**Dialectometric Analysis of Hungarian Dialects**

Fruzsina Sára VARGHA

## 1. Background, unresolved problems, work which has led to the project

The present project has several foundations. 1. It is inspired by dialectometric research (carried out mainly in Salzburg and Groningen), which has in recent years become a central topic in dialectology. 2. It will exploit the available Hungarian dialect databases created in several former computational dialectology projects. 3. The main technologies that will be used in the analysis have been already developed and are available to the project. 4. Preliminary research has been carried out on a huge quantity of data with promising results.

Dialect classification and the research of inter-dialectal relations have always been a central issue of linguistic geography. Classical methods that aimed at retrieving dialect boundaries are based on the analysis of a few linguistic variables chosen by the researcher, inevitably favouring his or her preconceptions. Thus classical methods offer a limited level of objectivity (see Nerbonne & Heeringa 2010), especially if the number of variables involved is limited. Another problem is – especially in territories where originally different Hungarian dialects are present in neighbouring locations – that there are practically no overlapping isoglosses, which makes it nearly impossible to define dialect boundaries using the traditional methods. The analysis of aggregate data, called dialectometry, makes dialect classification more objective. It aims to abstract a basic pattern from a linguistic atlas seen as a huge empirical database. The term was first used by Jean Séguy who determined linguistic distances by categorical data analysis (1973). Since the first application of such a method, several techniques have been developed (see also Chambers & Trudgill 1998: 137–140, Goebl 2006, Heeringa 2004). Lately the application of the Levenshtein algorithm (a string edit distance measurement) made possible the automatic comparison of words (strings of phonetic symbols) stored in appropriately digitised data sets. When comparing two words we calculate the number of operations needed to transform one string to another. That way we compare map by map the data collected at one location with data collected at other locations. The result of such comparisons is a similarity matrix showing how (dis)similar the collected data in one location are to data recorded in all other locations. In other words, linguistic (dis)similarity between every pair of locations is expressed by a numerical value or a percentage (for a detailed description of the method see Heeringa 2004, Nerbonne & Heeringa 2010, for its application to Hungarian dialect data see Vargha & Vékás 2009).

The era of computational dialectology in Hungary began with the development of a linguistic software called Bihalbocs (http://www.bihalbocs.hu/) for the time aligned transcription, analysis and automatic mapping of Hungarian dialect data. (This is a continuous, non-taxpayer-funded development made and supported by the interested researchers themselves, including the present applicant, in response to the evolving research needs.) The tool is used to build databases (oral corpora if sound is available) and also to hold together the connected pieces of information. A digitised, phonetically coded piece of data keeps its connections to the locality where it was collected, to the sociolinguistic metadata (such as the informant's age and gender), and also to the source recording it is extracted from. Thanks to the development efforts and the new methodology they generated (with the determinant role played in recent years by the present applicant), computational dialectology has become the main approach in Hungarian dialect research. The technology mentioned above is currently used at several universities in three different countries (HU, RO, SK) for the digitisation of already published dialect atlases and also for data processing in recent projects. A major advantage of this technology is the increased compatibility and convertibility (portability) of the constructed data structures. Time-aligned transcriptions are also convertible to PRAAT TextGrid files for further analysis, or can be published on the Internet (see for example http://geolingua.elte.hu/hkonyv/Hangoskonyv_index.html). For a more detailed overview of computational dialectology in Hungary see Vargha (2009) and Vékás (2007).

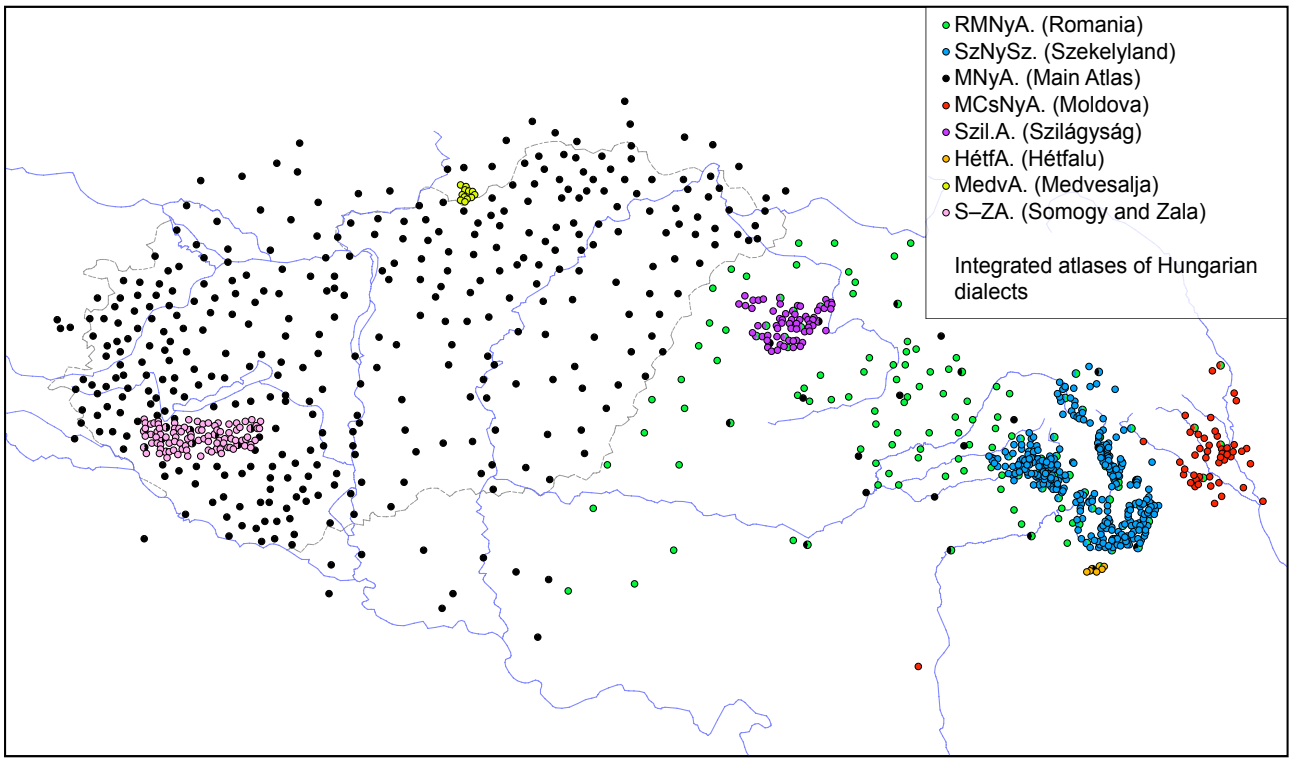


**Figure 1:** The already (or partially) digitised Hungarian dialect atlases. The 495 locations investigated in The Atlas of Hungarian Dialects appear in black.

Research on Hungarian dialects presupposes the availability of special software for two main reasons: tools developed in other laboratories do not meet our complex research needs, and they do not support the Hungarian traditional transcription system either. In the Hungarian linguistic tradition, we use a special narrow phonetic transcription, but thanks to an analytic encoding system the transcribed data is easily searchable (and convertible into broader transcription forms) by Bihalbocs. A reliable conversion method to IPA is not trivial to construct due to some significant underlying theoretical differences between the two approaches, but it is possible (this is still an unresolved question).

In the last fifteen years nearly one million data items from The Atlas of Hungarian Dialects and several regional Hungarian linguistic atlases have been appropriately digitised in a series of interrelated projects. The investigation points of these digitised atlases are mapped in Figure 1. Quantitative analyses on this corpus of rich dialect atlas transcripts and the automatic mapping of the results (showing spatial distribution of the investigated phenomena) are possible with Bihalbocs. The analytic capabilities and visualisation techniques present in this technology led the way to preliminary dialectometric research on Hungarian dialect atlases, with the use of the Levenshtein algorithm, the creation of similarity matrices and interactive dialectometric maps (Vargha & Vékás 2009). The applicability of dialectometry in dialect classification has been investigated in comparison with informants' dialect similarity judgements and by contrasting the dialectometric maps with the spatial distribution of dialect data and historical place names (Bodó–Vargha–Vékás 2012, Vargha 2010).

From the already (or partially) digitised atlases (Figure 1) four will be involved in the present project: The Atlas of Hungarian Dialects [A magyar nyelvjárások atlasza, MNyA., with 495 localities and 1137 digitised maps and 25 maps that will be digitised in the project]; The Atlas of Hungarian Dialects in Romania [A romániai magyar nyelvjárások atlasza, RMNyA., with 136 localities and 1200 digitised maps and about 300 maps that will be digitised in the project]; The Linguistic Atlas of Somogy and Zala Counties [Somogy–zalai nyelvatlasz, S–ZA., with 99 localities and 282 previously digitised maps that will be reviewed and corrected in the project]; the Moldavian Csángó Dialect Atlas [A moldvai csángó nyelvjárás atlasza, MCsNyA., 43 localities and 1054 digitised maps].

## 2. Hypotheses, key questions, aims of the project

The general aim of the project is the dialectometric analysis of duly digitised Hungarian dialect atlases. The following questions (1 to 6) can be formulated.

1. Could the automatic string edit distance analysis be validated by a manual classification of linguistic variables in the same corpus? – A regional atlas, The Linguistic Atlas of Somogy and Zala Counties (*Somogy–zalai nyelvatlasz*) will be investigated. Two different dialectometric analyses will be compared: the first will be based on the author's (Lajos Király) own manual classification, the second on different matrices created with the use of the Levenshtein algorithm. The impact of phonetic narrowness on the dialectometric outcome will also be investigated in order to find out what degree of narrowness (of the data items to be compared by the Levenshtein algorithm) correlates best with the results of the manual classification (see also 2. below).

2. What kind of linguistic questions could be answered by the comparison of different similarity matrices generated from narrow and broader transcriptions? – The digitised narrow transcriptions can be automatically converted into broader forms, thus different matrices can be generated. According to preliminary results we can formulate the following hypothesis: the similarity matrices generated from narrower or broader transcripts exhibit appreciable differences especially in the case of dialect enclaves and speaker communities that are distant in time or space from the originally related dialect area (figure 3 and 4 are showing the dialectometry of the same locality, Csíkrákos, according to a narrow and a broad form of dialect data). According to our hypothesis, the broader transcripts lead to an analysis that accentuates the effects of linguistic phenomena less resistant to change due to recent influences from the neighbouring dialects or languages.

3. How can linguistic atlases be integrated when they are different in transcription narrowness? – Hungarian dialects could not be analysed in a comprehensive way without the integration of The Atlas of Hungarian Dialects (495 localities) and The Atlas of Hungarian Dialects in Romania (136 localities), as in the former there are only 22 investigation points in Transylvania and it has no locations from the Hungarian speaking territory of Moldavia (eastern Romania). The transcription method is slightly diverging in the two atlases mainly because of some differences in the representation of diphthongs. A conversion method will be elaborated to narrow the gap between the transcriptions as much as possible. Several approaches will be considered, including the one proposed in Wieling and Nerbonne (2011). The method will be applied in the integration of other linguistic atlases as well.

4. How can loan-words borrowed from the surrounding language environment influence the outcome of the dialectometric analysis? – The effect of loan-words on the integrated dialectometric analysis of The Atlas of Hungarian Dialects in Romania and the Moldavian Csángó Dialect Atlas will be investigated. Loan-words typically have a considerable impact on the average similarity values. Bodó's research (2007) about the spacial distribution of loan-words in Hungarian dialects of Moldavia will provide a starting point for the analysis.

5. How can dialectometry be used for the investigation of the linguistic relations in the case of language or dialect enclaves? – The outcome of dialectometric analysis will be compared to the findings of previous dialect, onomasiological and settlement history research.

6. How can dialectometric analysis be involved in the determination of dialect areas? – Preliminary results suggest that dialect boundaries can be determined by the use of dialectometric maps (Vargha 2010, Bodó–Vargha–Vékás 2012). We can presume the existence of a dialect border between two neighbouring locations when the geographic "center of gravity" of their dialect similarity points into opposite directions. The planned conversion of the Hungarian dialect data into IPA format will also allow the use of methods available in Gabmap (a web application that visualises dialect variations, http://www.gabmap.nl/) for the automatic determination of dialect areas.

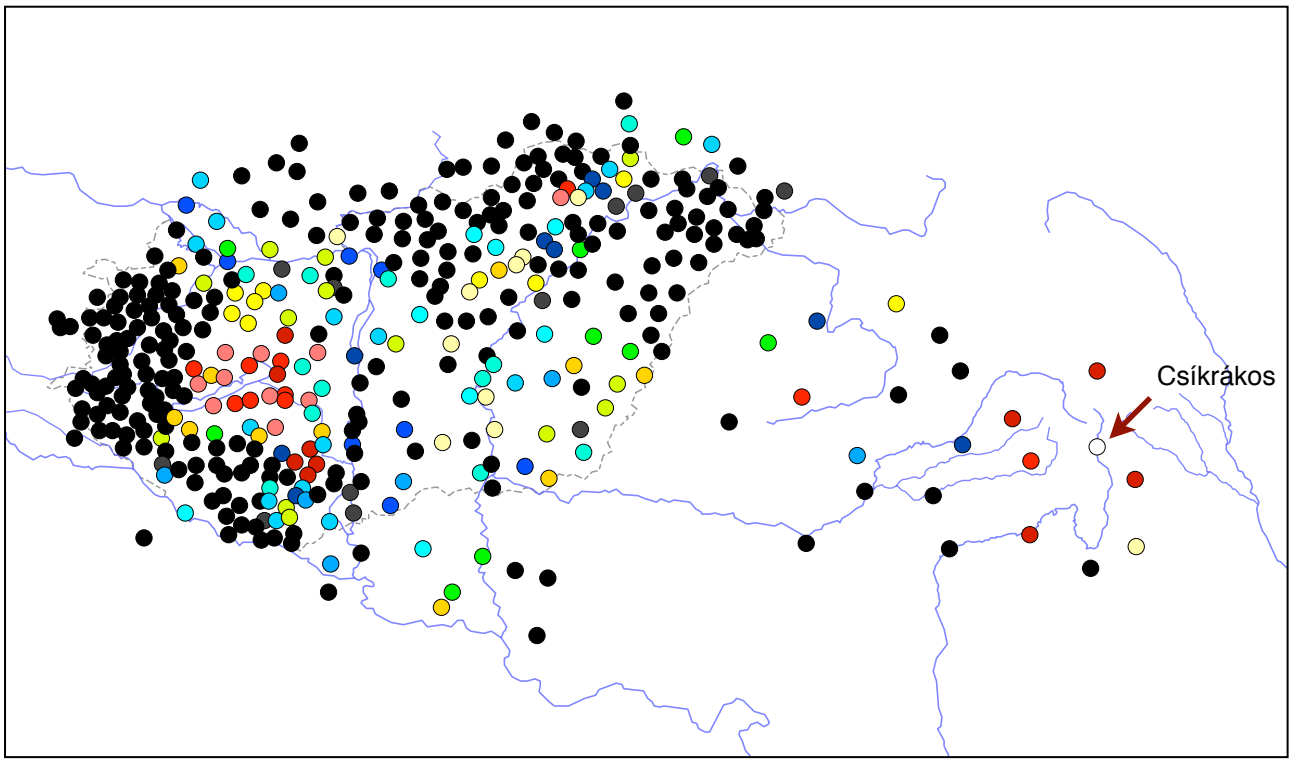


**Figure 2:** Dialectometric analysis of the location Csíkrákos using digitised data from The Atlas of Hungarian Dialects. The matrix is generated from the narrow transcription.


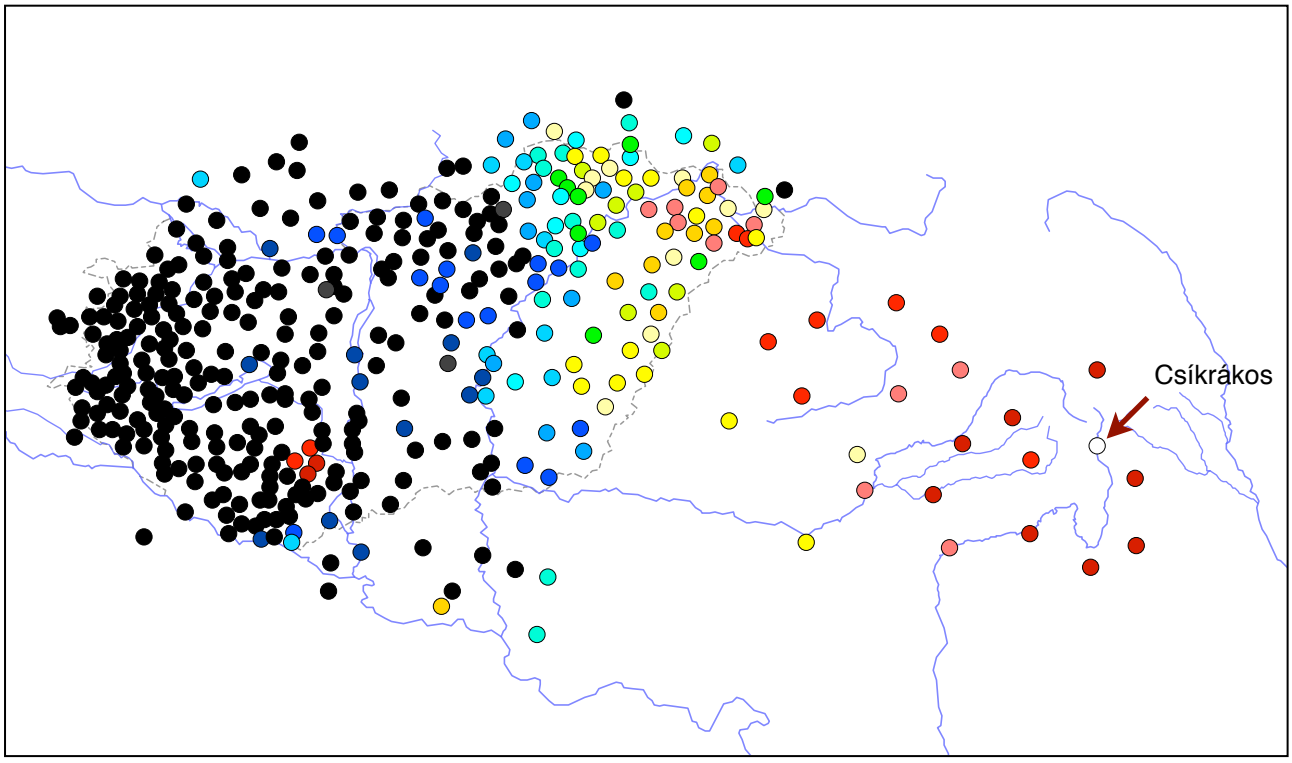


**Figure 3:** Dialectometric analysis of the location Csíkrákos using digitised data from The Atlas of Hungarian Dialects. The matrix is generated from a broad form of the original narrow transcription where all vowels are treated as equals.

## 3. Method

Previously developed research tools for the analysis of Hungarian digitised dialect data will be applied in the project. This dialectometric data analysis method is inspired by the works of Hans Goebl (interactive maps based on a similarity matrix), John Nerbonne and Wilbert Heeringa (string edit distance methods and general theoretical framework). The matrices underlying the dialectometric maps are based on previous manual classification of dialect data or, alternatively, they will be created with the use of string edit distance techniques (the Levenshtein algorithm).

The application of the research tools mentioned is also cost-effective because it is compatible with other computational dialectology projects. Analysis will be carried out mainly on already digitised Hungarian dialect data, thus the present project valorises the outcome of previous basic researches, some of which was financed by the Hungarian OTKA and NKA funding agencies.

## 4. Expected results

The main result of the present research will be a detailed dialectometric analysis of Hungarian dialects while answering the questions 1 to 6 in section 2 above. The confirmation of some assumptions, especially the applicability of dialectometry in dialect classification might have general implications in the field and means that these findings might be a reference point for Hungarian dialect researches in the future.

In order to make Hungarian data analysable with dialectometric techniques developed in other research laboratories, a conversion method of Hungarian narrow transcriptions to IPA symbols will be elaborated. As a result, all entries of our databases (including those digitised in the present project) will be converted to IPA symbols.

A simplified version of the dialectometric maps will be published on the Internet for educational purposes.

Results will be presented at national or international conferences. The most important findings will be published in Hungarian and international journals. A monograph will be written on the dialectometric analysis of Hungarian dialects.

See the attached work plan for the timeline and some more technical details.

## 5. Research infrastructure

The method for data processing and analysis to be used in the research has already been developed. The data bases are complete or will be completed in the first half of the project. No special research tools are needed except for a computer running Mac OS X and data storage tools. In addition to the full time employment of the leading researcher a PhD student will be partially employed in the first half of the project (database construction and integration).

## References

Bodó, Csanád 2007. A moldvai magyar nyelvjárások román kölcsönszórétegének területisége (= A territorial examination of Romanian loanwords in Moldavian Hungarian dialects). In: Benő, Attila – Fazakas Emese – Szilágyi N., Sándor (szerk.). *Nyelvek és nyelvváltozatok. Köszöntő kötet Péntek János tiszteletére.* I. kötet. Anyanyelvápolók Erdélyi Szövetsége Kiadó, Kolozsvár. 160–174.

Bodó, Csanád – Vargha, Fruzsina Sára – Vékás, Domokos 2012. Classifications of Hungarian dialects in Moldavia. In: Peti, Lehel & Tánczsos, Vilmos (eds.). *Language Use, Attitudes, Strategies: Linguistic Identity and Ethnicity in the Villages of the Moldavian Csángós.* The Romanian Institute for Research on National Minorities, Cluj-Napoca. 2012. 51-69. http://frufru.web.elte.hu/bodo_et_alii_2012.pdf

Chambers, J. K. & Trudgill, Peter 1998. *Dialectology*. Second edition. Cambridge University Press.

Goebl, Hans 2006. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21. 411–35.

Heeringa, Wilbert 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen Dissertations in Linguistics 46. Groningen.

Nerbonne, John & Heeringa, Wilbert 2010. Measuring Dialect Differences In: Schmidt, Jürgen Erich & Auer, Peter (eds.). *Language and Space. An International Handbook of Linguistic Variation: Theories and Methods.* Mouton de Gruyter, Berlin. 550–567.

Séguy, Jean 1973. La dialectométrie dans l'atlas linguistique de la Gascogne. *Revue de linguistique romane* 37. 1–24.

Vargha, Fruzsina Sára 2009. The New Oral Corpus and Related Talking Maps of Hungarian Dialects from the 1960s. SIDG Congress, Maribor, September 14. 2009. http://frufru.web.elte.hu/Vargha_Maribor2009.pdf

Vargha, Fruzsina Sára 2010. A dialektometria alkalmazása és történeti helynevek nyelvföldrajzi vizsgálata a Székelyföldön (= The application of dialectometry and the geolinguistic study of historical place-names in Székelyland). *Helynévtörténeti Tanulmányok* 5. 223–233.

Vargha, Fruzsina Sára & Vékás, Domokos 2009. Magyar nyelvjárási adattárak vizsgálata interaktív dialektometriai térképekkel (= The investigation of Hungarian dialect databases with interactive dialectometric maps). Előadás a Magyar Nyelvtudományi Társaság felolvasóülésén. 2009. március 24. http://bihalbocs.hu/eloadas/dialektometria_20090324.pdf

Vékás, Domokos 2007. Számítógépes dialektológia (= Computational dialectology). In: Guttmann Miklós – Molnár Zoltán (szerk.): *V. Dialektológiai Szimpozion.* Berzsenyi Dániel Főiskola, Szombathely. 289–293.

Wieling, Martijn & Nerbonne, John 2011. Measuring Linguistic Variation Commensurably. *Dialectologia*. *Special Issue II: Production, Perception and Attitude*. 141–162.