sciendo
DOI 10.2478/jazcas-2021-0046

MAPKA: A MAP APPLICATION FOR WORKING WITH CORPORA OF SPOKEN CZECH

HANA GOLÁŇOVÁ – MARTINA WACLAWIČOVÁ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

GOLÁŇOVÁ, Hana – WACLAWIČOVÁ, Martina: Mapka: A map application for working with corpora of spoken Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 502 – 509.

Abstract: A new interactive map-based web application named Mapka was published by the Institute of the Czech National Corpus in 2020. It aims to serve linguists, as well as schools and the general public, and it features various functions described in this paper. Mapka was designed as a supplement to the CNC spoken corpora, starting with the DIALEKT corpus (more to come in the future). Its main function is to display various types of territorial division (primarily in terms of dialect, but also administrative) and networks of localities associated with the corpus. The main dialect regions are provided with overviews of their typical dialectal features and two samples of dialectal discourse – one slightly historical and one contemporary. The application offers the possibility of searching for municipalities, plotting the points on the map and creating a custom map. The paper concludes with future prospects concerning an enhanced and improved version of the application.

Keywords: corpus, map, Czech language, spoken language, dialect

1 INTRODUCTION

In July 2020, the Institute of the Czech National Corpus published a new tool: a web application named Mapka [1]. It is an interactive map-based application primarily designed as a supplement to the spoken corpora of the Czech National Corpus, however it features various functions beyond this framework. The Mapka application is intended to serve both linguists and the general public. It is accessible without registration and it is available at: https://korpus.cz/mapka/.

Currently, its main function is to display the various types of language boundaries (and additionaly administrative borders) and nets of localities represented in the DIALEKT corpus ([2], [3], [4]) using an interactive map of the Czech Republic. The next phase, planned for the current year of 2021, will include (amongst other things) additional data from other spoken CNC corpora, e.g., ORTOFON [5] and ORAL [6]. The application includes presentations of characteristic features of the main Czech dialect regions illustrated by authentic speakers' utterances – slightly historical, as well as contemporary ones. Users are enabled to search for municipalities, add these points to the map, and create their own map.

The goal of this paper is to showcase the current version of the Mapka application, introduce its possible uses, and outline future prospects.

2 FEATURES AND FUNCTIONS OF THE MAPKA APPLICATION

2.1 Territorial division

The Mapka application displays various types of territorial division on the background map. The most important of these is the dialect-based territorial division of the Czech-speaking language territory, i.e., 10 regions (Central Bohemia, Northeast Bohemia, West Bohemia, South Bohemia, Bohemian-Moravian transient region, Central Moravia, East Moravia, Silesia, Bohemian borderland, Moravian and Silesian boderland) including the Bohemian borderland and the Moravian and Silesian borderland, although they do not belong to the group of traditional dialect regions. In this context, borderland refers to the historical area defined by the former numerical prevalence of the German-speaking population, massive population relocation after World War II, and the lack of a traditional Czech dialect substrate. Moreover, the Czech-speaking language territory is not only the territory of the Czech Republic, but a few localities – Czech language islands belonging to the dialectal region of Northeast Bohemia or Silesia – are located in Poland.

If needed, it is possible to choose a mode showing even more detailed dialect categories: dialect region / *nářeční oblast*, dialect subgroup / *nářeční podskupina*, dialect area / *nářeční úsek*, dialect type / *nářeční typ*. This detailed division is important mainly for the area of Moravian and Silesian dialects. Concerning the Bohemian territory, these detailed dialect categories can be found almost solely in the border areas of its dialectal regions. The Mapka application shows the exact position in the dialect system and a numeric code (used by Bělič [7]) for each dialect area of any type. The system of territorial dialect division employed in both the DIALEKT corpus and Mapka is based on Bělič's approach, *the Czech Linguistic Atlas* ([8], [9], [10]), *the Encyclopaedic Dictionary of Czech* [11] and its presentday online version: *The New Encyclopaedic Dictionary of Czech* [12].

Besides this, the map also provides an option of displaying the boundaries of the Czech Republic's administrative units, i.e., districts / *okresy* or regions / *kraje*.

The basic background map can be easily switched from to the relief map which, e.g., enables comparing natural and dialectal boundaries and discovering their connections.

The process of developing the application required a thorough revision of dialect regions delimitation, mainly particularization of their borders. It was based on a large amount of dialect monographs, studies, and consultations with dialectologists. Additional verification was made using the statistical lexicons of municipalities ([13], [14]). As far as the boundary delimination is concerned, problems of transition zones must be mentioned. In dialectology, the transition

zones are defined by a gradual decrease in the number of dialectal features typical for dialectal subgroup, area, or type. Classification of some municipalities can be difficult, for there can be a lack of certain features typical for one dialect area on the one hand, and on the other hand, features typical for the neighbouring area may be missing. For example, the municipality Brodek u Přerova belongs to the Core Central Moravian Subgroup (*hanácká*), however the phonetic changes y > e and u > o typical for the subgroup do not occur in the above mentioned municipality. In spite of this, it cannot be subsumed into another subgroup, as it does not evince appropriate dialectal features. In the future version of the application, this problem will probably be solved by graphic highlighting of the transition zones along the borders.

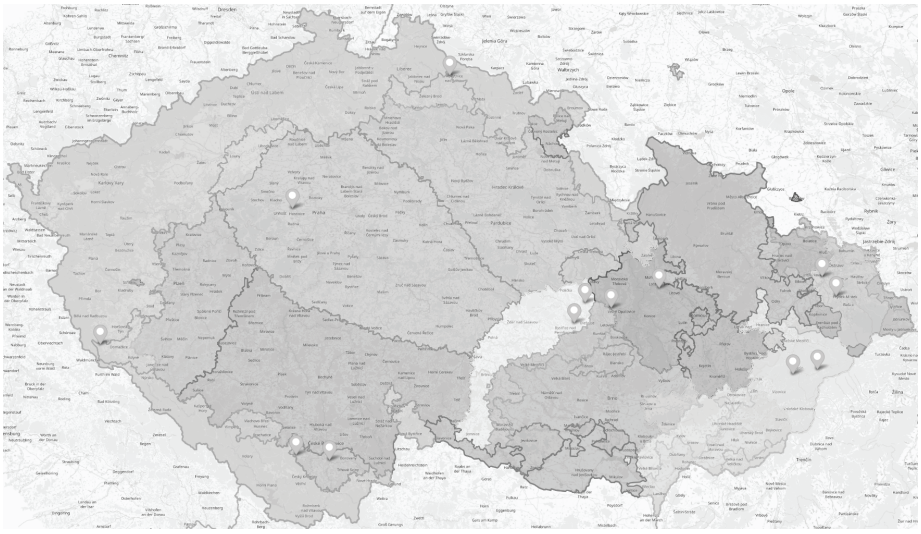

Fig. 1. Detailed dialect division of the Czech-speaking language territory. The samples of dialectal discourses are plotted as white pins

Currently, new types of territorial division based on historical data are being prepared to be added to the Mapka application. One of them is plotting the historical Bohemian-Moravian border and the Moravian-Silesian border, as they were formed at the end of the 12th century and stabilized in the 14th century. These borders have not been used by administrative authorities since 1948, nevertheless, they have a huge historical significance. Another new layer of the map will display the German language islands in the Czech Republic. These are eight areas where there was a historical numerical prevalence of the German-speaking population and the German language was used. They are located mainly in Moravia, some of them have

urban characteristics, some of them rural. They disappeared in the 20th century – partly in the first half of the century, partly after the World War II.

2.2 Overviews of dialectal features and samples of dialectal discourses

The Mapka application encompasses several overviews of typical dialectal features pertaining to the main three regions of the Czech Republic (Bohemia, Moravia and Silesia) and eight dialect regions of the Czech-speaking language territory (see above). The overviews are mainly focused on phonological and morphological features as they are fundamental for the dialect division. Examples of dialectal phenomena were primarily selected from the current version of the DIALEKT corpus, but some examples were supplementarily taken from transcripts which will be published in the upcoming version of the corpus.

Furthermore, each of the eight main dialectal regions is illustrated by two samples featuring authentic dialectal discourses chosen from the DIALEKT corpus. The samples consist of an audio recording and its two transcripts – dialectological (based on the Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak [15]) and orthographic. The samples were chosen in order to demonstrate the most typical dialectal features of the given regions. The recordings of the DIALEKT corpus are divided into two time strata, hence for each dialect region, one sample was selected from the older stratum of dialect material (the period between the late 1950s and the 1980s) and one from the newer one (from the 1990s up to the present day). The samples were chosen so that both recordings representing a particular region were made in the same municipality or a neighbouring location. This guarantees maximum comparability of discourses and users can trace changes of the local dialect in time. Each sample is followed by an analysis that describes relevant dialect features (phonological, morphological, syntactic and lexical) occurring in the discourse. In the future, samples for other prominent dialectal sections and types will be added. Our goal is to capture the variability of the traditional regional dialects as much as possible.

2.3 Mapka as a supplement to the DIALEKT corpus

While designing the Mapka application, the primary aim was to create a supplement for the spoken corpora of CNC, that would integrate data from these corpora with a map-based interface. For the time being, Mapka serves as a supplement for the DIALEKT corpus. Above all, municipality networks that have certain connections to the DIALEKT corpus can be displayed on the background map. For example, users can observe a network of municipalities where recordings included in the current version of the DIALEKT corpus were produced. It is possible to display all of the concerned municipalities or to choose either the network of localities, where recordings from the older time stratum were produced, or the network of localities bound with the newer stratum. Another option is to visualize

the network of all municipalities where overall data collection for this corpus (published recordings as well as unpublished so far) took place. A special bonus is the possibility to display the network of research localities of the Atlas of the Czech Language. All these networks can be visualized simultaneously and compared. Users can choose which one of them will be displayed above all the others. If needed, the work with the map can be restricted to a particular dialect region or regions and the others will not be considered.

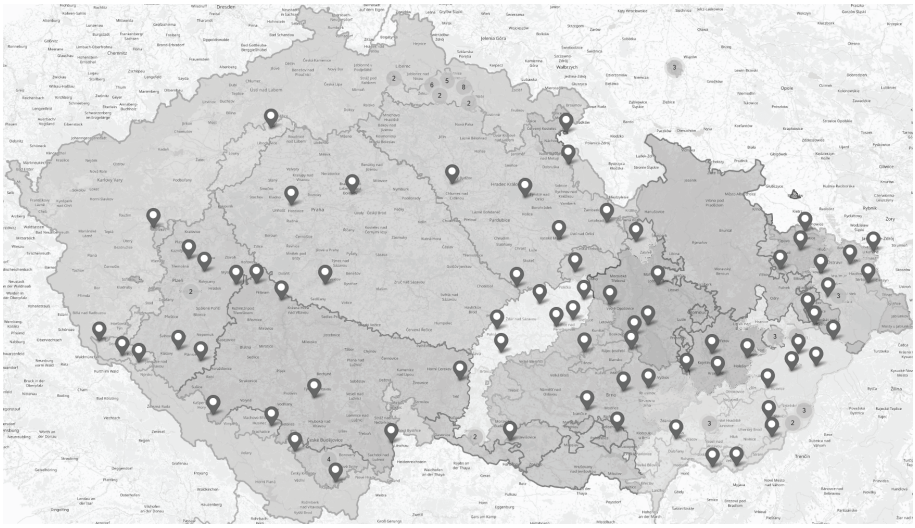

Fig. 2. Network of municipalities where overall data collection for the DIALEKT corpus took place. The dark grey pins containing numbers mark areas with a larger amount of points

2.4 Searching and creating your own map

The Mapka application offers the option to search for municipalities / parts of municipalities in the Czech Republic. The cadastral boundaries of the looked-up municipalities are visualized on the map, in order to be clear which parts belong to a certain municipality. Users can display information about the position of the municipality in the system for the division of dialectal regions.

The application enables users to proceed to plot these points on the map and create their own map. They can choose colours from the colour spectrum for differentiation of various groups of points.

The resulting maps can be downloaded and printed. In the future version, users will also be able to save their map with plotted points and continue working on it later, after loading the application again. We hope it will prove to be useful for linguists doing research or preparing their own map to illustrate their monographs or studies.

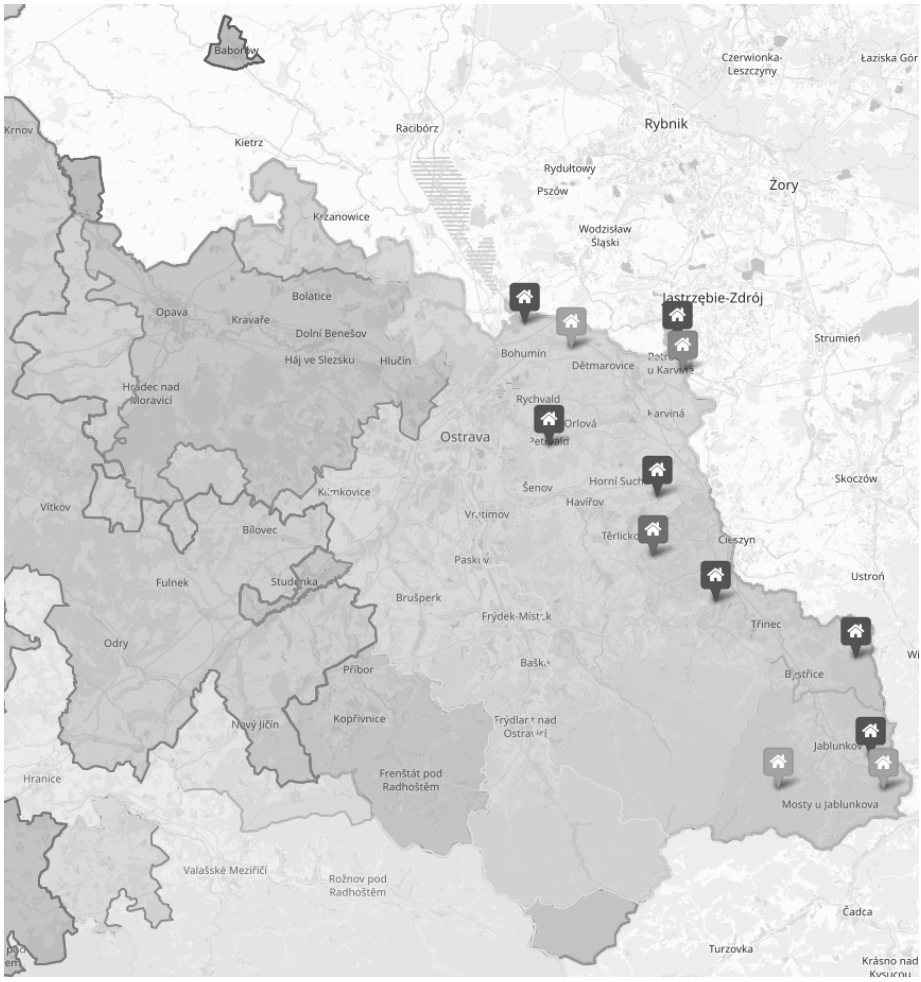

Fig. 3. Example of creating a map. Markers of four different colours identify four groups of locations in Silesian-Polish dialect subgroup. Each colour is bound with a certain dialect monograph and markers refer to the locations where language samples originated from

3 FUTURE PROSPECTS AND GOALS OF THE MAPKA APPLICATION

An enhanced version of the Mapka application is being prepared and hopefully will be published soon. Some of the planned innovations have been mentioned above, but in this section, we would like to sum up all future prospects related to Mapka.

As far as the DIALEKT corpus is concerned, new samples featuring authentic dialectal discourses for other prominent dialectal sections and types will be added to the application.

Considering the Mapka application was designed as a supplement to all of the spoken corpora of CNC, data from spoken corpora such as ORAL or ORTOFON will be included in the application prospectively. The data will encompass e.g. word counts, recording counts, statistics about the speakers, networks of locations of the speakers' childhood residence (until 15 years of age) or places of their longest residence, or networks of locations where recordings were made. Collections of data will probably be different for each corpus, since the corpora include different types of sociolinguistic metadata. We are planning on incorporating samples of authentic discourses chosen from the ORAL or ORTOFON corpora. The samples could be manually chosen and prepared or could be automatically generated random samples or both.

When speaking of creating a personalised map, the possibility to save the user's data will be integrated into the application.

We are considering creating an entertaining linguistic quiz focused on dialects or spoken language in general and its incorporation into the application. It could be attractive for the general public or schools.

The primary goal of the Mapka application is to capture the variability of spoken language across the Czech-speaking language territory. Taking the innovations planned for the next version of the application into account, we will be able to follow many particular aims such as to show differences between dialects captured by the DIALECT corpus and everyday spoken Czech language captured by other spoken corpora (e.g. ORAL and ORTOFON), differences between the language of the oldest generation of speakers and the other generations, between urban and rural speech or between monological and dialogical discourse. The application will help during the research of various aspects of spoken language, i.e. phonology, morphology, syntax, lexis, pragmatics or dialogue construction. The application is expected to serve language experts, all levels of schools as well as amateurs from the general public.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Goláňová, H., Waclawičová, M., and Pejcha, J. (2020). Mapka: Mapová aplikace pro korpusy mluvené češtiny. Version 1.0. Praha: ÚČNK FF UK. Accessible at: http://korpus.cz/mapka.

[2] Goláňová, H., Waclawičová, M., Komrsková, Z., Lukeš, D., Kopřivová, M., and Poukarová, P. (2017). DIALEKT: nářeční korpus, verze 1 z 2. 6. 2017. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[3] Goláňová, H. (2015): A new dialect corpus: DIALEKT. In K. Gajdošová and A. Žáková (eds.), Proceedings of the Eight International Conference Slovko 2015 (Natural Language Processing, Corpus Linguistics, Lexicography), pages 36–44. Lüdenscheid: RAM-Verlag.
[4] Goláňová, H., and Waclawičová, M. (2019). The DIALEKT corpus and its possibilities. Jazykovedný časopis, 70(2), pages 336–344.
[5] Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017). ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[6] Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., and Křen, M. (2017). ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[7] Bělič, J. (1972). Nástin české dialektologie. Praha: Státní pedagogické nakladatelství, 463 p.
[8] Balhar, J., Jančák, P. et al. (1992, 1997). Český jazykový atlas 1, 2. Praha: Academia, 427, 507 p.
[9] Balhar, J. et al. (1999, 2002, 2005). Český jazykový atlas 3, 4, 5. Praha: Academia, 577, 626, 680 p.
[10] Balhar, J. et al. (2011): Český jazykový atlas Dodatky. Praha: Academia, 579 p.
[11] P. Karlík, M. Nekula, and J. Pleskalová (eds.). (2002). Encyklopedický slovník češtiny. Praha: Nakladatelství Lidové noviny, 604 p.
[12] P. Karlík, M. Nekula, and J. Pleskalová (eds.). (2016). Nový encyklopedický slovník češtiny. Accessible at: ttps://www.czechency.org/.
[13] Statistický lexikon obcí v zemi České. (1934). Úřední seznam míst podle zákona ze dne 14. dubna 1920, čís. 266 Sb. zák. a nař. Statistický lexikon obcí v republice Československé. Praha: Orbis, 643 p.
[14] Statistický lexikon obcí v zemi Moravskoslezské (1935). Úřední seznam míst podle zákona ze dne 14. dubna 1920, čís. 266 Sb. zák a nař. Statistický lexikon obcí v republice Československé. Praha: Orbis, 236 p.
[15] Dialektologická komise České akademie věd a umění. (1951). Pravidla pro vědecký přepis dialektických zápisů českých a slovenských. Praha: Česká akademie věd a umění, 5 p.