Application of new classification methods in the study of bilingualism and ethnic identity in Hungary*

András Vargha
Professor
Institute of Psychology
Károli Gáspár University
of the Reformed Church,
Hungary
E-mail: vargha.andras@kre.hu

Anna Borbély
Senior Research Fellow
Research Institute for
Linguistics
Hungarian Academy
of Sciences,
Hungary
E-mail: borbely.anna@nytud.mta.hu

This study demonstrates the usefulness of new and promising modern techniques of cluster analysis in linguistics. Data are employed from a study on language shift and assimilation of six bilingual nationalities [1] living in Hungary (Boyash, Roma, German, Romanian, Serb, and Slovak). Using ethnic identity, attitude towards minority language, language choice in family, and minority language proficiency as variables, statistically valid and linguistically meaningful homogeneous types are explored, characterising bilingual adult persons from these national minorities. Based on cluster analyses, ten main types are identified in the process of language shift and assimilation.

KEYWORDS:
Nationalities in Hungary.
Bilingualism.
Cluster analysis.

DOI: 10.15196/hsr2018.01.en005

* The preparation of the present article was supported by a research grant obtained from the Faculty of Humanities, Károli Gáspár University of the Reformed Church (Person- and Family-oriented Health Study, Grant No. 2018/20643B800). It was also written in the framework of 'Languag-E-Chance': Development of language conscious schools, bilingual deaf education, and innovative methods and tools of knowledge exploitable by language – RIL-HAS (Research Institute for Linguistics, Hungarian Academy of Sciences) Languag-E-Chance Educational Research Group's project (2016–2020). The project has also been funded by a grant from the National Research & Development Programme (NKFP 5/126/2001; coordinated by *Csilla Bartha* and managed by the second author) and from the Hungarian Scientific Research Fund (OTKA K 81574).

[1] Term used by the Act No. CLXXIX/2011 on the rights of nationalities (see http://www.ilo.org/dyn/natlex/natlex4.detail?p_lang=en&p_isn=96678).

In the last Hungarian census carried out in 2011, about 7% of the population of Hungary declared themselves as belonging to one of Hungary's 13 officially accepted nationalities (national minority groups).[2] These groups, characterised by their own languages and cultures, also belong to the more inclusive category of bilingual people in Hungary. Though bilingualism is obviously equally useful for cultural, political, and economical reasons, several studies have shown that over time the national minorities of Hungary lose their minority languages and assimilate to the Hungarian majority. Thus, transforming from a monolingual minority status to, temporarily, becoming a bilingual minority, finally, they reach a monolingual majority status. This change occurs at differing, but increasing rhythms (e.g. *Bartha* [2003], *Borbély* [2015]). In sociolinguistics, this phenomenon is called language shift (e.g. *Gal* [1979]).

The present paper studies bilingualism and ethnic identity in six national minority groups living in Hungary (Boyash[3], Roma, German, Romanian, Serb, and Slovak). These groups are living the process of language shift and assimilation. For this reason, the focus of our study is to compare the sustainability of bilingualism in these six national minority populations. The concept of sustainable bilingualism does not refer to bilingualism as a stable, long-lasting, and static phenomenon of two co-occurring languages, but to a much more diverse and constantly and dynamically changing linguistic phenomenon, which can be experienced only under certain circumstances (*Borbély* [2015]; cf. linguistic sustainability by *Bastardas Boada* [2014][4]).

The central goal of our paper is to identify the main processes of language shift and forms of ethnic identity by means of new pattern oriented statistical methods.

[2] See Table 2.1.6.3. on http://www.ksh.hu/nepszamlalas/tables_regional_00

[3] Boyash is a Romanian-speaking Gypsy population living in Hungary, Romania, Bosnia, Bulgaria, Croatia, Greece, Serbia, and Slovakia. They represent the smallest group of the Gypsy population in Hungary, about 40,000-50,000 in number and mainly live in southern Transdanubia. The history of their origin is less known than the history of the Hungarian-speaking Roma or the Romani-speaking groups.

[4] *Bastardas Boada* defines linguistic sustainability in the following way: 'To paraphrase *Ramon Folch*, we could say that *linguistic sustainability should be a process of gradual transformation from the current model of the linguistic organisation of the human species, a transformation whose objective would be to avoid that collective bilingualism or polyglottism of human beings must require the abandonment by different cultural groups of their own languages*' (*Bastardas Boada* [2014] pp. 137–138.).

1. Research design

In the ensuing sections, we present the methods, project, variables and subjects of our empirical investigation, and statistical analyses applied.

1.1. Project

The empirical method for obtaining the data sample submitted to cluster analysis was carried out in accordance with the framework of the project 'Dimensions of linguistic otherness: prospects of minority language maintenance in Hungary', created by researchers in the Department of Sociolinguistics, Research Institute for Linguistics, Hungarian Academy of Sciences and the Department of Modern Hungarian Language, Eötvös Loránd University, between 2001 and 2004. The project has been funded by a grant from the National Research & Development Programme (NKFP 5/126/2001), and it is called, in short, the HuBiLing study. In this project, six (Boyash, Roma, German, Romanian, Serb, and Slovak) nationalities living in Hungary were involved, each with one or two representative settlements/local communities.[5] The fieldwork of the comparative sociolinguistic research was carried out between the fall of 2003 and spring of 2004. The data were collected with a questionnaire by persons belonging to the same nationality who knew the local minority language variety (with one exception of a member of the Slovak research group). Their professions were as follows: linguists, university teachers, elementary and secondary school teachers, and university students. In this questionnaire, questions[6] were asked about different components of minority and majority language choices in different domains (family, religion, work place, etc.), language proficiency, ethnic identity, and language attitude (*Bartha* [2003]).

1.2. Variables

The following are the four sociolinguistic variables of our study – related to sustainable bilingualism and ethnic identity – that were involved in the pattern-oriented analyses.

> *Ident*: Ethnic identity (based on participants' subjective opinions of their own nationality and first language);

[5] These settlements from Hungary were as follows: Mánfa and Bogyiszló (Boyash), Tarján (German), Mezőtúr (Roma), Kétegyháza (Romanian), Pomáz (Serb), and Tótkomlós (Slovak). The authors are thankful to *Anna Orsós* and *Borbála Egregyi* (Boyash), *Maria Erb* (German), *Andrea Kiss* (Roma), *Mária Abrudán* (Romanian), *Milosné Szimics* (Serb), *Erzsébet Hornokné Uhrin*, *Sándor Tóth*, *Tünde Tuska*, and *Mária Zsilák* (Slovak) for the data collection.

[6] This paper does not list each question in the questionnaire.

Attit: Aesthetic and emotional attitude referring to minority language (based on answers to questions about the beauty and likability of respondents' own minority language);

Family: Frequency of minority language choice with family members (averaging corresponding data with respect to parents, siblings, spouses, children, and grandchildren);

LanProf: Level of language proficiency in minority language based on a self-evaluation (averaging proficiency ratings of local minority language variety and that of the standard variety [the latter only if it existed]).

The internal consistency and psychometric reliability of these four variables, regarded as items of the same scale, was confirmed by a Cronbach-α value of 0.68 (*Cohen* [1977], [1992]).

1.3. Subjects

Respondents/Subjects were adult persons (at least 20 years old at the time of the investigation) who voluntarily participated in the investigation and were members of one of the six nationality groups mentioned above. The total number of subjects participating in the investigations was 421. Each nationality was represented by 70–71 subjects. An attempt was made to balance the representation of males and females, the four educational levels, and the three age ranges. Some basic statistics for gender, age, and level of education are summarised in Table 1.

Table 1

Basic statistics for gender, age, and level of education in the six nationality samples

Nationality	Size	Male-female percentage	Age mean (years)	Range of age (years)	Percentage of four education levels
Boyash	70	50-50	46.2	21–81	1.7-94.8-3.7-0.0
Roma	70	49-51	46.5	20–84	40.0-45.7-12.9-1.4
German	70	50-50	51.0	23–80	0.0-47.1-38.6-14.3
Romanian	71	51-49	52.1	28–80	0.0-42.3-42.3-15.5
Serb	70	50-50	49.9	21–81	0.0-42.9-42.9-14.3
Slovak	70	49-51	51.3	21–76	0.0-44.3-41.4-14.3

Note. In the last column of the table, the meaning of the presented percentage values are as follows: first number – proportion of those persons who did not complete the eight elementary school grades; second number – completed eight grades in elementary school; third number – completed secondary school; fourth number – completed college or university.

Source: Here and in following tables and figures, data acquired from the 'Dimensions of linguistic otherness: prospects of minority language maintenance in Hungary' project conducted in 2003 and 2004.

The six samples are very similar in terms of gender distribution and age. However, in terms of the education level, they differ substantially due to the fact that in the Boyash and Roma samples the proportion with the highest level of education was negligible (in contrast to the other samples, where it was above 14%), whereas the proportion of the lowest level was substantial (in the Roma sample, for example, it was 40%).

1.4. Statistical analyses

The main goal of our investigation was to explore the different types of language shift and ethnic identities in the process of assimilation. This will enable researchers to obtain relevant information on sustainable bilingualism and ideas for how to maintain ethnic identity. The type exploration was done by using cluster analysis on the four sociolinguistic variables (see Section 1.2.) of the 421 subjects. The statistical analyses were carried out using ROPstat software (*Vargha–Torma–Bergman* [2015]).

An additional goal of our study was to demonstrate the correct run and interpretation of cluster analysis, clarifying important decisions on the number of clusters as well as the adequacy and significance of the obtained cluster structure.

2. Results

In the following, we describe the process of obtaining an appropriate cluster structure and its internal and external validity, and we detail the main characteristics of the obtained clusters.

2.1. The process of obtaining an appropriate cluster structure

The strength of the four sociolinguistic variables (*Ident*, *Attit*, *Family*, *LanProf*), assessed by means of Spearman's ρ, was 0.13–0.47. A weakest rank correlation was obtained between *Ident* and *LanProf* ($r = 0.13$) and *Attit* and *LanProf* ($r = 0.24$). The strongest ($r = 0.47$) was obtained between *Family* and *LanProf*. The variance proportions of the four variables explained by the other three remaining were 24%, 24%, 40%, and 24%, respectively. This shows that the four variables are in a relationship of a weak to moderate level, despite the relatively high Cronbach-α, which reflects a respectable common portion of the shared variance. Consequently, each

variable has some special information, not shared by the others. This implies that our set of variables is an optimal input for a classification analysis.

In the first step we checked, via residual analysis, for the presence of outliers, which can distort the cluster structure – no outliers were found. Then a HCA (hierarchical cluster analysis) was carried out using Ward's method, applying the ASED (average squared Euclidian distance) of cases (*Bergman–Magnusson–El-Khouri* [2003] Chapter 4). For assessing the goodness of the different cluster structures, the HCA module of ROPstat computes the following adequacy measures, the so-called QCs (quality coefficients) (*Bergman–Vargha–Kövi* [2017], *Vargha–Bergman–Takács* [2016]):

– EESS%: explained variance proportion;
– PB: point-biserial index, measuring the extent to which inter-cluster cases are closer to each other than cases belonging to different clusters;
– XBmod: modified Xie-Beni index, measuring the extent to which the average distance from one's own cluster centre is smaller than the distance of the two nearest cluster centres;
– SC: Silhouette-coefficient, measuring the extent to which the average distance from one's own cluster centre is smaller than the average distance of cases from the nearest foreign cluster centre;
– HCmean: the weighted mean of the cluster homogeneity coefficients (weights are the cluster sizes).

The QCs for cluster numbers 5 to 12 are summarised in Table 2. The best structure seems to be the one belonging to cluster number $k = 10$, since, for this solution, the two most important QCs measuring the homogeneity of the clusters are appropriate (EESS% = 70.47 that exceeds 65.0, and HCmean = 0.60, which is well less than 1.00), and no HC-value[7] is greater than 1. From the QCs measuring the separation of the clusters that are of secondary importance, the PB reaches the expected 0.30 level, and both the XBmod and SC are close to the acceptable 0.50 level.

We attempted to improve upon this 10-cluster HCA solution via relocation (KCA – k-means cluster analysis). This happened to be successful since EESS% increased and HCmean decreased to a sensible extent. In Table 3, one can see that the QCs of this KCA-structure are good (EESS% > 74, PB > 0.30, XBmod > 0.40, SC > 0.60, HCmean < 0.55).

[7] For any cluster, the HC homogeneity coefficient is the average pairwise within-cluster distance, so the smaller the HC-value of a cluster, the more homogeneous it is.

Table 2

The quality coefficients of the hierarchical cluster analysis for five to twelve clusters

Number of clusters	EESS%	PB	XBmod	SC	HCmean	HCmin–HCmax
12	74.08	0.288	0.320	0.471	0.533	0.18–0.88
11	72.40	0.288	0.276	0.469	0.565	0.18–1.00
10	**70.47**	**0.317**	**0.327**	**0.449**	**0.604**	**0.40–1.00**
9	68.54	0.319	0.283	0.444	0.641	0.42–1.00
8	65.40	0.346	0.211	0.446	0.703	0.42–1.00
7	62.11	0.389	0.263	0.478	0.768	0.50–1.00
6	58.60	0.382	0.195	0.478	0.838	0.50–1.33
5	54.72	0.377	0.119	0.462	0.915	0.67–1.33

Note. Here and in the following tables, EESS: explained variance proportion; PB: point-biserial index; XBmod: modified Xie-Beni index; SC: Silhouette-coefficient; HC: cluster homogeneity coefficient; HCmean: the weighted mean of the cluster homogeneity coefficients. The line of the best hierarchical structure is highlighted with boldface font type.

Table 3

QCs of the hierarchical and non-hierarchical (k-means) cluster structure for ten clusters

Type of CLA	EESS%	PB	XBmod	SC	HCmean	HCmin–HCmax
Hierarchical	70.47	0.317	0.327	0.449	0.604	0.40–1.00
K-means	74.48	0.325	0.438	0.583	0.523	0.26–0.86

Note. Here and in the following tables, QC: quality coefficient; CLA: cluster analysis. Except for the HC-values, the higher the QC, the better the cluster structure.

2.2. Results obtained for internal validation

The validity investigation of a cluster structure consists of statistical analyses through which the goodness and adequacy of the structure is confirmed. This may be done through significance tests and the computation of specific coefficients. In an internal validity analysis, we only consider the input variables of the cluster analysis and characteristics of the obtained structure. Since the value levels of the QCs (EESS%, HCmean, etc.) depend strongly on the number of input variables and number of clusters, it is proposed that the QCs of the cluster structure, obtained for a data set should be related to the QCs obtained by parallel control cluster analyses of data

sets characterised by independent random variables (*Vargha–Bergman–Takács* [2016]). Performing this analysis, we applied 25 random repetitions of generating independent random variables in three different ways: *1*. random permutation of input variables values, *2*. generation from multidimensional uniform continuous distribution, and *3*. generation from multidimensional normal distribution.

The QCs of our 10-cluster structure were, in most cases, (and always for EESS%, SC, and HCmean) significantly better than those obtained for similar cluster structures based on simulated random control data at $p < 0.001$ level.

Beyond significance tests, *Vargha–Bergman–Takács* [2016] also proposed to compute MORI (measure of relative improvement) coefficients in order to assess the extent to which the QCs of an obtained structure are better than those obtained for structures of random data control. Their suggested criterion thresholds for MORI values for interpreting a cluster structure were as follows: minimally acceptable level: 0.15, moderate level: 0.30, and high level: 0.50.

Table 4

The internal validation of the ten-cluster solution by means of random data controls

Real sample/MORI	QCs and their relative improvements				
	EESS%	PB	XBmod	SC	HCmean
Real sample	74.48	0.325	0.438	0.583	0.523
MORI – control: random permutation of the values of the input variables	0.20	–0.05	0.12	0.07	0.20
MORI – control: independent random uniform continuous variables	0.20	–0.09	–0.29	0.07	0.20
MORI – control: independent random normal variables	0.34	0.04	0.04	0.16	0.34

Note. MORI: measure of relative improvement. The obtained MORI values are based on 25 independent repetitions of random data generations.

Table 4 contains the QC and MORI values for five QCs. We see that the advantage of our 10-cluster structure compared to random data controls is moderate in terms of the most important QCs, the EESS% and HCmean, and weak in terms of the PB, XBmod, and SC. Consequently, we can declare that, in term of the homogeneity of the clusters, the internal validity of the obtained structure is of moderate size, which deserves the interpretation of most of the clusters. The low MORI values of PB, XBmod, and SC indicate that the separation of some clusters is not optimal, which necessarily occurs if the cluster number is high (in our case it is ten).

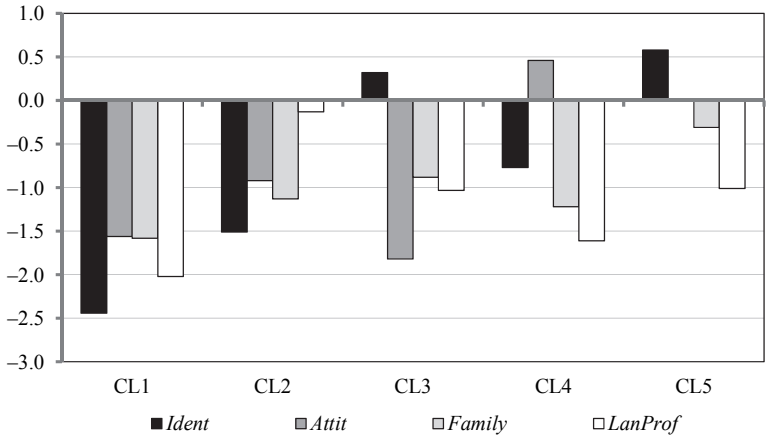
Some recent studies (*Bergman–Vargha–Kövi* [2017], *Vargha–Borbély* [2017]) argue that, beyond the three types of random data control presented in Table 4, it is worth applying a fourth one, where random control data are generated by means of a multivariate normal distribution that can be characterised with the same inter-correlations as the set of input variables. Using this more stringent criterion, the obtained MORI values were all significantly greater than 0.00, and, for EESS% and HCmean, they reached the 0.20 level.

Finally, we performed similar validation analyses with the best 8- and 9-cluster solutions to confirm that the 10-cluster solution can best be adopted for the clusterisation of the current sample. These analyses showed that the MORI values of EESS% and HCmean were always somewhat lower than those obtained in the case of the 10-cluster structure, thus confirming the superiority of the 10-cluster solution over other possible alternatives.

2.3. The explanation of the obtained clusters

After finding the good internal validity values of the previous section, we plotted the pattern of standardised means of the final 10-cluster solution separately for the first five clusters (CL1–CL5), representing the dominantly Hungarian (language and identity) groups (see Table 5 and Figure 1) and the last five clusters (CL6–CL10), representing the dominantly minority (language and identity) groups (see Table 5 and Figure 2).

Figure 1. The pattern of standardised means of the final ten-cluster solution for the five clusters of mainly Hungarian dominance (CL1–CL5)



Note. For cluster explanations, see Table 5. Here and in the following table, *Ident*: ethnic identity; *Attit*: minority language attitude; *Family*: language choice in the family; *LanProf*: minority language proficiency. Lower values for all variables indicate a higher level of Hungarian assimilation.

Table 5

Explanation, short label, size, proportion, and homogeneity of the ten clusters

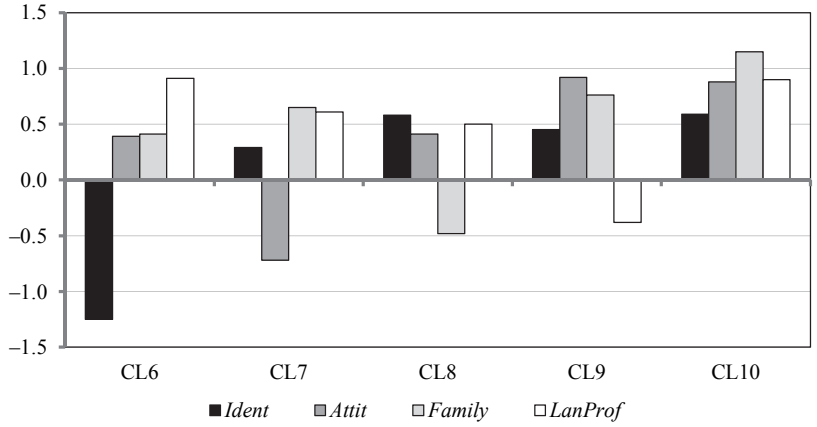
Cluster	Explanation and brief label	Size	Proportion (%)	HC
CL1	Completed language shift, finished Hungarian assimilation: C1H++	15	3.6	0.62
CL2	Hungarian assimilation with some minority language proficiency: C2LanP+	46	10.9	0.86
CL3	Hungarian assimilation with some minority ethnic identity: C3Id+	24	5.7	0.68
CL4	Hungarian assimilation with some positive minority language attitude: C4Att+	20	4.8	0.74
CL5	Intermediate status with slight minority ethnic identity and low-level minority language proficiency: C5LanP–	44	10.5	0.42
CL6	Sustained minority language proficiency, but lost minority ethnic identity: C6Id–	25	5.9	0.68
CL7	Sustained minority language proficiency and ethnic identity, but negative attitude toward own minority language: C7Att–	61	14.5	0.59
CL8	Sustained minority language proficiency and ethnic identity, but missing minority language choice in the family: C8Fam–	67	15.9	0.43
CL9	Sustained ethnic identity and minority language choice in the family with deteriorating language proficiency: C9LanP–	50	11.9	0.43
CL10	Sustained minority language and ethnic identity: C10Min+	69	16.4	0.26
	Total	*421*	*100.0*	–

Note. Here and sometimes in other tables, the total of all percentages differs from 100.0% due to rounding.

The cluster patterns in Figure 1 all reflect a more or less advanced level of language shift. Among them CL1 is of Hungarian dominance in each sociolinguistic component/variable. However, in clusters CL2–CL4 we find just one component (CL2 – minority language proficiency, CL3 – ethnic identity, CL4 – minority language attitude), which shows a slight but existent presence of the minority language and ethnic identity in the lives of the respondents. Respondents belonging to CL5 primarily use the Hungarian language. However, they still keep their ethnic identities as reflected by the somewhat increased attitude level toward their minority language.

The patterns in Figure 2 all reflect clusters where at least three out of the four sociolinguistic variables of minority language choice and ethnic identity show minority dominance. CL10 reflects the strongest attachment to the group's own nationality, but CL6, CL7, CL8, and CL9 are also at a high level for this attachment, which tends toward Hungarian assimilation only in one aspect (ethnic identity, minority language attitude, language choice in the family, or minority language proficiency).

Figure 2. The pattern of standardised means of the final ten-cluster solution for the five clusters of minority dominance (CL6–CL10)



Note. For cluster explanations, see Table 5. Lower values for all variables indicate a higher level of Hungarian assimilation.

In order to test the stability of the obtained cluster structure, we compared the 10-cluster solution with the best 8- and 9-cluster ones by means of the Centroid module of ROPstat. The analyses showed that six clusters of the 10-cluster solution (the first three and the last three) almost perfectly matched six clusters in the 8-cluster solution, and eight clusters (the first four and the last four) matched very well to eight clusters in the 9-cluster solution. In this latter comparison, even the weakest match (with a distance of 0.014 of the corresponding centroids) was quite acceptable. The only difference between the 9- and 10-cluster solutions was that CL4 and CL5 were fused into one common cluster in the 9-cluster solution. However, this fusion is not acceptable from a sociolinguistic point of view because of the clearly opposing trends of ethnic identity level for CL4 and CL5. (See Figure 1.) Consequently, we cannot explain the differences among the minority peoples studied with less than ten clusters.

These comparisons show also that in the process of language and identity shift the most stable types are the extreme ones. The intermediate types in the middle are the most flexible. The exploration and identification of the numerous types indicate the complexity of the process of assimilation, but at the same time, provide opportunities for various possible interventions to attain and help preserve the condition of sustainable bilingualism.

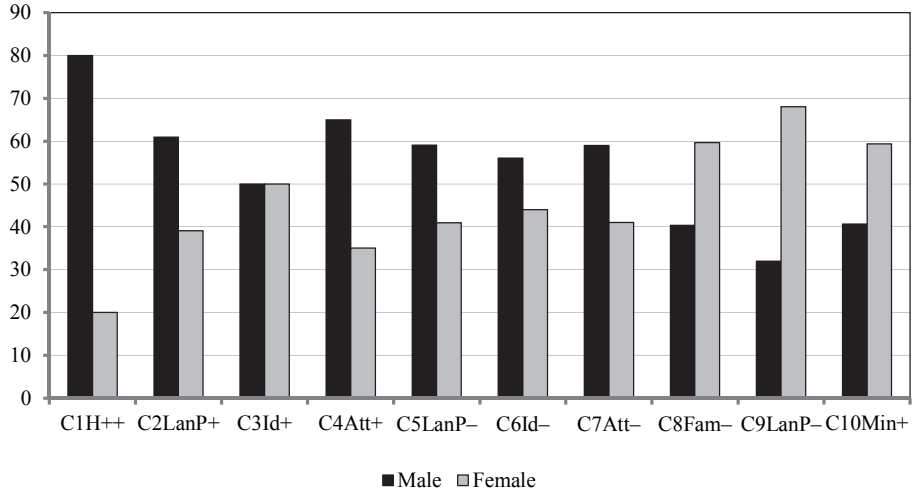
The explanations and the frequencies of the ten final clusters are summarised in Table 5. Here, we see that the proportion of the first five clusters (those of Hungarian dominance) is 35%, whereas the proportion of the last five clusters (those of minority dominance) is 65%. This has the positive message that about two-third of our subjects could preserve their minority language choice and ethnic identity in the form of a sustainable bilingual state.

2.4. Results of external validity investigations

The aim of the external validity investigations is to confirm the adequacy and meaningfulness of a cluster structure by means of external variables; they are not involved in the technical process of cluster analysis. In this context, we related the following four variables to the final 10-cluster solution: gender, age, education level, and nationality group.

Male-female proportions are illustrated in Figure 3. From this, we can see that in CL1–CL7 (C1H++ – C7Att–) males are dominant (in the case of CL1 [C1H++] with 80%) with one exception (CL3 [C3Id+]), whereas there is a clear female dominance in the clusters of the highest levels of minority language choice and ethnic identity (CL8–CL10 [C8Fam– – C10Min+]). This relationship is strongly significant as well ($\chi^2(9) = 24.60$, $p = 0.0034$, Cramér's $V = 0.24$).

Figure 3. Male and female percentages in the ten clusters

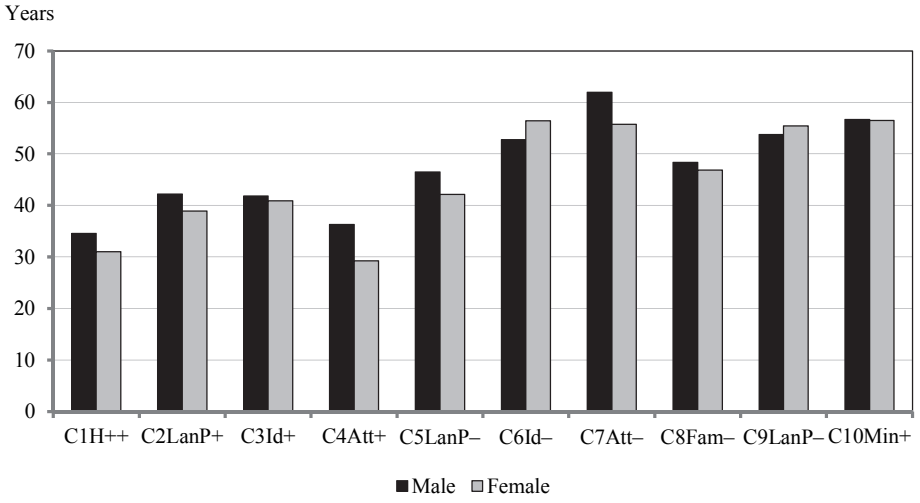


Note. For cluster explanations, see Table 5.

Figure 4 shows the mean ages of the ten clusters for the two genders. It is noticeable that the mean ages of the most Hungarian dominant CL1–CL4 (C1H++ – C4Att+) clusters are all below 45 years, whereas the means of clusters CL6–CL10 (C6Id– – C10Min+), containing people maintaining their minority language choice and ethnic identity to the highest possible extent, are all greater than 45. It is not by chance that the mean ages among these five clusters is the smallest in CL8 (C8Fam–), where the minority language choice within the family is the least frequent. This influence of age was also strongly statistically significant (in two-way ANOVA [analysis of variance] $F(9; 401) = 14.35$, $p < 0.0001$; partial $\eta^2 = 0.244$). It is worth noting as well that the age pattern was almost the same for males and females.

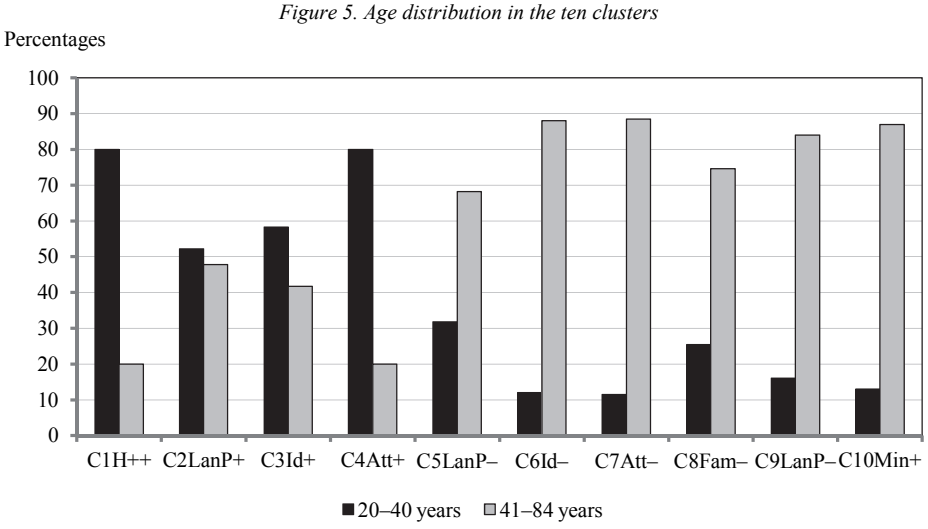
The impact of age was also tested by comparing the proportions of two age groups (20–40 and 41–84) across the ten clusters. Figure 5 illustrates the high dominance of elders in the minority dominant clusters (CL6–CL10 [C6Id– – C10Min+]) and the majority of younger people in the Hungarian dominant clusters (CL1–CL4 [C1H++ – C4Att+]).

Figure 4. Mean ages in the ten clusters



Note. For cluster explanations, see Table 5.

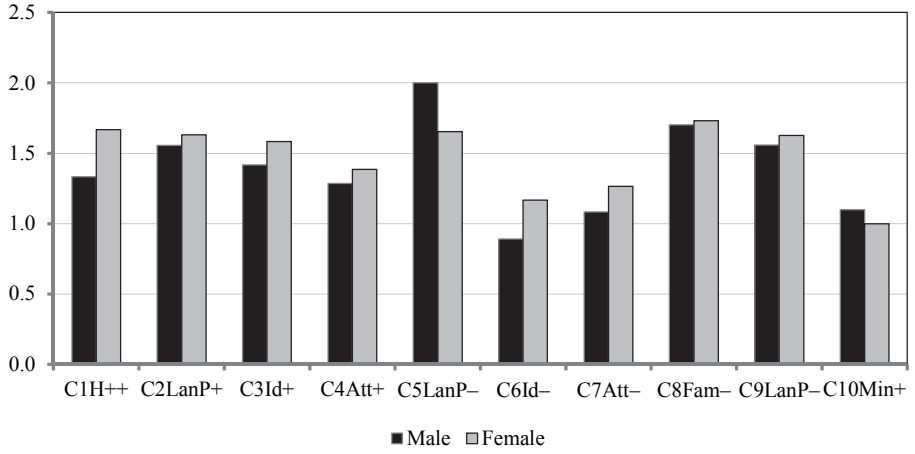
Figure 5. Age distribution in the ten clusters



Note. For cluster explanations, see Table 5.

The mean education levels of the ten clusters can be seen in Figure 6. The influence of this factor was also significant (in two-way ANOVA $F(9;\ 389) = 6.27$, $p < 0.0001$; partial $\eta^2 = 0.127$). When we compared the cluster means pairwise, the education level advantage of CL5 (C5LanP–) against clusters CL6 (C6Id–), CL7 (C7Att–), and CL10 (C10Min+), all being 'right' from CL5 (C5LanP–) in Figure 6, was significant at $p < 0.01$ level. The education level of CL10 (C10Min+) was also significantly smaller than the mean of the other united clusters (using Scheffé-type contrast method $F(9;\ 389) = 1.951$, $p = 0.0439$; *Scheffé* [1959]). It is again worth noting that the pattern of education level means – just as it was with age – is almost the same for males and females.

Figure 6. Averages of education levels in the ten clusters for males and females



Note. For cluster explanations, see Table 5. 0 – elementary school not completed; 1 – completed eight-grade elementary school; 2 – completed secondary school; 3 – completed college or university.

The percentage distribution of the six nationality groups within the ten clusters are summarised in Table 6. It is apparent that each nationality group is dominant in one of the clusters with a proportion of 34.0% or higher. Slovaks are represented in CL1 (C1H++) by 40.0%, Germans in CL2 (C2LanP+) by 45.7%, Romanians in CL3 (C3Id+) by 50.0%. In addition, Romas are represented in the most highly minority affiliated CL10 (C10Min+) cluster by 42.0%, and similarly Serbs in CL9 (C9LanP–) by 34.0%. It is also worth mentioning the 60.0% proportion of Boyashes in cluster CL6 (C6Id–).

Table 7 shows the cluster percentages in the different minority groups. According to this table, clusters CL1–CL5 (C1H++ – C5LanP–), which all show a higher level of language shift, are represented in the total sample by a proportion of 35.5%. It is noticeable that they form a wider group within the Slovak (50.1%), Romanian (46.4%), and German (43.0%) minorities, whereas their proportion is only 14.3%

among the Romas. That means that a vast majority (85.7%) of Romas can still be characterised by maintaining their language and identity. However, to a slightly smaller extent, it is also typical among Serbs (71.5%) and Boyashes (70.0%).

Table 6

The percentage distribution of the six nationality groups within the ten clusters

Cluster	Boyash	Roma	German	Romanian	Serb	Slovak	Total
C1H++	6.7	13.3	13.3	20.0	6.7	**40.0**	*100.0*
C2LanP+	13.0	8.7	**45.7**	8.7	4.3	19.6	*100.0*
C3Id+	16.7	4.2	8.3	**50.0**	0.0	20.8	*100.0*
C4Att+	30.0	5.0	15.0	15.0	25.0	10.0	*100.0*
C5LanP–	9.1	4.5	4.5	25	27.3	29.5	*100.0*
C6Id–	**60.0**	8.0	16.0	4.0	0.0	12.0	*100.0*
C7Att–	19.7	26.2	19.7	14.8	8.2	11.5	*100.0*
C8Fam–	13.4	11.9	23.9	16.4	14.9	19.4	*100.0*
C9LanP–	8.0	10.0	10.0	28.0	**34.0**	10.0	*100.0*
C10Min+	13.0	**42.0**	4.3	4.3	26.1	10.1	*100.0*
Total	16.6	16.6	16.6	16.9	16.6	16.6	*100.0*

Note. For cluster explanations, see Table 5. In the table, proportions exceeding 33.0% are denoted by boldface font type.

Table 7

Cluster percentages in the different nationality groups

Cluster	Boyash	Roma	German	Romanian	Serb	Slovak	*Total*
C1H++	1.4	2.9	2.9	4.2	1.4	8.6	3.6
C2LanP+	8.6	5.7	30.0	5.6	2.9	12.9	10.9
C3Id+	5.7	1.4	2.9	16.9	0.0	7.1	5.7
C4Att+	8.6	1.4	4.3	4.2	7.1	2.9	4.8
C5LanP–	5.7	2.9	2.9	15.5	17.1	18.6	10.5
Total	*30.0*	*14.3*	*43.0*	*46.4*	*28.5*	*50.1*	*35.5*
C6Id–	21.4	2.9	5.7	1.4	0.0	4.3	5.9
C7Att–	17.1	22.9	17.1	12.7	7.1	10.0	14.5
C8Fam–	12.9	11.4	22.9	15.5	14.3	18.6	15.9
C9LanP–	5.7	7.1	7.1	19.7	24.3	7.1	11.9
C10Min+	12.9	41.4	4.3	4.2	25.7	10.0	16.4
Total	*70.0*	*85.7*	*57.0*	*53.5*	*71.5*	*50.0*	*64.6*

Note. For cluster explanations, see Table 5. The sums of the totals of CL1–CL5 (C1H++ – C5LanP–) and CL6–CL10 (C6Id– – C10Min+) occasionally differ from 100.0 due to rounding.

3. Discussion

The main goal of the present paper was to demonstrate the usefulness of modern, person-oriented cluster analytic techniques in linguistics. The data came from a study of language shift and national minority assimilation of six nationalities (Boyash, Roma, German, Romanian, Serb, and Slovak) living in Hungary (*Bartha* [2003], *Borbély* [2015]). We attempted to identify homogeneous types that characterise adult bilingual minority persons who are undergoing the process of language shift and who were able to speak in an interview in their minority language.

Cluster analysis is a preferred, older method through which any objects (persons, variables, etc.) can be categorised into homogeneous groups, so-called clusters (*Hartigan* [1975]). The speciality of our method was the application of appropriate cluster quality coefficients (EESS%, PB, etc.) by means of which we could find a cluster structure that is acceptable by mathematics-based validation criteria (*t*-tests, MORI coefficients) and could be explained in a sociolinguistic framework as well.

The ten clusters explored identify ten types of speakers in the clearly not linear process of language shift (see Figures 1 and 2, and Table 5). Within the ten clusters, CL6–CL10 are definitely minority affiliated, and they cover about two-thirds of the total sample. (See Figure 2.) It is important to note that in four out of these five clusters the respondents have taken some steps in the direction of language shift, including a weakened minority ethnic identity (CL6), attitude toward their minority language (CL7), minority language choice within their family (CL8), or minority language proficiency (CL9). However, the existence of these clusters is a positive and hopeful fact compared to CL1–CL4, which all represent a final or a near final stage of language shift and assimilation. (See Figure 1 and Table 5.) It is noteworthy that we found only one single cluster (CL5) representing an intermediate state, but still this shows a clear tendency of language shift. (See Figure 1.) This may indicate that the middle part of the language shift process is unstable, individual, and fast, which does not create an easily identifiable intermediate type.

Our results confirm that the gradual but irreversible process of language shift within the nationalities living in Hungary is not of a linear type. It has several identifiable components (language choice in the family, language proficiency, language attitude, ethnic identity, etc.). If they are influenced in a clever and efficient manner, opportunities for sustainable bilingualism may be created. In this way, parallel to the development of the dominant Hungarian language choice, bilingualism could be sustained in terms of relevant linguistic components and attachment to the one's own minority group and language, which has three equal components: an emotional component (attitude toward one's own language and ethnic identity), a language competence component (language proficiency), and a language behaviour component (language choice).

The correlations between the input variables of the cluster analysis were only of weak to moderate size, which provide an opportunity to influence these minority components separately. For example, ethnic identity can be strengthened by various cultural, scientific, social, or sport-related minority programs; minority language competence can be increased by special life-long language courses; positive attitude toward one's own minority language can be increased through public declarations by well-known minority persons about the importance of minority language competence and positive evaluation of minority ethnic identity. Self-governments for each nationality can play a key role in these positive interventions, and the process can also be supported by the national media, if they offer more positive opinions and news and demonstrate a willingness to cooperate with the minority groups living in Hungary and neighbouring countries, along with their respective nations.

Data submitted to statistical analyses in the current paper reflect state of the art information from 2003–2004. The results concerning the relationship between the clusters and the age of respondents (see Figures 4 and 5) justify the statement that in the last 10-12 years, after the death of many older people who were eager to maintain their minority language, the situation has changed. Coincidentally, the process of language shift is not a strict linear process keeping on one fixed trend. The nationalities in Hungary are composed of special, smaller groups, where the process of language shift has slowed down, and ethnic identity and a positive attitude toward one's own minority language have been maintained/sustained, as demonstrated by the results of a current longitudinal study carried out with Romanians living in Hungary (*Vargha–Borbély* [2017]).

References

BARTHA, CS. [2003]: Die Möglichkeiten der Bewahrung der Minderheitensprachen in Ungarn. Über eine soziolinguistische Zweisprachigkeitsuntersuchung im Landesmasstab. In: *Glatz, F.* (ed.): *Sprache und die kleinen Nationen Ostmitteleuropas* (*Begegnungen, Band 21*). Europa Institut. Budapest. pp. 225–236.

BASTARDAS BOADA, A. [2014]: Linguistic sustainability for a multilingual humanity. *Darnioji daugiakalbystė – Sustainable Multilingualism*. Vol. 5. pp. 134–163. http://dx.doi.org/10.7220/2335-2027.5.5

BERGMAN, L. R. – MAGNUSSON, D. – EL-KHOURI, B. M. [2003]: *Studying Individual Development in an Interindividual Context. A Person-Oriented Approach.* Lawrence-Erlbaum Associates. Mahwa, London.

BERGMAN, L. R. – VARGHA, A. – KÖVI, Z. [2017]: Revitalizing the typological approach: some methods for finding types. *Journal for Person-Oriented Research*. Vol. 3. No. 1. pp. 49–62. http://dx.doi.org/10.17505/jpor.2017.04

BORBÉLY, A. [2015]: Studying sustainable bilingualism: comparing the choices of languages in Hungary's six bilingual national minorities. *International Journal of the Sociology of Language*. Issue 236. pp. 155–179. https://doi.org/10.1515/ijsl-2015-0025

COHEN, J. [1977]: *Statistical Power Analysis for the Behavioral Sciences* (*Revised Edition*). Academic Press. New York.

COHEN, J. [1992]: A power primer. *Psychological Bulletin*. Vol. 112. No. 1. pp. 155–159. http://dx.doi.org/10.1037/ 0033-2909.112.1.155

GAL, S. [1979]: *Language Shift: Social Determinants of Linguistic Change in Bilingual Austria*. Academic Press. New York.

HARTIGAN, J. A. [1975]: *Clustering Algorithms.* John Wiley and Sons. New York.

SCHEFFÉ, H. [1959]: *The Analysis of Variance.* John Wiley and Sons. New York.

VARGHA, A. – BERGMAN, L. R. – TAKÁCS, SZ. [2016]: Performing cluster analysis within a person-oriented context: some methods for evaluating the quality of cluster solutions. *Journal for Person-Oriented Research*. Vol. 2. Nos. 1–2. pp. 78–86. http://dx.doi.org/10.17505/jpor.2016.08

VARGHA, A. – BORBÉLY, A. [2017]: Application of modern classification methods in the study of bilingualism. *Glottotheory*. Vol. 8. No. 2. pp. 203–216. https://doi.org/10.1515/glot-2017-0013

VARGHA, A. – TORMA, B. – BERGMAN, L. R. [2015]: ROPstat: a general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research*. Vol. 1. Nos. 1–2. pp. 87–98. http://dx.doi.org/10.175 05/jpor.2015.09