

ANOPHONE: AN ANNOTATION TOOL FOR PHONEMES AND L2 ANNOTATION SYSTEMS FOR CZECH

RICHARD HOLAJ¹ – PETR POŘÍZKA²

¹ Department of Czech Language, Faculty of Arts, Masaryk University,
Brno, Czech Republic

² Department of Czech Studies, Faculty of Arts, Palacký University,
Olomouc, Czech Republic

HOLAJ, Richard – POŘÍZKA, Petr: ANOPHONE: An Annotation Tool for Phonemes and L2 Annotation Systems for Czech. *Journal of Linguistics*, 2023, Vol. 74, No 1, pp. 333 – 344.

Abstract: The goal of this text is the presentation of the ANOPHONE annotation system, which allows for the management and annotation of speech data to develop a tool for the automatic transcription of speech of non-native speakers of Czech. This system is currently designed for annotations on the segmental level of recordings of non-native speakers of Czech, with the aim to train automatic speech recognition (ASR) models used in this tool.

After an introductory section that discusses the use of technology in pronunciation teaching and mentions some of the e-learning applications for teaching the pronunciation of second languages (L2), we address both general and more specific aspects of speech data annotation to train ASR models and mention attributive and synthetic segmental systems of speech data annotation for Czech as L2. We also briefly introduce the annotation system of non-native speakers of Czech called BV1, which is used for testing the ANOPHONE tool. The main part of this text focuses on presenting the annotation tool itself, while the conclusion describes the experience of testing the speech data annotation tool using BV1 annotation system for Czech as L2.

Keywords: annotation, ANOPHONE, pronunciation, L2, ASR, Czech, e-learning

1 THE USES OF TECHNOLOGY IN PRONUNCIATION TEACHING

The importance of e-learning has been growing dramatically in recent years, as has been its use in foreign language teaching (Blake 2011, cf. also Holaj 2018, Holaj – Pořízka 2021). One of the challenges is in tools that would help with the issues of teaching pronunciation to non-native speakers, especially in providing individualized feedback and knowledge verification, as these elements are extremely time- and skill-intensive for teachers. Technology presents itself here as the ideal solution. Derwing and Munro (2015) come to the same conclusion, stating that such tools should be able to identify and automatically evaluate pronunciation errors that have a negative effect on the intelligibility and accessibility of speech production, or provide the user with exercises with feedback and monitor their progress through practice. This is referred to as computer-assisted pronunciation training (CAPT) and appears to be the most promising method for improving the level of teaching

pronunciation to non-native speakers (cf. Levis 2007). Tools that meet these requirements involve the use of machine learning technology.

Although there already exist several tools that address Czech pronunciation learning, notably CzechME (Holaj et al. 2021) and ProCzeFor (Veroňková et al. 2022), none of these tools provide an automatic feedback on pronunciation. It should be mentioned that feedback on Czech L2 pronunciation is provided by the Czech version of the Duolingo application, but it works on the basis of native speaker speech recognition tools and its output is heavily biased (cf. Nováčková 2017, 2018). The situation is better in the teaching of, for example, English pronunciation, where several advanced tools exist, notably ELSA Speak (Elsa Corp.), which focuses only on pronunciation teaching, but even this state-of-the-art application is prone to significant errors in feedback (Becker – Edalatishams 2019). This may be due to the fact that essentially all automatic feedback tools for native speaker pronunciation only work with segments that occur in the pronunciation of native speakers of a given language (Holaj – Pořízka 2021).

Such annotation is time consuming and places specific demands on the annotators. In this paper, we want to show that this process can also be greatly facilitated with the use of technology, specifically by using the ANOPHONE annotation tool we have developed, which reflects these specific requirements. In the following section, we first briefly describe the issue of speech annotation and then use the example of several speech annotation systems for non-native speakers of Czech to show what such annotation can look like. In the last part we will introduce the ANOPHONE tool itself and show how it can facilitate the process of annotating the speech of non-native speakers using one of the described systems.

A solution to this problem could be a speech recognition system trained on data from non-native speakers, which would take into account speech segments that do not occur in the pronunciation of native speakers. However, such a system, which is already under development (Holaj 2023), requires a huge amount of annotated audio data (recordings). Such annotation is time-consuming and places specific demands on the annotators. In this paper, we want to show that this process can be greatly facilitated by using technology, specifically by using the ANOPHONE annotation tool, which reflects these specific requirements. In the following section, we first briefly describe the issue of speech annotation and then, using the example of several systems for annotating the speech of non-native speakers of Czech, we show what such annotation can look like. In the last section, we introduce the ANOPHONE tool itself and show how it can facilitate the process of annotating the speech of non-native speakers using one of the mentioned L2 Czech systems.

2 SPEECH DATA ANNOTATION AND L2 ANNOTATION SYSTEM FOR CZECH

Compared to typical linguistic annotation, speech annotation is different in that it uses audio data rather than text data. An essential criterion for speech annotation is

deciding what the purpose of the annotation is and what aspects or features of the audio recordings we want to capture or annotate. In most cases, sequential annotation is used, which annotates speech as a sequence of phonemes (or speech sounds), syllables, words or other sound segments. This design also includes all transcription-based annotations of recordings. The essential feature of sequential annotation is that the recording is annotated as a sequence of tags (or labels). This implicitly presumes that the recording itself is also a sequence of segments, with each such segment corresponding to one label in the recording's annotation. The order of labels (or tags) and speech segments mutually corresponds in the vast majority of cases. However, aligning sequential annotations represents somewhat of a problem, as the individual segments (speech sounds, syllables, words) are not of the same length. Thus, the labels correspond to sections of different lengths, and a simple sequential annotation alone does not provide information on how to align these labels with the recording. Currently, this issue is most often resolved by using a programme which uses the annotation, such as the CTC (Graves et al. 2006) algorithm. Thanks to this algorithm, there is no need to laboriously include the manual alignment in the annotation, and all that is required is merely to indicate the sequence of the labels (tags) themselves, regardless of their length and alignment with the recording. This is the approach we chose for our ASR application.

In the context of segmental annotation, it is necessary to mention the form of the labels for the individual segments. The most common are simple synthetic labels (tags) that have no internal structure and cannot be further divided. Each segment is then assigned exactly one of these labels. Within this approach, one can imagine, for example, a simple annotation of speech sounds where each speech segment (corresponding speech sound) is assigned exactly one of the possible labels (the one that corresponds to the given speech sound or is closest to it).

In addition to the sequential synthetic label system, there are, for example, attributive or positional systems: an example of an attributive annotation of spoken language are the variant annotations AV1 and AV2 (see Holaj – Pořízka 2021 for more), which we tested in the first phase of the project, but based on the results of the trained models, we decided to change the annotation and replace it with a synthetic one called BV1 (Holaj 2023). An illustrative comparison of a few selected labels for annotations AV1, AV2 and BV1 is shown in Tab. 1.

AV1	AV2	BV1
e::a	a	a
k:z1	g	g
t:vM	t:vM	tʲ

Tab. 1. Illustrative comparison of a few selected annotation labels of AV1, AV2 and BV1 annotation systems

The new annotation system BV1 is a sequence of synthetic tags (labels) without an internal structure. A recording (sound segment) is always assigned a sequence of these

tags separated by a space and corresponding to the sequence of speech sounds in the recording. This system meets the requirements for simplicity (the annotation corresponds only to the identifier of the respective speech sound), using as few tags as necessary (only sounds selected by a user are part of the inventory) and extensibility (the annotation can be extended with new tags and it does not interfere with other labels).

For the purpose of clarity, these labels were divided into several groups based on two criteria. The primary criterion is the type of speech sound corresponding to the label. Along those lines, we divide the tags into vowels, consonants and other tags (such as aspiration, quantity, etc.), which always refer to the previous segment. The secondary criterion indicates whether this is a speech sound from the standard Czech inventory or a non-standard speech sound. The individual tags available in BV1 are described in a work on a tool for automatic speech transcription of non-native speakers of Czech (Holaj 2023).

The standard procedure for annotation in cases where the pronounced speech sound is not among the available tags is to choose the perceptually closest speech sound. If the pronunciation is perceived as acceptable (i.e., it is not a fundamental error in terms of intelligibility), standard speech sounds are preferred if they are roughly similar. Non-standard speech sounds are preferred in cases of unacceptable pronunciation.

A clear advantage of this system is its simple extensibility. Without affecting the already existing tags, additional segments can be added into this system on the basis of new findings, which has already occurred several times over the course of data annotation with BV1. This makes it easy, for example, to add segments specific to certain nationalities as they are included among the data being annotated. The described system was used to test our annotation tool ANOPHONE.

3 ANOPHONE AND SPEECH DATA ANNOTATION

In order to create a tool based on ASR technology, it is first necessary to collect enough speech data from non-native speakers (in our case Czech data) and subsequently annotate it. This annotated data is then trained via neural networks and used to create a speech recognition model. For this annotation, we developed the ANOPHONE application,¹ which fulfils a dual purpose. Firstly, it serves as an online database of recordings of non-native speakers and a repository of their annotations with the ability to filter data using metadata. The second purpose is the annotation of these recordings itself through the annotation tasks that the tool allows us to create; i.e., we can add the task for the annotation of speech sounds or another task for the annotation of stressed and unstressed syllables etc. Although this annotation is manual, ANOPHONE provides a user-friendly interface and makes the entire annotation process simpler and faster minimizing annotation errors.

¹ <https://anophone.evetechnology.cz>

The application is divided into an administrator section and a user section. The user section allows annotators to browse the recordings database, including metadata and information of the final annotation for the individual annotation tasks and recordings. These recordings can also be filtered by metadata using regular expressions.

Fig. 1 shows the most important part of the application, the annotation interface itself (with the already mentioned BV1 annotation):

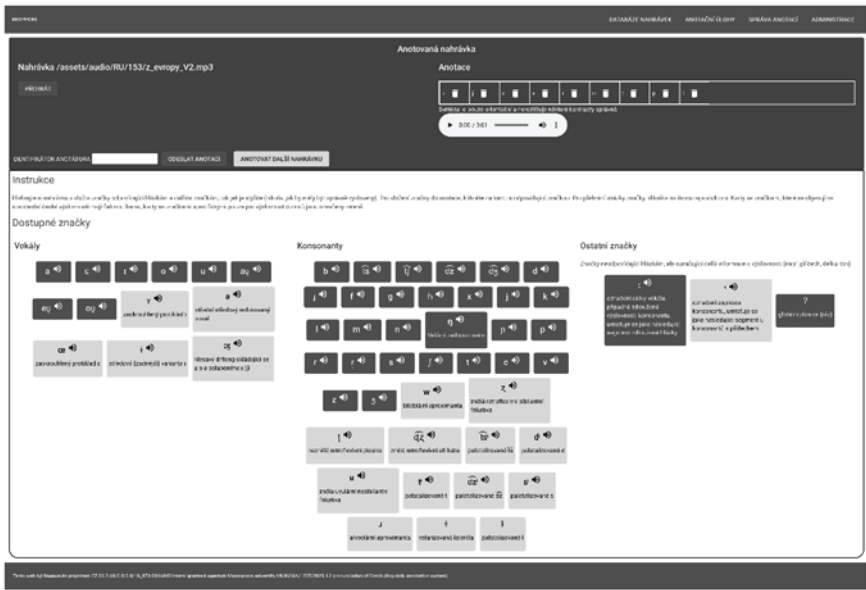


Fig. 1. ANOPHONE: Interface for annotation of recordings

Explanatory notes: anotate = annotation; (anotovaná) nahrávka = (annotated) recording; přehrát = to play; identifikátor anotátora = annotator's identifier; odeslat anotaci = to send an annotation; anotovat další nahrávku = to annotate another recording; dostupné značky = available labels; vokály = vowels; konsonanty = consonants; ostatní značky = other labels.

The lower part of the interface contains annotation instructions and the individual (available) tags. These tags (or labels) are divided into three sections: vowels, consonants and other tags (labels). Standard speech sounds (L2 language inventory) are marked in green (here in light gray) and non-standard sounds are in purple (here in dark grey). In most cases, the labels are accompanied by sample recordings (see speaker icon) or a description of the speech sound to facilitate orientation during annotation (to simplify the annotation process).

Both the recording and the current annotation are located at the top of the interface. The upper left corner contains the name, the recording's path within the source data and a button for playing back the recording being annotated. The right

side contains the current sequence of labels in the annotation. After clicking on a label card, the selected label is added to the end of the sequence corresponding to the annotation of the current segment. These labels can also be removed from the sequence by clicking on the bin icon. Below the current annotation is the experimental option to play back the entire sequence of tags using simple speech synthesis. In the lower left part, there is a box for entering the annotator's identifier (the annotation cannot be saved without it being filled) and buttons for saving the annotation and for moving on to annotating the next recording.

Recordings for annotation are automatically selected from the provided database based on the number of already existing annotations for a given segment within the currently selected annotation task. A random recording is always selected from the recordings with the fewest annotations. A single recording can be used in multiple annotation tasks, while the annotations are always saved for the corresponding annotation task. Before beginning the annotation, the user chooses from among the available annotation tasks.

ANOPHONE also allows (within the administrator section) to edit available labels, add recordings or create new annotation tasks. The editing window essentially allows us to define all aspects of the annotation task. Specifically, we can select a subset (or all recordings) from the entire database of recordings based on set parameters, which will be annotated as part of the annotation task, specify instructions and define labels available for annotation (labels can optionally be divided into groups named by the user). The colour of the label card and the text colour of the given label can also be defined. An example of annotation task administration can be seen in Fig. 2. For a specific annotation task, labels are selected from a database of all available labels, which can also be edited by the administrator (see Fig. 3). In this case, the administrator sets the label identifier, name, description, and path to the recording.

ANOPHONE
DATABÁZE NAHRÁVEK
ANOTAČNÍ ÚLOHY
SPRÁVA ANOTACÍ
ADMINISTRACE

Administrace

Datové sady / Tasks / Anotování fonémů

Anotační úlohy properties

Identifikátor úlohy *

phoneme-annotation

Jméno úlohy *

Anotování fonémů

Instrukce *

Přehrajte si nahrávku a vložte značky odpovídající hláskám a dalším značkám, tak jak je slyšíte (nikoliv, jak by měly být správn

☒ Syntéza řeči pro labely *

Sekce

Sekce 1 properties

Název sekce *

Vokály

Dostupné značky + Značka

Značka 1
Značka 2
Značka 3
Značka 4
Značka 5
Značka 6
Značka 7
Značka 8
Značka 9
Značka 10
Značka 11
Značka 12
Značka 13

Značka 1 properties

Odpovídající značka *

a

Barva značky *

Barva textu *

🗑 Značka ↓

🗑 Sekce ↓

Sekce 2 properties

Název sekce *

Konsonanty

Fig. 2. ANOPHONE: Project administration

Explanatory notes: **Header labels** – databáze nahrávek = database of recordings; anotační úlohy = annotation tasks; správa anotací = annotation administration.

Interface – administrace = administration; identifikátor/jméno úlohy = task identifier/name; instrukce = instructions; název sekce = section name; (odpovídající) značka = (corresponding) label; barva značky/textu = label/text colour.

ANOPHONE
DATABÁZE NAHRÁVEK
ANOTAČNÍ ÚLOHY
SPRÁVA ANOTACÍ
ADMINISTRACE

Administrace

Datové sady / Labels / ts

Značka properties

Název *

Popis značky properties

Popis

Ilustrační nahrávka (URL)

Identifikátor značky *

Metainformace

+ Atribut

✓ ULOŽIT

Tento web byl financován projektem CZ.02.2.69/0.0/0.0/19_073/0016943 Interní grantová agentura Masarykovy univerzity MUNI/IGA/1225/2020, L2 pronunciation of Czech

(linguistic annotation system)

Fig. 3. ANOPHONE: Tag (label) administration

Explanatory notes: administrace = administration; datové sady = datasets; značka = label (tag); název = name; popis = description; ilustrační nahrávka = sample recording; identifikátor značky = tag identifier; metainformace = metadata.

The final relevant function of ANOPHONE available in the administration section is annotation management. In addition to browsing recordings and manual simplified annotation, it also allows for viewing all assigned annotations for a given recording within all annotation tasks. This part of the application also provides functions for setting and editing the final annotation, by selecting from different annotations of the same segment (speech sound, word, phrase), which is then displayed in the recording database. Annotations are grouped by recording and can again be filtered by meta-information. The interface of this screen is shown in Fig. 4.

ANOPHONE
DATABÁZE NAHRÁVEK
ANOTAČNÍ ÚLOHY
SPRÁVA ANOTACÍ
ADMINISTRACE

Filtrování nahrávek

language
speaker
repeating
word
Filtrovat

/assets/audio/UKR/157/lovit_v1.mp3
Atributy | language:ukrainian | speaker:157 | repeating:true | word:lovit
MD2000: l o v i t
Finální anotace: + l + o + v + t + ULOŽIT

/assets/audio/UKR/157/zdimat_v1.mp3
Atributy | language:ukrainian | speaker:157 | repeating:true | word:zdimat
22010: z j i : m a t
navrv: z j i : m a t
Finální anotace: + z + j + i + : + m + a + t + ULOŽIT

/assets/audio/UKR/157/věc_v2.mp3
Atributy | language:ukrainian | speaker:157 | repeating:false | word:věc
vyskma: v j e t s
glabas: v j e t s
Finální anotace: + v + j + e + t + s + ULOŽIT

/assets/audio/UKR/157/zhořet_v1.mp3
Atributy | language:ukrainian | speaker:157 | repeating:true | word:zhořet
Vráňová: z h o : r e
Finální anotace: + z + h + o + : + r + e + ULOŽIT

/assets/audio/UKR/157/lano_v2.mp3
Atributy | language:ukrainian | speaker:157 | repeating:false | word:lano
526303: l a n o
olbret: l a n o
Finální anotace: + l + a + n + o + ULOŽIT

/assets/audio/UKR/157/symfonie_v1.mp3
Atributy | language:ukrainian | speaker:157 | repeating:true | word:symfonie
526936: s i n f o n i j e
Finální anotace: + s + i + n + f + o + n + i + j + e + ULOŽIT

/assets/audio/UKR/157/útok_v2.mp3
Atributy | language:ukrainian | speaker:157 | repeating:false | word:útok
pospve: u : t o : k
sosnovak22: ? u : t o : k
Finální anotace: + u : + t o : + k + ULOŽIT

/assets/audio/UKR/157/na_obraze_v2.mp3
Atributy | language:ukrainian | speaker:157 | repeating:false | word:na obraze
E7: n a o : b r a z e
Finální anotace: + n + a + o : + b + r + a + z + e + ULOŽIT

/assets/audio/UKR/157/v_ulici_v2.mp3
Atributy | language:ukrainian | speaker:157 | repeating:false | word:v ulici
496977: f u l v c i
Finální anotace: + f + u + l + v + c + i + ULOŽIT

/assets/audio/UKR/157/sbírka_v1.mp3
Atributy | language:ukrainian | speaker:157 | repeating:true | word:sbírka
K4321: s b i r k a
Finální anotace: + s + b + i + r + k + a + ULOŽIT

Fig. 4. ANOPHONE: Annotation management
Explanatory notes: filtrování nahrávek = filtering of recordings; atributy = attributes; finální anotace = final annotation.

3.1 Data Annotation Process Using ANOPHONE

During the annotation process using BV1 tagset, it is necessary to create appropriate data structures in ANOPHONE for the individual recordings (complete with metadata) and import them into the tool along with the recordings themselves. Furthermore, one needs to create data for individual labels of BV1: vowels, consonants, and other tags.

The subsequent annotation of the data prepared in the system typically occurs in two rounds. The first, „draft“ version of the annotation, where there are several annotations from different annotators for each recording (segment), is followed by the second phase, in which the final annotation of the recording is determined (and optionally modified) from the available annotations. These resulting annotations are subsequently exported from the ANOPHONE system and, using Python scripts (created by the authors of this text), converted to files named after the individual recordings containing the corresponding annotation. The annotation files processed in this way along with the recordings (both files share the same name and differ only in the file extension: *.label* for annotations and *.wav* for recordings) were subsequently used as input for machine learning (ASR model training). The output is always a recording and a text file containing the relevant annotation (with labels separated by spaces). This format matches the input data requirements of the speech recognition (ASR) library Persephone (Adams et al. 2018), which we used to train all speech recognition models to create the tool for automatically transcribing the speech of non-native speaker of Czech. (A comparison of the results of the individual models is presented in Holaj – Pořízka 2021 and Holaj 2023.)

Based on a comparison of the fully manual annotation process using a text editor (and the AV1/AV2 attributive tagsets, see Holaj – Pořízka 2021) with the annotation process using ANOPHONE, it can be said that ANOPHONE provides a significant simplification and acceleration of the annotation process. Moreover, compared to text editor-based annotation, typing errors in the annotation are minimized. Thus, the tool proved to be very successful in testing the annotation process of recording non-native speakers of Czech with BV1 annotation.

4 CONCLUSION

The main goal of this contribution was to present the ANOPHONE annotation system, a flexible tool that allows for sound segments to be uploaded into a database for further processing, i.e., speech data annotation. It also serves as a database of recordings with their metadata. Additionally, it enables not just data annotation, but also creating or defining sets of annotation tags or modifying them to suit the users' needs and applying these tags onto data (sound segments) through a browser-based and freely available interface. ANOPHONE also allows users to manage multiple annotation tasks for available recordings or to create new

annotation tasks, independently of the existing ones. This language-agnostic (i.e., relatively universal) tool can thus be used by other researchers for other, similarly oriented projects, helping them to improve and speed up the process of annotating speech data.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic MUNI/IGA/1225/2020 “L2 Pronunciation system (Linguistic Annotation System)”.

References

- Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In LREC 2018 (Language Resources and Evaluation Conference), Japan, pages 3356–3365.
- ANOPHONE application. Accessible at: <https://anophone.evetechnology.com/>.
- Arora, V., Lahiri, A., and Reetz, H. (2018). Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America*, 143(1), pages 98–108. Accessible at: doi 10.1121/1.5017834.
- Becker, K., and Edalatshams, I. (2019). ELSA Speak – Accent Reduction [Review]. In *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, Iowa State University, pages 434–438.
- Blake, R. (2011). Current Trends in Online Language Learning. *Annual Review of Applied Linguistics*, 31, pages 19–35. Accessible at: doi 10.1017/S026719051100002X.
- Derwing, T., and Munro, M. (2015). *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. John Benjamins. Accessible at: doi 10.1075/Illt.42.
- Duolingo Inc. The free, fun, and effective way to learn a language! Accessible at: <https://www.duolingo.com/>.
- ELSA Corp. Meet ELSA – Your personal AI-powered English speaking coach. Accessible at: <https://elsaspeak.com/>.
- Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning (ICML ‘06)*, New York, USA, pages 369–376. Accessible at: doi 10.1145/1143844.1143891.
- Holaj, R. (2018). *Počítačová podpora osvojování české ortoepické výslovnosti*. Master thesis. Masarykova univerzita, Brno. Accessible at: <https://is.muni.cz/th/k5tej/>.
- Holaj, R. (2023). *Nástroj pro automatickou transkripci řeči nerodilých mluvčích češtiny*. Dissertation. Masarykova univerzita, Brno.
- Holaj, R., et al. (2021). CzechME. Accessible at: <https://www.phil.muni.cz/en/research/publishing-and-editorial-activities-of-the-faculty/overview-of-publishing-and-scientific-activities/1778138>.

Holaj, R., and Pořízka, P. (2021). L2 Czech Annotation for Automatic Feedback on Pronunciation. *Jazykovedný časopis*, 72(2), pages 510–519. Accessible at: doi 10.2478/jaz-cas-2021-0047.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, 27, pages 184–202. Accessible at: doi 10.1017/S0267190508070098.

Nováčková, S. (2017). *Nové aplikace*. Unpublished. Masarykova univerzita, Brno.

Nováčková, S. (2018). *Návrh cvičení k výuce češtiny pro cizince přes mlearning*. Accessible at: <https://is.muni.cz/th/m0q42/>.

Veroňková, et al. (2022). *ProCzeFor*. Accessible at: <https://proczefor.cz>.