber available online.[2]

However, the transcriptions obtained in this way are transcriptions of word forms in isolation. Czech is a language where voicing assimilation occurs even across word boundaries (final voiceless obstruents become voiced when preceding a voiced obstruent or, in some regions, a sonorant or a vowel (Šimáčková et al., 2012)). This means that the word form *hrad* 'castle' is pronounced as /hrat/ in isolation or when followed by a voiceless obstruent but becomes /hrad/ before words starting with a voiced obstruent: /hrad byl/ 'castle was'. This is not taken into account by the script. While the effects of connected speech are not taken into consideration in the original calculator by Vitevitch and Luce (2004) either, voicing assimilation across word boundaries does not occur in English.

In addition to the frequencies obtained from the aforementioned corpora, the script can also be executed using a custom frequency list. This might be useful for future experiments when new corpora are released, but it can also serve researchers who want to calculate phonotactic probabilities based on the speech of specific groups of speakers (e.g., children) or use manually transcribed or more controlled frequency lists in order to obtain more accurate phonotactic probability estimates.

[2] See <http://tom.brondsted.dk/text2phoneme/?vieweval&l=English>.

3.2 Calculation and script characteristics

To calculate phonotactic probabilities, Vitevitch and Luce (2004) use two measures, positional unigram frequency and position-specific bigram frequency. The present script goes one step further, meaning that the calculation can be executed based on any *n*-grams. The unigram positional probability is calculated by searching the selected frequency list for all words that contain a given segment in a given position. To give an example, for the word *srna* /sŗna/ 'roe deer,' this means searching for every word with /s/ in the first position, /r/ in the second one, /n/ in the third one and /a/ in the fourth one. Then, for every segment, log-transformed (base 10) frequencies of the respective words are summed and divided by the sum of log-transformed frequencies of all words which contain any segment in the same position (in the case of /sŗna/, for /s/ this means all words that are at least one segment long, for /r/ all words that are at least two segments long, and so forth). The arithmetic mean of these values for every segment in a word serves as the final estimate of phonotactic probability. The same applies to bigram or *n*-gram position-specific frequency, the only difference being that the procedure works with a sequence of two or *n* segments in a given position instead of one. For the position-specific bigram frequency of the word *srna*, this would mean searching for every word with /sr/ in the first position, /rn/ in the second one and /na/ in the third one and then executing the calculation as described in the case of the unigram. The calculation works the same for orthotactics, in which case the *n*-grams consist of graphemes rather than phonemes.

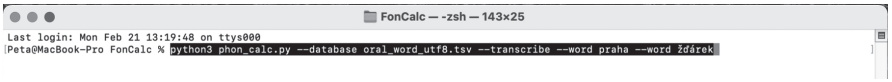

Figure 1: Input. The highlighted command specifies that the script phon_calc.py is to be run with the file oral_word_utf8.tsv as the chosen frequency list, transcription is enabled, and we are interested in the words *praha* and *žďárek*.

The input for the script is any existing or non-existing word or an entire list of (pseudo)words. By default, the script computes position-specific bigram frequency based on orthography, hence calculating orthotactic probability on the basis of the SYN2015 frequency list. The use of position-specific bigram frequency as the basis for the calculation can be changed to any *n*-gram, and phonetic transcription can be enabled by means of specific options. Other, either ready-made or custom, frequency lists can be employed. The script also allows for transcribing the entire frequency list into IPA and printing it out, which can be useful especially when one uses a custom frequency list and wants to check its accuracy.

The output of the script consists of several lines of text.

The first line states the name of the frequency list file that the calculation was based on. The following lines list the *n*-grams of the input, their position in the word and the number of words containing the same *n*-gram in the same position in the frequency list. The final line repeats the input followed by the final estimate of orthotactic or phonotactic probability, which is always a number between 0 and 1. For example, based on the SYN2015 frequency list, the phonotactic probability of the word *Praha* /praɦa/ 'Prague' is 0.0118, while the estimate for the word *Ždárek* /ʒɟaːrɛk/ (another toponym) with the not very frequent syllabic onset *žď* /ʒɟ/ is far lower: 0.0017.

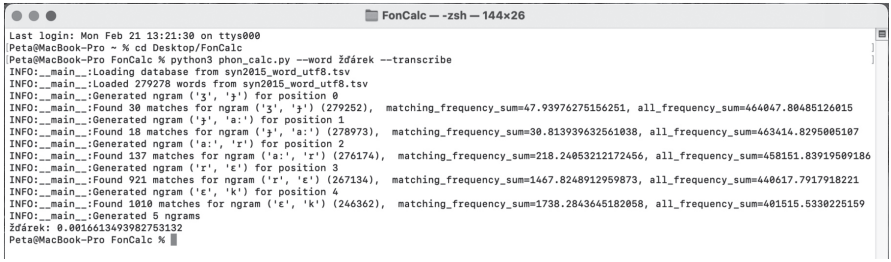

Figure 2: Output: Phonotactic probability (position-specific bigram frequency) of the word *žďárek*.

When working with the script, we noticed that the method of calculation gives more weight to the beginnings of words than to their endings. However, in Czech, a large number of words end with the same sequences of segments due to morphological reasons, e.g., the sequence *-ka* /ka/ in words such as *žačka* /ʒa͡ʧka/ 'a female pupil,' *učitelka* /uʧ͡ʃɪtɛlka/ 'a female teacher,' and *hokejistka* /ɦokɛjɪstka/ 'a female hockey player.' For the calculator, on the other hand, the sequence appears in different positions due to variations in word length (in the three examples, the sequence *-ka* appears in the fourth, seventh and ninth position of each word respectively). This means that the overall phonotactic or orthotactic probability estimate of a word ending in *-ka* reflects only those words ending in *-ka* that have the same length and ignores all the other words ending with *-ka*, even though this kind of regularity is certainly abstracted out from the input by speakers. We decided to address this problem with an optional feature of reversed calculation. In this case, the script first reverses the order of segments within words in the frequency list (when calculating phonotactic probability it does so after transcribing to IPA), then it calculates the probabilities as usual. Averaging the result of the standard and reversed calculation should lead to phonotactic or orthotactic probability estimates that give the same weight to the initial and the final segments of words.

4 Limitations and future directions

While we believe that the calculator represents an important addition to the phonetic and psycholinguistic tools available for the Czech language, some of its features are not ideal and may be improved in its future versions.

An obvious drawback of the script lies in the accuracy of the IPA transcription, which does not reach 100%. This is caused by two factors: firstly, by the absence of the effects of connected speech (such as inter-word voicing assimilations) in the source frequency list, and secondly, by the fact that the mostly rule-based automatic transcription cannot capture pronunciations of foreign proper nouns or certain other words of foreign origin. Future improvements could include initial automatic transcription of the whole corpus into IPA (one would need to have access to the entire text collections of the corpus) with manual corrections of the erroneously transcribed words and the subsequent generation of a frequency list based on those transcriptions. This approach would provide a more accurate representation of the phonetic structure of the language and thus yield more accurate estimates of phonotactic probability.

A second limitation of this kind of calculator is the fact that the calculation does not take into account syllabification. Vitevitch and Luce (2004) identify factors that may come into play in English such as stress pattern, onset-rime syllable structure and relationships between phonology and grammatical part of speech that are not included as constraints in the calculation. They, however, argue (and their statements are equally important for our calculator), that this neutrality towards other factors may be actually helpful to some extent; the calculation itself is not required to change every time a new factor is described, and by not implementing other factors into the calculation, researchers can examine the interaction of phonotactic probability with other factors more easily.

In summary, further developments of the script may benefit from improvements in the automatic grapheme-to-phoneme transcription and from a more advanced treatment of other factors such as syllabification. Nonetheless, the calculator already provides very important metrics that had not been available for Czech before.

REFERENCES

Aljasser, Faisal – Vitevitch, Michael S. (2018): A web-based interface to calculate phonotactic probability for words and nonwords in Modern Standard Arabic. *Behavior Research Methods*, 50(1), pp. 313–322. <https://doi.org/10.3758/s13428-017-0872-z>.

Baayen, R. Harald (2001): *Word Frequency Distributions*. Dordrecht – Boston, MA – London: Kluwer Academic Publishers. <https://doi.org/10.1007/978-94-010-0844-0>.

Bryda, Elizabeth C. (2013): The Mighty Mouse: the impact of rodents on advances in biomedical research. *Missouri Medicine*, 110(3), pp. 207–211.

Davis, Colin J. – Perea, Manuel (2005): BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, 37(4), pp. 665–671. <https://doi.org/10.3758/BF03192738>.

Dziubalska-Kołaczyk, Katarzyna – Pietrala, Dawid (2020): The NAD phonotactic calculator: an online tool to calculate cluster preferability across languages. In: Hans Sauer – Piotr P. Chruszczewski (eds.) *Mostly Medieval: In Memory of Jacek Fisiak*. San Diego, CA: Æ Academic Publishing, pp. 445–458.

Easterday, Shelece (2019): *Highly Complex Syllable Structure: A Typological and Diachronic Study*. Berlin: Language Science Press.

Gries, Stefan T. (2010): Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily… *International Journal of Corpus Linguistics*, 15(3), pp. 327–343. <https://doi.org/10.1075/ijcl.15.3.02gri>.

Hall, Kathleen C. – Mackie, J. Scott – Lo, Roger Y. (2019): Phonological CorpusTools: Software for doing phonological analysis on transcribed corpora. *International Journal of Corpus Linguistics*, 24(4), pp. 522–535. <https://doi.org/10.1075/ijcl.18009.hal>.

Hayes, Bruce (2012): BLICK: a phonotactic probability calculator [online; manual]. Last accessed: 23 January, 2023. <https://linguistics.ucla.edu/people/hayes/BLICK/index.htm>.

Janse, Esther – Newman, Rochelle S. (2013): Identifying nonwords: effects of lexical neighborhoods, phonotactic probability, and listener characteristics. *Language and Speech*, 56(4), pp. 421–441. <https://doi.org/10.1177/0023830912447914>.

Jusczyk, Peter W. – Luce, Paul A. – Charles-Luce, Jan (1994): Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), pp. 630–645. <https://doi.org/10.1006/jmla.1994.1030>.

Kopřivová, Marie – Lukeš, David – Komrsková, Zuzana – Poukarová, Petra – Waclawičová, Martina – Benešová, Lucie – Křen, Michal (2017): *ORAL v1: A Corpus of Informal Spoken Czech* [online]. Prague: Institute of the Czech National Corpus. Last accessed: 28 June, 2022. <http://www.korpus.cz>.

Křen, Michal – Cvrček, Václav – Čapka, Tomáš – Čermáková, Anna – Hnátková, Milena – Chlumská, Lucie – Jelínek, Tomáš – Kovaříková, Dominika – Petkevič, Vladimír – Procházka, Pavel – Skoumalová, Hana – Škrabal, Michal – Truneček, Petr – Vondřička, Pavel – Zasina, Adrian (2015): *SYN2015: A Representative Corpus of Written Czech* [online]. Prague: Institute of the Czech National Corpus. Last accessed: 28 June, 2022. <http://www.korpus.cz>.

Křen, Michal – Cvrček, Václav – Henyš, Jan – Hnátková, Milena – Jelínek, Tomáš – Kocek, Jan – Kovaříková, Dominika – Křivan, Jan – Milička, Jiří – Petkevič, Vladimír – Procházka, Pavel – Skoumalová, Hana – Šindlerová, Jana – Škrabal, Michal (2020): *SYN2020: A Representative Corpus of Written Czech* [online]. Prague: Institute of the Czech National Corpus. Last accessed: 28 June, 2022. <http://www.korpus.cz>.

Luce, Paul A. – Large, Nathan R. (2001): Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16(5–6), pp. 565–581. <https://doi.org/10.1080/01690960143000137>.

Lukeš, David (2021): *CorPy 0.3.1* [online; Python package]. Last accessed: 23 January, 2023. <https://github.com/dlukes/corpy>.

Mattys, Sven L. – Jusczyk, Peter W. (2001): Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2), pp. 91–121. <https://doi.org/10.1016/s0010-0277(00)00109-8>.

Messer, Marielle H. – Leseman, Paul P. M. – Boom, Jan – Mayo, Aziza Y. (2010): Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology*, 105(4), pp. 306–323. <https://doi.org/10.1016/j.jecp.2009.12.006>.

Norcliffe, Elisabeth – Harris, Alice C. – Jaeger, T. Florian (2015): Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience*, 30(9), pp. 1009–1032. <https://doi.org/10.1080/23273798.2015.1080373>.

Pitt, Mark A. – McQueen, James M. (1998): Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39(3), pp. 347–370. <https://doi.org/10.1006/jmla.1998.2571>.

Rogers, Derek – d'Arcangeli, Luciana (2004): Italian. *Journal of the International Phonetic Association*, 34(1), pp. 117–121. <https://doi.org/10.1017/S0025100304001628>.

Schenker, Alexander M. (1993): Proto-Slavonic. In: Bernard Comrie – Greville G. Corbett (eds.), *The Slavonic Languages*. London: Routledge, pp. 60–124.

Skarnitzl, Radek – Volín, Jan (2012): Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy*, 18(1), pp. 7–11.

Šimáčková, Šárka – Podlipský, Václav Jonáš – Chládková, Kateřina (2012): Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association*, 42(2), pp. 225–232. <https://doi.org/10.1017/S0025100312000102>.

Toma, Ştefan-Adrian – Munteanu, Doru-Petru (2009): Rule-based automatic phonetic transcription for the Romanian language. *COMPUTATIONWORLD '09: Proceedings of the 2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns*. Washington, D.C.: IEEE Computer Society, pp. 682–686. <https://doi.org/10.1109/ComputationWorld.2009.59>.

Vitevitch, Michael S. – Luce, Paul A. – Charles-Luce, Jan, – Kemmerer, David (1997): Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech*, 40(1), pp. 47–62. <https://doi.org/10.1177/002383099704000103>.

Vitevitch, Michael S. – Luce, Paul A. (2004): A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, 36(3), pp. 481–487. <https://doi.org/10.3758%2Fbf03195594>.

Vitevitch, Michael S. – Chan, Kit Ying – Goldstein, Rutherford (2014): Using English as a 'model language' to understand language processing. In: Nick Miller – Anja Lowit (eds.), *Motor Speech Disorders: A Cross-Language Perspective*. Bristol – Buffalo, NY – Toronto: Multilingual Matters, pp. 58–73. <https://doi.org/10.21832/9781783092338-006>.

Weber, Andrea – Cutler, Anne (2004): Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50(1), pp. 1–25. <https://doi.org/10.1016/S0749-596X(03)00105-0>.

Nový nástroj pro výpočet fonotaktické probability v češtině

Pojem fonotaktická probabilita odkazuje k frekvenci, s níž se fonologické segmenty a sekvence těchto segmentů objevují ve slovech v určitém jazyce (Vitevitch – Luce, 2004). Předchozí výzkumy ukázaly, že fonotaktická probabilita slov hraje důležitou roli při zpracování a akvizici jazyka (Jusczyk et al., 1994; Mattys – Jusczyk, 2001; Pitt – McQueen, 1998). Slova s vysokou fonotaktickou probabilitou

jsou například rychleji zpracována rodilými mluvčími v úloze „same-different“ (Luce – Large, 2001) a pseudoslova s vysokou fonotaktickou probabilitou jsou dospělými hodnocena jako pravděpodobnější slova daného jazyka (Vitevitch et al., 1997). V tomto článku představujeme nástroj pro výpočet fonotaktické probability pro češtinu, který je volně dostupný jako skript v programovacím jazyce Python. Nástroj vychází z údajů o frekvenci slov ze tří volně dostupných korpusů českého jazyka: korpus psaného jazyka SYN2015 (Křen et al., 2015), korpus psaného jazyka SYN2020 (Křen et al., 2020) a korpus mluveného jazyka ORAL v1 (Kopřivová et al., 2017). Výpočet replikuje postup původního kalkulátoru pro anglický jazyk (Vitevitch – Luce, 2004) a výstupem je odhad fonotaktické (a navíc také ortotaktické) probability pro jakékoliv české slovo či pseudoslovo. Skript je dostupný z internetové stránky <https://phonocalc.github.io>.

Department of Linguistics, Faculty of Arts, Charles University
náměstí Jana Palacha 1/2, 116 38 Praha 1
petra.cechova@ff.cuni.cz
luca.cilibrasi@unisi.it
jan.henys@ff.cuni.cz