

Marta Šimečková

Czech Language Institute, Czech Academy of Sciences
simeckova@ujc.cas.cz

Martin Karafiát

Faculty of Information Technology, Brno University of Technology
karafiat@fit.vut.cz

Oldřich Plchot

Faculty of Information Technology, Brno University of Technology
iplchot@fit.vut.cz

USING MACHINE LEARNING FOR AUTOMATIC DIALECT DETECTION. NEW METHODS IN CZECH DIALECTOLOGY¹

1. Introduction

Applied dialectology is being developed in relation to dialect coaching, natural language processing, or forensic dialectology especially in some Anglophone countries (e.g. Nolan et al. 2009; Köster et al. 2012; Watt 2018), whereas in the Czech context, it represents a novelty. However, with current technological advancement, opportunities for cooperation between dialectologists and natural scientists are opening up.

The significant potential of such collaboration is particularly evident in advances in machine learning, which is considered part of artificial intelligence. Machine learning enables recognition, classification, segmentation, or analysis of a large volume of data. The processed data can take various forms, including linguistic data, which may be textual or audio-based. In the field of linguistics, software systems for automatic language detection or automatic transcription of spoken languages (and dialects) are being developed. Such systems can be invaluable tools for linguists and other professionals working with spoken language.

This chapter focuses on the creation of a specific software developed within the framework of the project ‘Language Memory of the Regions of the Czech

1 The chapter was supported by the Czech Ministry of Culture, NAKI III, project ‘Language Memory of the Regions of the Czech Republic’, No. DH23P03OVV010.

Republic'. In this project, machine learning methods are employed for the preservation, documentation, and presentation of Czech language dialects, involving collaboration among three institutions: the Czech Language Institute of the Czech Academy of Sciences (CLI CAS), Brno University of Technology, and Palacký University Olomouc. The developed software will serve as a detector adapted for the identification of Czech language dialects based on audio recordings.

The first part of the chapter gives a detailed account of the data source, crucial for the development of the dialect recognizer software – the Archive of Dialect Audio Recordings, deposited in the Dialectology Department of CLI CAS. The second part briefly outlines the history of research on automatic language identification. The following sections are focused on the progress of work on the software and highlight selected results. Given the objectives being pursued, this chapter extends into the realm of engineering, showcasing the potential of applied dialectology and interdisciplinary collaboration.

2. The Archive of Dialect Audio Recordings

2.1 The history of the Archive

One of the main tasks of the Dialectology Department is the systematic documentation of the Czech language dialects. The most accurate way of doing this is to make audio recordings. The initial attempts commenced in the 1950s.

In the 1960s and 1970s, systematic efforts to create the actual archive began. At that time, intensive field research was conducted to gather linguistic material for the upcoming atlas of Czech language dialects (*Czech Language atlas*, 1992–2011; the updated version is available on the website cja.ujc.cas.cz/e-cja). Thanks to this initiative, authentic dialectal speech recordings were collected, covering the entire Czech-speaking territory and selected areas abroad in Czech language enclaves in Poland, Romania, and former Yugoslavia. Recordings were made in approximately 300 different locations within a relatively short timeframe, representing the linguistic variation of the Czech national language at that time.

The recordings primarily captured the speech of rural speakers using an archaic dialect layer. These speakers were from the older generation aged 65 and above, who were born and raised before World War I. In the surveyed villages, a questionnaire survey for the aforementioned atlas was conducted first, followed by the recordings of dialect speech. Selected speakers were prompted to narrate about life in the village, local customs, regional cuisines, vanishing crafts, etc. (Šimečková 2024).

In subsequent years, recording and enrichment of the archive continued. However, compared to the era of nationwide research in the 1960s and 1970s, these were more

localized actions. Areas for these verification surveys were usually chosen according to the needs of ongoing tasks of the Dialectology Department, especially in relation to the emerging *Dictionary of Czech Language Dialects* (2016–).

In 2023, a new phase of intensive research started, financially supported by the mentioned project. Within the first year of the project, 85 surveys were conducted, enriching the archive with over 300 new recordings.

The archived records are continually expanded through collaboration with schools, and some recordings were acquired through public donation or through open science projects (e.g. the project ‘Become a superdialectologist!’, see Fodor 2023).

2.2 Digitization of the sound archive and cataloging

The oldest recordings were made on reel-to-reel tapes. Since the 80s, cassette voice recorders have become the prevalent choice, offering easier handling and greater accessibility compared to the older reel-to-reel tape recorders. Later, it became necessary to transfer the old recordings from magnetic tapes to more modern audio media. In collaboration with Czech Radio, all the recordings were digitized. These digital copies were stored on compact discs labeled with basic annotations, totaling 460 physical units. However, not all the older recordings underwent this process successfully; some reels and tapes were in such poor technical condition that the transferred sound was assessed as incomprehensible and distorted, rendering them unsuitable for further research. Presently, dialectologists exclusively work with digitizations stored on several external hard drives. Contemporary research involves the use of digital voice recorders or occasionally recorders built into mobile phones with lossless WAV recording, enabling high-quality archiving of audio data.

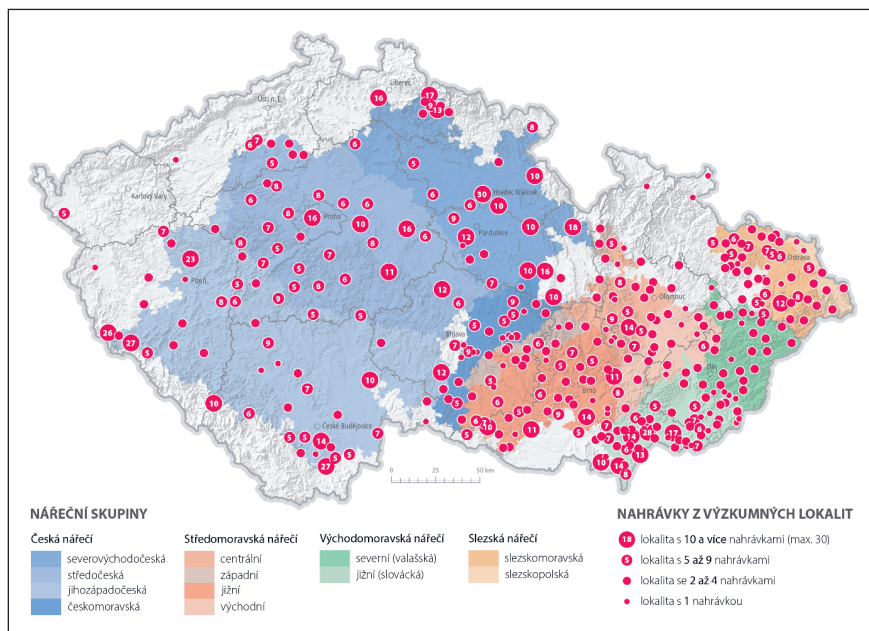
In 2022, Marta Šimečková became the head of the archive. Under her leadership, comprehensive cataloging of digitizations was performed, leading to the creation of the electronic *Database of Dialect Speeches*. Now, this Database records 1741 audio files with a total duration of 620 hours, averaging 22 minutes and 15 seconds per recording. Detailed and standardized descriptions are created for all recordings in the database. Cataloging entries include the name of the location, recording date, explorer’s name, speaker-related metadata (birthplace, residence, year of birth, gender, occupation), or information regarding the recording’s content and language. Names of recorded subjects are anonymized in accordance with the General Data Protection Regulation (GDPR).

2.3 Language of recordings and Czech language dialects

The archive contains recordings of the Czech language dialects, showcasing varying degrees of their preservation. Some recordings, especially from the earlier phase of dialectal research, reflect traditional dialects with relatively rigid structures. In new recordings, the tendency to standard or semi-standard language is noticeable. Nevertheless, even in these instances, the speech retains regional language characteristics. Only a handful of recordings can be identified as non-dialectal, semi-literary, or entirely standard.

For machine learning purposes, it was crucial to assign recordings to specific dialect types unequivocally. A code list was created to categorize archive recordings into specific dialects. The basic hierarchy consists of four dialectal groups: Czech dialectal group (1), Central Moravian (2), Eastern Moravian (3), and Silesian (4). A special class includes recordings from non-dialectal locations (5), such as border areas and former German-speaking enclaves, along with recordings from foreign locations (6) and recordings with a non-dialectal or mixed speech style (7). These groups were further divided into smaller units – dialectal subgroups, sections, and types. A total of 49 dialectal units were defined within groups 1 to 4.

Geographically, recordings from the Moravian and Silesian regions predominate. In contrast, the documentation of the Central Bohemian region is limited



Map 1: Geographical distribution of audio recordings deposited in the Archive of Dialect Audio Recordings (24 September 2023)

(see Map 1). This was due to an earlier belief that this area was linguistically uninteresting – traditional dialects here largely transitioned into interdialectal common Czech. Only sporadic recordings were made in the past in border areas, which were excluded from systematic dialect research due to becoming resettled territories after 1945.

3. Introduction to automatic dialect detection

Research in automatic dialect identification of spoken audio utterances is not an area with a high focus in speech studies, therefore not too many resources are available. Fortunately, the dialect identification can be treated as a sub-task of language identification (LID). The use of this technology on English dialects is well analyzed in Etman and Beex (2015). Similarly, Ali et al. (2016) applied LID techniques on Arabic dialects. Both studies showcased these techniques performing well. Thus, in this chapter, we aim to adopt and explore LID methodologies for the automatic detection of Czech dialects. Our team possesses extensive expertise in LID, consistently yielding commendable results in global assessments sponsored by NIST.

3.1 Brief history of automatic language identification

To better understand our system's functionality, we would like first to present the evolutionary trajectory of these techniques, spanning from their inception in the 1990s to the present day. Furthermore, there exists the possibility that some of the currently obsolete approaches, such as phonotactics, could be revived for potential application in dialectal analyses in the future.

During the period **from the 1990s to the early 2000s**, spectral feature sequences were used to train a statistical generative model, specifically the Gaussian Mixture Model (GMM), as outlined by Zissman (1996) and Torres-Carrasquillo et al. (2002). The process of creating an LID system based on GMMs can be put simply as:

1. Initially, a general GMM underwent training encompassing all data across various languages. The model is called the Universal Background Model (UBM), and it represents collective data distribution. For simplicity, only the means from all Gaussians were extracted and concatenated into so-called supervectors.
2. Then, the model underwent further adaptation using specific language data from the training set, thereby generating language-specific models (GMMs).
3. During recognition, an unknown speech utterance represented by a sequence of spectral features is classified by computing the log-likelihoods (which serve as scores) of all language-specific models (GMMs) given this sequence.

Between 2003 and 2009, the development of a new generation of phoneme recognizers based on Neural Networks (NNs) started. Utilization of straightforward statistical modeling (counting) of individual phonemes (Jayram et al. 2002; Pellegrino et al. 1999) was always a widely used approach to LID, and phoneme transcriptions were often obtained from Hidden Markov Model-based Automatic Speech Recognition (ASR) systems. In Zissman and Singer (1994), six phoneme recognizers running concurrently were used. Each of these recognizers generated a language-specific likelihood based on a phonotactic model. The final score was obtained by averaging of corresponding likelihoods in logarithmic space. This is known as Parallel Phone Recognition, followed by Language Modeling (PPRLM). Later, the NN-based phoneme recognizers became very efficient (Schwarz et al. 2004) and the PPRLM phonotactic system was often used as a complementary technique to the previously presented GMM-UBM system.

Between 2012 and 2018, the i-vector approach transformed the high dimensional UBM-GMM supervector into a subspace vector known as an i-vector, characterized by a considerably smaller dimensionality, approximately 600 dimensions (Dehak et al. 2011).

The classifiers based on the i-vectors (Martínez et al. 2011) surpassed the phonotactic approach even without any extraction of explicit phonetic information.

Since 2018, the NN embeddings, initially introduced in Snyder et al. (2018) for speaker recognition tasks, resulted in models producing highly informative features for speaker classification where the phonetic information was implicitly encoded. The idea was based on processing the feature vectors through NN layers, incorporating a distinctive “special” pooling layer. This layer accumulates statistics from input

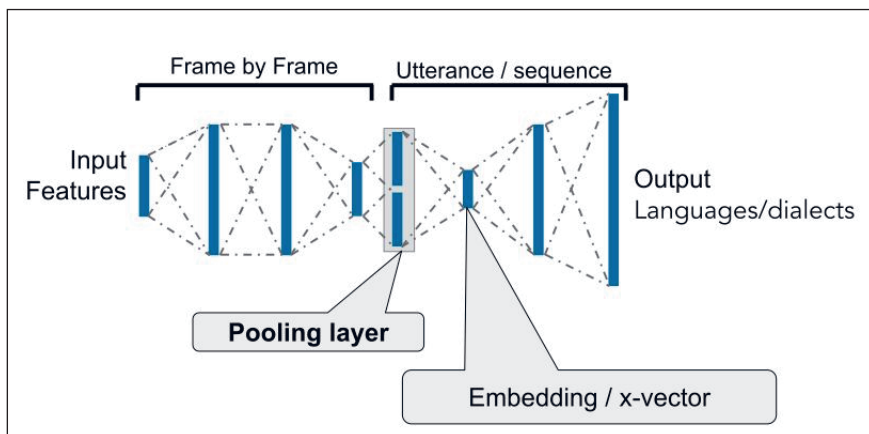


Figure 1: The x-vector extractor Neural Network

audio data of variable duration, typically spanning an entire utterance, and outputs mean and variance vectors. Subsequently, these statics undergo transformation via a series of affine layers; the vector from either of these layers is an embedding that is often called x-vector.

The final layer represents the classes, originally targeting speakers but adapted to our task, we simply use language labels. The whole architecture is illustrated in Figure 1.

During the NN training, the language of each utterance is known, and we can train the whole NN by optimizing the cross-entropy. Consequently, information crucial for the final language classification task is embedded in the ensuing layers and therefore also in extracted x-vectors. These vectors alone can be used as features for general language classification purposes, even if the languages in the x-vector training data differ from the target languages/dialects.

4. Our dialect detection system

4.1 Pre-processing of the data

First, the audio data are transformed into a flow of spectral feature vectors mentioned above in the GMM-UBM technique (see 3.1). It is based on the presumption that the human vocal tract can be considered static in a short time window of length about 20–25 ms. Therefore, the acoustic signal is cut into 20 ms segments with a hamming window (to suppress edges). The segments are typically overlapped by 10 ms to get more smooth time development. Short-time Fast Fourier Transform is applied to estimate spectra and further averaged by triangular filter distributed in the Mel-frequency domain. The resulting Mel-filterbank features reflect the fact that the human ear is more sensitive to lower frequencies than to higher and speech is generated to be heard. Conversion into log-domain and Discrete Cosine Transform

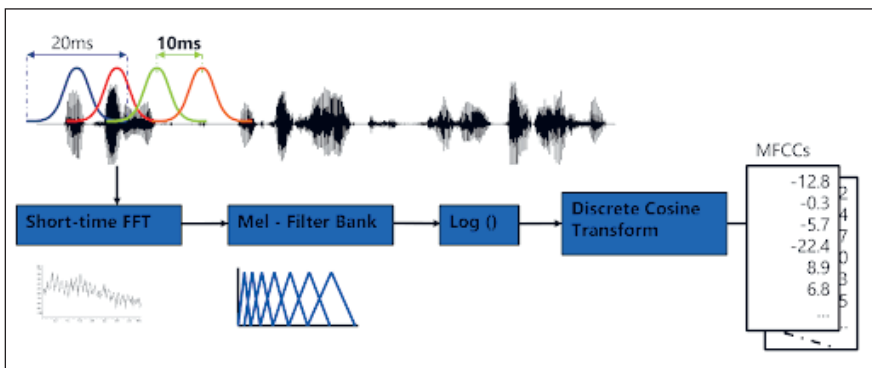


Figure 2: Feature extraction for NNs

generates final features well known as Mel-Frequency Cepstral Coefficients in the speech processing field. The whole process is outlined in Figure 2.

4.2 Embedding vectors

Our system is x-vector based as it is giving state-of-the-art performance nowadays. But we also trained the traditional i-vector based system for comparison. Our i-vector system represents a technology that has been state-of-the-art until c. 2017 for language identification. The i-vector approach was applied on top of information-rich features extracted from a NN that was trained on multiple languages. These features are called bottleneck features (Kramer 1991) and are typically extracted from a narrow layer of NN. They have brought speech signal parametrization to a quantitatively different level (Grézl et al. 2007) as they convey information about phonetic content in a nonlinearly compressed form which can be directly used for the task of language identification, where they have demonstrated state-of-the-art performance (Matejka et al. 2014; Jiang et al. 2014; Ferrer et al. 2016).

The x-vector extractor was trained on 107 languages from the VoxLingua107 database (Valk – Alumäe 2021). Note, that no Czech dialectal data were visible during the x-vector training.

The backbone architecture of our x-vector extractor is derived from the standard computer vision model suggested by He et al. (2016). Unlike the original ResNet34, the first convolution has a kernel size of 3 and a stride of 1. The stages of the network have 64, 128, 256 and 256 channels. 64-dimensional Mel-filterbank features constitute the input to the model, where window length and shift are 25 ms and 10 ms, respectively. We limit the input training example length to 300 frames. After the initial frame-wise processing, internal features are aggregated with statistics (mean and standard deviation) pooling. Statistics are projected to 256-dimensional embeddings. The models were trained for language classification on 8-GPU compute nodes by stochastic gradient descent with a momentum of 0.9. We regularized the models with a weight decay with a factor of $1e-4$. Each minibatch comprised 64 examples. During the initial warm-up stage (first 10k training steps), the learning rate increased from 0 to 0.2. Subsequently, it was multiplied by a factor 4 of 0.5 every time a plateau on a cross-validation loss was reached. The cross-validation set was derived from the training set. The network was trained to minimize Additive Angular Margin loss (Deng et al. 2019) with a margin set to 0 and a scale to 30.

After the training, the Czech dialectal data was processed through the x-vector extractor. Each recording was represented by a single embedding x-vector, which was further used for the language detection system.

4.3 LID system

As a backend for our experiments, we use Gaussian Linear Classifier (GLC) (Martínez et al. 2011). This model assumes that the embeddings for each dialect follow Gaussian distribution. Each Gaussian is assumed to have a language-dependent mean vector, and fixed covariance matrix shared across all languages. This model is trained on the held-out training set extracted from the collection of the Czech dialectal data. In our experiments, we do not apply any pre-processing to the embeddings: they are used as they come out of the embedding extractor. Notice that the extractor produces already length-normalized embeddings.

4.4 Displaying of x-vectors in 2D space

Plotting the analyzed vectors in standard figures can be useful to show how distant dialects are from each other in the x-vector space and how well they are separated. To be able to do it, the high dimensional (256 dim.) x-vectors need to be projected into two-dimensional space. The most common techniques are: t-distributed stochastic neighbor embedding commonly used for visualizing high-dimensional data and Linear Discriminant Analysis (LDA) which converts the data into space where the classes (dialects) are better separated. We experimented with both, but LDA produced slightly better results. Labeled projections of the x-vectors into the first two LDA bases are presented in Figure 3. It clearly shows the bigger distance of the Bohemian dialects group from others. Moravian dialects are closer to each other with strong overlap, but the centroids are shifted showing average distances between the dialect groups.

5. Results

For testing purposes, we randomly selected 30 recordings from each main Czech language dialects group and 15 recordings from each subgroup (for the hierarchy of Czech language dialects, see section 2.3). The rest of the data was used for the training. Note, that the whole recordings are typically the interviews between the target subject (dialect speaker) and the interviewer (researcher, dialectologist, non-dialect speaker). These non-dialect parts could cause data contamination, therefore, some automatic clustering methods based on speaker change detection would be useful. Fortunately, the amount of interviewer data is minor to interviewee one, and therefore, we are ignoring this effect so far. We plan to experiment with automatic data cleaning in the near future.

Our automatic Dialect ID system is based on a GLC classifier trained on top of x-vectors from the training part of the data. The extractor is described above in section 4.2. Note, no Czech dialectal data was used for the extractor training and all classification burden is put on GLC shoulder.

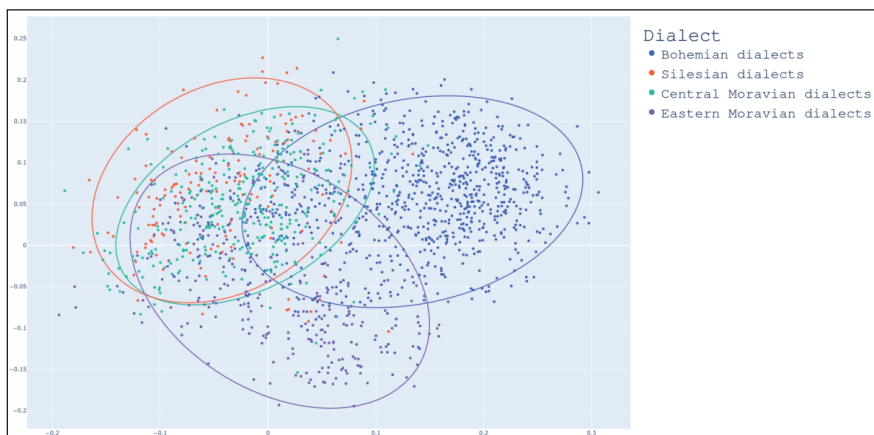


Figure 3: X-vectors projected in LDA space

The final results are shown in Table 1. The results demonstrate significant gains over standard i-vector based systems with reasonable accuracy of over 80% in the main dialectal groups (groups 1–4) and almost 70% in the dialectal subgroups.

	4 classes (groups) [%]	13 classes (subgroups) [%]
Baseline i-vectors	81	68
x-vectors	86	60

Table 1: Accuracy of dialect detection system

5.1 Confusion matrix

To be able to better analyze and understand the errors of our Dialect ID system, we plot the confusion matrix of the system results in Figure 4. The rows present a ground-true dialect identifier and columns the system output. Therefore, the central diagonal shows the number of correctly recognized data, and the out of diagonal shows the count of errors. Interestingly, it presents that most of the errors lie inside of the main dialectal groups, therefore the errors are caused mainly by disfluencies between dialects from the same groups.

6. Conclusions and future work

Due to the significant costs associated with collecting and annotating training linguistic data, the dialectology has practically minimal support in modern artificial intelligence and machine learning technologies, primarily represented by automatic speech recognition. This situation, at least in the Czech context, might be changed with the new project ‘Language Memory of the Regions of the Czech Republic’.

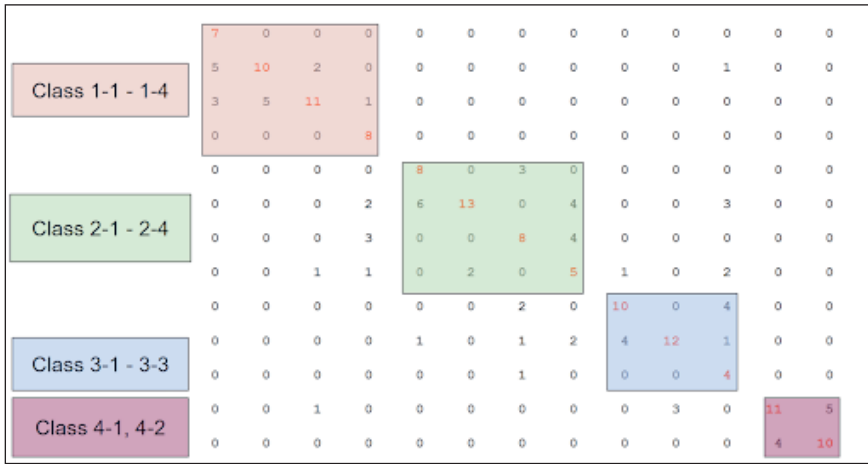


Figure 4: Confusion matrix of automatic dialect identification system

This project is based on a collaboration between dialectology, machine learning, and geoinformatics experts. The Archive of Dialect Audio Recordings, owned by the Dialectology Department of CLI CAS, serves as a data source for software development capable of dialect detection from audio recordings. The outcome will be a speech recognizer adapted for automatic detection of Czech language dialects and regionally accented speech based on audio recordings.

The initial outcomes of the dialect detection system, based on i-vectors and x-vectors, are presented in this chapter. The results follow expectations from the language identification field, therefore the x-vectors were found superior to i-vectors, but both methods were giving reasonable performance on the main dialect groups as well as also on dialect subgroups ones.

Developed dialect detectors have the ability to produce not just one best detection but also distances from all other tested dialects. It could be a useful measure of newly analyzed linguistic data coming in future.

In further development, we are planning to focus our work into these tasks:

1. Employing automatic segmentation of the data into the interviewer/interviewee could be used to clean the data.
2. X-vector extractor could be fine-tuned on the dialectal data as the current extractor is purely trained only on language identification data.
3. Further improvement can be achieved by using large pre-trained models.

These results demonstrate the applicability of the machine learning techniques in Czech dialectology and new approaches in the applied linguistics in general.