Recognition of Linguistic Features by Hidden Markov Model (HMM)

Božidar Tepeš[1], Lajos Szirovicza[2], Anita Sujoldžić[2] and Martina Primorac[1]

[1] Faculty of Philosophy, University of Zagreb, Zagreb, Croatia
[2] Institute for Anthropological Research, Zagreb, Croatia

Based on recent results in creating automatic taggers for different European languages, including the Croatian language, an attempt has been made to use Hidden Markov Model (HMM) for analyzing linguistic (dialectal) microdifferentiation of reproductively isolated populations in the Eastern Adriatic. As in this geographic area two main dialects are spoken, two different HMM were created, one for the recognition of the "čakavian" dialect, and the other one for the recognition of the "štokavian" dialect. The recognition of the dialects is based on their differential phonetic characteristics. The paper gives a short introduction of HMM as a potential mathematical background for future research and results, the development of HMM for dialect classification ("čakavian" and "štokavian"), description of the corpora available at the moment, and the results obtained.

Keywords: Hidden Markov Model, HMM, stochastic tagging, language processing

1. Introduction

In the natural language processing community, there has been a growing awareness of the key importance that lexical and corpora resources play in the advancement of research in this area as well as in the development of relevant products. During the last two decades of the research of natural languages models have been formed which are based on the frequencies of linguistic traits as well as on their real sequence (e.g., word patterns within a sentence), while the observed linguistic traits are considered to be a sequence of consecutive states (e.g. adjective precedes noun and is followed by verb, etc.). Since the late 1960's hidden Markov models (HMM) have been applied rather successfully for the recognition of speech, as shown in an extensive review by Rabiner (1989). The applicability of hidden Markov models in the recognition of the Croatian language has been tested so far by tagging the Croatian language texts (Tepeš et al., 1996) included in the COST (Corpus of School Texts) collection of Croatian school texts, selected and prepared in machine-readable form. The COST has 51 853 words and 4 671 sentences. In this experiment, a tag set comprising the main grammatical categories: noun, adjective, numeral, pronoun, adverb, verb, preposition, conjunction and sentence break was used. The results obtained by the experiment were within the range of results obtained for other European languages (Dermatas and Kokkinakis, 1995; Rentzepopoulos and Kokkinakis, 1996). Encouraged by this, an attempt has been made to use HMM for analyzing linguistic (dialectal) microdifferentiation of reproductively isolated populations in the Eastern Adriatic. As in this geographic area two main dialects (čakavian and štokavian) are spoken, two different HMM were created, one for the recognition of the "čakavian" dialect, and the other one for the recognition of the "štokavian" dialect. The prerequisites for this included linguistically acceptable state selection for HMM and the objective selection of two groups of villages, each representing one of the dialects in question. The material used for this purpose was collected by the Institute for Anthropological Research, Zagreb, Croatia, within the research aimed at the analysis of linguistic differentiation conducted in the Eastern Adriatic for the last twenty years.

The data base consists of 95 words of the basic vocabulary (including universal items such as mother, father, ashes, dog, rib, etc.), from 48 villages of the islands of Brač, Hvar, Korčula,

Code	Symbol	Distinctive Phonetic Features
0		without accent
1	˜	čakavian acute
2	″	short-falling accent
3	⌒	long-falling accent
4	`	short-rising accent
5	′	long-rising accent
6	-	vowel length
7	•	vowel closure

Table 1

Pag, Silba and Olib, and the peninsula of Peljesac, listed in Table 4. The choice of words is based on the assumption that in all languages there is a "basic vocabulary" related to some universal human categories, the items of which can be equally well applied to all languages. As compared to the so-colled "cultural vocabulary", the basic fund of words is more resistant to changes. It includes a list of approximately 100 common words which exist in every culture. Another assumption of this method is that changes in the basic vocabulary occur at a constant rate which is the same in all languages. Determination of the relationship between two languages is based on the analysis of shared cognates, items which are similar at phonetic, morphological and lexical levels. This data base was previously described in detail and analysed by various clustering methods previously by Sujoldžić et al.: Hvar (1983), Korčula (1986), Olib and Silba (1987), Brač (1988), Peljesac (1989) and Pag (1990). As such a data base appears to be very suitable for the construction and verification of different stochastic models, in this paper the attempt is made to apply the hidden Markov model for the recognition of the dialects, as a contribution to the study of regional linguistic microdifferentiation.

The application of hidden Markov models in the recognition of dialect was based on the construction of two models using distinctive features of the word syllables from the basic vocabulary determined with respect to the čakavian and the štokavian dialects. The villages were subject to both models and their closeness to the observed dialects was determined on the basis of the recognition achievement. The words were divided into syllables containing vowels with their distinctive features. The vocalic "r" was also used as syllabic, regardless of its distinctive features. The model defined in this way produced acceptable results in agreement with the results of previous linguistic investigations, providing at the same time the opportunity of a more detailed analysis at the level of the model. The testing of the model was based on phonetic features presented in Table 1. Table 2 presents 45 coded vowels used in the analysis.

2. Hidden Markov model for distinctive phonetic features

The discrete hidden Markov model may be described with the set of states, the set of observations, the state transition probability distributions, the observation probability distribution and the initial state distribution. In our model the N states are elements of tags $S = \{S(1), S(2), \ldots, S(N)\}$ comprising different vowels with distinctive phonetic features in the syllables of the basic vocabulary words. By combining the vowels with all possible sorts of distinctive features N states were found (Table 2). The set of observations $W = \{W(1), W(2), \ldots, W(K)\}$ comprising the syllables in the words of the basic vocabulary in the analyzed villages or $W = \{W(k)\}$ for $1 \leq k \leq K$, where K is the number of different syllables regardless of their phonetic features.

In the same word we observed the sequence of syllables, $w = \{w(1), w(2), \ldots, w(T)\}$, where T means the number of vowels in the word or $w = \{w(t)\}$ for $1 \leq t \leq T$. T varies from word to word and in the material used its range was 2-8.

In the same word there is a state sequence $s = \{s(1), s(2), \ldots, s(T)\}$ of different vowels with distinctive phonetic features in the syllables or $s = \{s(t)\}$ for $1 \leq t \leq T$. The state

Model states

No.	State	Code	No.	State	Code
1	other		24	3-i-0	312 0
2	0-a-0	0 1 0	25	4-i-0	412 0
3	1-a-0	1 1 0	26	5-i-0	512 0
4	2-a-0	2 1 0	27	6-i-0	612 0
5	3-a-0	3 1 0	28	0-o-0	018 0
6	4-a-0	4 1 0	29	1-o-0	118 0
7	5-a-0	5 1 0	30	2-o-0	218 0
8	6-a-0	6 1 0	31	3-o-0	318 0
9	1-a-7	1 1 7	32	4-o-0	418 0
10	0-a-7	0 1 7	33	5-o-0	518 0
11	3-a-7	3 1 7	34	6-o-0	618 0
12	6-a-7	6 1 7	35	0-o-7	018 7
13	0-e-0	0 8 0	36	3-o-7	318 0
14	2-e-0	2 8 0	37	0-u-0	024 0
15	3-e-0	3 8 0	38	1-u-0	124 0
16	4-e-0	4 8 0	39	2-u-0	224 0
17	5-e-0	5 8 0	40	3-u-0	324 0
18	6-e-0	6 8 0	41	4-u-0	424 0
19	3-e-7	3 8 7	42	5-u-0	524 0
20	6-e-7	6 8 7	43	6-u-0	624 0
21	0-i-0	012 0	44	0-u-3	024 3
22	1-i-0	112 0	45	–r–	20
23	2-i-0	212 0			

Table 2

transition probability distribution are rows of matrix $A = \{a(i,j)\}$ for $1 \leq i,j \leq N$. The elements $a(i,j)$ are conditional probabilities:

$$a(i,j) = P[s(t+1) = S(j) | s(t) = S(i)] \quad \text{for } 1 \leq i,j \leq N \text{ and } 1 \leq t \leq T.$$

The observation state distributions are rows of the matrix $B = \{b(j,k)\}$ for $1 \leq j \leq N$ and $1 \leq k \leq K$. The elements $b(j,k)$ are conditional probabilities:

$$b(j,k) = P[w(t) = W(k) | s(t) = S(j)] \quad \text{for } 1 \leq j \leq N \text{ and } 1 \leq k \leq K.$$

The initial state distribution is matrix $\Pi = \{\pi(i)\}$ for $1 \leq i \leq N$. The elements $\pi(i)$ are probabilities:

$$\pi(i) = P[s(1) = S(i)] \quad \text{for } 1 \leq i \leq N$$

The complete parameter set of hidden Markov model for distinctive phonetic features is $h = (A, B, \Pi)$. Our problem is how to adjust the model parameters $h = (A, B, \Pi)$ to maximise the probability of the observation sequence. The probability of the observation sequence is:

$$P[w|h] = \sum_{s(1),\ldots,s(t)\in S} p(s(1))b(s(1),w(1))a(s(1),s(2)) \ldots a(s(T-1),s(T))b(s(T),w(T))$$

The parameters of the first model $h = (A, B, \Pi)$ are:

$a(i,j)$ = (number of transitions from S(i) to S(j)) / (number of transitions from S(i))

$b(j,k)$ = (number of W(k) in S(j)) / (number of S(j))

$\pi(i)$ = (number of S(i) at $t = 1$)

As the determination of all possible state sequences for a series of syllables leads to a great number of mathematical operations, it is necessary to introduce backward and forward variables as elements of the matrices $F = \{f(t,i)\}$, $G = \{g(t,i)\}$:

$$f(t,i) = P[w(1), w(2), \ldots, w(t), s(t) = S(i), h] \quad \text{for } 1 \leq t \leq T \text{ and } 1 \leq i \leq N,$$

Selection of villages for training

Čakavian villages		Štokavian villages	
Code	Name	Code	Name
01	Bobovišća	14	Sumartin
02	Ložisća	23	Bogomolje
04	Dračevica	31	Račišće
07	Škrip i Splitska	35	Kuna
09	Pražnice	36	Pijavičino
10	Gornji Humac	37	Potomje
15	Dol	42	Dinjiška, Vrčići i Vas
16	Vrbanj	43	Vlašići i Smokvica
17	Svirče	44	Povljana
18	Vrisnik		
19	Pitve		
20	Poljica		

Table 3

i.e., the conditional probability that a series of observations after t syllables is found in the state S(i) starting from the initial state:

$$g(t,i) = P[w(t+1), w(t+2), \ldots, w(T), s(t) = S(i), h] \text{ for } 1 \le t \le T \text{ and } 1 \le i \le N,$$

i.e., the conditional probability that a series of observations after $T-(t+1)$ syllables is found in the state S(i) starting from the last syllable in the word. If we accept the optimal criterium that the most probable individual state is taken into account, the matrices $C = \{c(t,i)\}$ and $D = \{d(t,i,j)\}$ should be defined by the following elements:

$$c(t,i) = P[s(t) = S(i), w, h] \text{ for } 1 \le t \le T \text{ and } 1 \le i \le N,$$

i.e., the conditional probability that t-syllable is in the state S(i) in the word W and model λ:

$$d(t,i,j) = P[s(t) = S(i), s(t+1) = S(j), w, h] \text{ for } 1 \le t \le T-1 \text{ and } 1 \le i,j \le N,$$

where elements d(t, i, j) are conditional probabilities of t-syllable of the word W in the state S(i) and t+1 in the state S(j).

(We can find the matrices F and G using the forward — backward procedure shown by Rabiner, 1989.). The elements of matrices C and D can be expressed in terms of forward-backward variables in the following way:

$$c(t,i) = f(t,i)g(t,i)/P[w|h] \text{ for } 1 \le t \le T \text{ and } 1 \le i \le N,$$

$$d(t,i,j) = f(t,i)a(i,j)b(j,w(t+1))g(t+1,j)/P[w|h] \text{ for } 1 \le t \le T-1 \text{ and } 1 \le i,j \le N.$$

Since the sum c(t, i) for t from 1 to T-1 can be interpreted as the expected number of transitions from the state S(i), and the sum d(t, i, j) as the expected number of transitions from the state S(i) into the state S(j), we can estimate the parameters of model $h = (A, B, \Pi)$ with parameters $h' = (A', B', \Pi')$ where:

$$a'(i,j) = \frac{\sum_{t=1}^{T-1} d(t,i,j)}{\sum_{t=1}^{T-1} c(t,i)} \quad \text{for } 1 \le i,j \le N$$

$$b'(j,k) = \frac{\sum_{t=1}^{T} c(t,j)}{\sum_{t=1}^{T} c(t,i)} \text{ for } 1 \le j \le N \text{ and } 1 \le k \le K$$

$$\pi'(i) = c(1,i) \text{ for } 1 \le i \le N$$

We iteratively use $h' = (A', B', \Pi')$ in place of h and repeat the estimate calculations. The final

Percentage of correctly recognised syllables (POCRS)

Island	Village code	Village name	POCRS by ČHMM	POCRS by ŠHMM	Island	Village code	Village name	POCRS by ČHMM	POCRS by ŠHMM
BRAČ	S01	BOBOVIŠĆA	83.03	31.52	KORČULA	S26	SMOKVICA	54.60	42.94
BRAČ	S02	LOŽIŠĆA	85.37	31.71	KORČULA	S27	ČARA	53.61	40.96
BRAČ	S03	SUTIVAN	57.93	45.73	KORČULA	S28	PUPNAT	40.83	60.95
BRAČ	S04	DRAŽEVICA	86.06	35.15	KORČULA	S29	ŽRNOVO	52.98	47.02
BRAČ	S05	DONJI HUMAC	83.64	30.30	KORČULA	S30	LUMBARDA	43.79	60.95
BRAČ	S06	NEREŽIŠĆA	83.13	33.73	KORČULA	S31	RAČIŠĆE	24.39	84.76
BRAČ	S07	ŠKRIP I SPLITSKA	86.14	29.52	PELJEŠAC	S32	LOVIŠTE	46.06	58.18
BRAČ	S08	DOL I POSTIRA	82.93	31.10	PELJEŠAC	S33	VIGANJ	35.54	74.10
BRAČ	S09	PRAŽNICE	83.73	32.53	PELJEŠAC	S34	KUĆIŠĆE	30.95	76.79
BRAČ	S10	GORNJI HUMAC	85.63	31.74	PELJEŠAC	S35	KUNA	32.74	85.12
BRAČ	S11	SELCA	63.64	45.45	PELJEŠAC	S36	PIJAVIČINO	29.17	87.50
BRAČ	S12	NOVO SELO	62.42	46.67	PELJEŠAC	S37	POTOMJE	31.14	82.63
BRAČ	S13	POVLJA	63.86	46.99	PAG	S38	LUN	49.09	33.94
BRAČ	S14	SUMARTIN	31.33	84.34	PAG	S39	NOVALJA	40.48	36.90
HVAR	S15	DOL	80.79	28.25	PAG	S40	KOLAN, MANDRE I ŠIMUNI	44.31	37.72
HVAR	S16	VRBANJ	80.72	36.14	PAG	S41	PAG	38.32	42.51
HVAR	S17	SVIRČE	84.34	35.54	PAG	S42	DINJIŠKA, VRČIĆI I VAS	35.15	81.21
HVAR	S18	VRISNIK	83.13	35.54	PAG	S43	VLAŠIĆI I SMOKVICA	30.54	84.43
HVAR	S19	PITVE	84.24	32.12	PAG	S44	POVLJANA	33.53	76.65
HVAR	S20	POLJICA	85.98	31.71	PAG	S45	ZUBOVIĆI	47.62	33.93
HVAR	S21	ZASTRAŽIŠĆE	78.44	37.72	PAG	S46	METAJNA	42.17	45.78
HVAR	S22	GDINJ	69.33	47.24	SILBA	S47	SILBA	38.95	27.91
HVAR	S23	BOGOMOLJE	56.21	66.27	OLIB	S48	OLIB	38.10	32.14
KORČULA	S24	VELA LUKA	57.14	36.90					
KORČULA	S25	BLATO	53.61	34.34					

Table 4

result of this estimation procedure is called maximum likelihood estimate of hidden Markov model for distinctive phonetic features.

3. Recognition of Linguistic Features by Hidden Markov Models (HMM)

The corpus of basic vocabulary had 4 560 words. Two HMM were built, one for recognition of the "štokavian" dialect (ŠHMM) and the other one for recognition of the "čakavian" dialect (ČHMM). Each of them had its own training set of villages.

The training set of ŠHMM consisted of the basic vocabulary of 9 villages (Table 3) that represent the so called conservative "štokavian" group. Its learning consisted of three iterations with 855 words divided into 1501 syllables. The training set of ČHMM consisted of the basic vocabulary of 12 villages (Table 3) that represented the so called conservative "čakavian" group and its learning consisted of four iterations with 1140 words divided into 1997 syllables. Both models were tested on each of all 48 villages and the properly recognised percentage of syllables is given in Table 4.

The two hidden Markov models obtained were defined as follows:

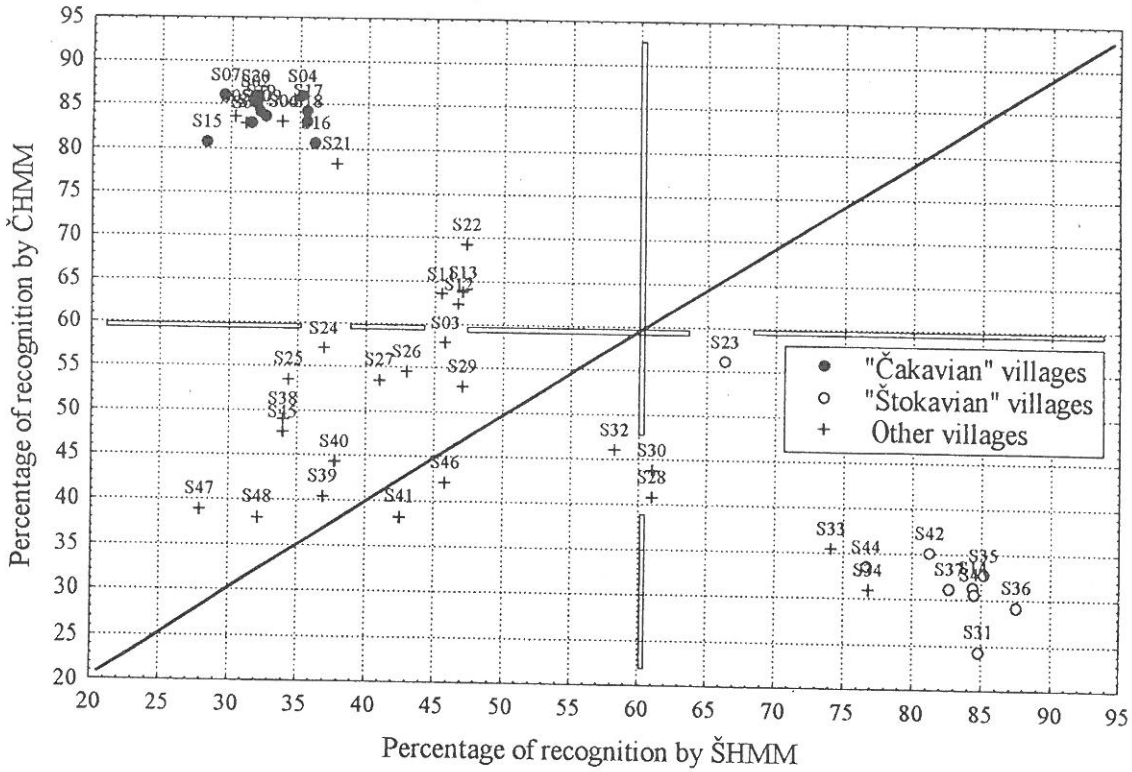

Fig. 1.

- N = 45, number of states;
- S = {S(1), S(2), ..., S(N)}; states defined by distinctive vocalic features given in Table 1.
- K = 201, or 202; number of different syllables for selected 12 čakavian or 8 štokavian villages given in Table 3;
- W = {W(1), W(2), ..., W(K)}, syllables with different phonetic features.

In Figure 1, we can see the partition of villages based on their dialectal differences recognized by HMM. As it can be seen, none of the villages of the conservative "čakavian" group is placed among the villages of the conservative "štokavian" group. On the other hand, none of the villages of the conservative "štokavian" group is placed among the villages of the conservative "čakavian" group. The results obtained by this method were compared to the results obtained previously by other methods, such as clustering, and they were within the range.

If the boundary between the čakavian and štokavian groups is set to more than 60% of successfully recognised syllables, the situation shown in Figure 1 is obtained, where the mentioned criterion is presented by bold horizontal and vertical lines. The consistency of the model is confirmed by the fact that there are no villages in the upper right quadrant. The lower right quadrant, in addition to selected "štokavian" villages, includes also villages S33 (Viganj) and S34 (Kučišće), while the group of villages S28 (Pupnat), S30 (Lumbarda) and S32 (Loviste) are situated on the boundary of the "štokavian" area. According to previous linguistic analyses (Sujoldžić, 1991) all these settlements have been subjected to a process of strong stokavization. The upper left quadrant, in addition to selected "čakavian" villages, includes also other conservative "čakavian" villages, which make quite a compact group. This quadrant includes also another compact group of villages: S11 (Selca), S12 (Novo Selo), S13 (Povlja) and S22 (Gdinj). All these villages are situated in the eastern parts of the islands of Brač and Hvar which have been exposed to immigration of the "štokavian" population since the 17th century. The lower left quadrant contains the villages from the islands of Korčula, Pag, Silba and Olib. The position of these villages reflects their autochtonous "čakavian" substrata and different

extents of immigrating "štokavian" superstrata as well as their geographic, reproductive and socio-cultural isolation, which contributed to the development of specific local features within each local speech (Sujoldžić, 1991). The villages of the south- eastern part of the island of Pag were selected in the "štokavian" group at the beginning of the analysis. The position of the islands of Silba and Olib reflects their long period of isolation which brought about specific "čakavian" features.

4. Conclusion

This paper shows that it is possible to use HMM for the recognition of dialect differences in the Croatian language. It is necessary to say that both models were created primarily for recognition of extreme groups of villages of both "štokavian" and "čakavian" groups. However, as the main problem lies in the recognition of villages whose dialect is neither purely "štokavian" nor "čakavian", it should be solved with a new and larger training and testing set including also an extended set of dialect categories, which will be done in our future work.

References

[1] Dermatas, E., G. Kokkinakis, (1995) Automatic Stochastic Tagging of Natural Language Texts, Computational Linguistic **21**, 137–163.

[2] Rabiner, R. L., (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE **77**, 257–386.

[3] Rentzepopoulos, P. A., G. Kokkinakis, (1996) Efficient Multilingual Phoneme-to-Grapheme Conversion based on Hidden Markov Model, Computational Linguistic **22**, 351–376.

[4] Sujoldžić, A., L. Szirovicza, P. Simunović, B. Finka, D. F. Roberts, P. Rudan, (1983) Lingvističke udaljenosti na otoku Hvaru, Rasprave Zavoda za jezik **8–9**, 197–214.

[5] Sujoldžić, A., P. Simunović, B. Finka, P. Rudan, (1986) Bazični vokabular otoka Korčule kao pokazatelj jezične mikroevolucije, Filologija **14**, 313–328.

[6] Sujoldžić, A., B. Finka, P. Simunović, A. Chaventre, P. Rudan, (1987) Jezična mikroevolucija otoka Silbe i Oliba (Analiza bazičnog rječnika), Rasprave Zavoda za jezik **13**, 107–115.

[7] Sujoldžić, A., B. Finka, P. Simunović, P. Rudan, (1988) Sličnost i razlike u govorima otoka Brača kao odraz migracijskih kretanja, Rasprave Zavoda za jezik **14**, 163–184.

[8] Sujoldžić, A., P. Simunović, B. Finka, L. A. Bennett, P. Rudan, (1989) Jezične udaljenosti na popuotoku Peljescu, Zbornik rasprava iz slovanskega jezikoslovja Tinetu Logarju ob sedamdesetletnici, SAZU Ljubljana, pp. 275– 291.

[9] Sujoldžić, A., B. Finka, P. Simunović, P. Rudan, (1990) Lingvističke udaljenosti otoka Paga, Filologija **18**, 7–37.

[10] Sujoldžić, A., (1991) The Population Study of Middle Dalmatia: Linguistic History and Current Regional Differentiation of Croatian Dialects, Coll. Antropology **15**, 309–320.

[11] Tepeš, B., T. Žubrinić, L. Szirovicza, I. Hunjet, (1996) Hidden Markov Model for Tagging of Croatian Language Texts, ICITI, Pula.

Received: October, 1997
Accepted: January, 1998

Contact address:
Božidar Tepeš and Martina Primorac
Faculty of Philosophy
University of Zagreb
Ivana Lucića 3
10 000 Zagreb
Croatia
e-mail: btepes@ffzg.hr

Lajos Szirovicza and Anita Sujoldžić
Institute for Anthropological Research
Ulica grada Vukovara 72/4
10 000 Zagreb
Croatia

BOŽIDAR TEPEŠ graduated from the Faculty of Science, University of Zagreb in 1967, and received his M.Sc. from the same University in 1974 and Sc.D. degree in 1981. He is professor at the Faculty of Philosophy, University of Zagreb and Head of the University Studies of Business Informatics. His research interests include mathematical modelling and computational linguistics.

LAJOS SZIROVICZA graduated from the Faculty of Science, University of Zagreb in 1971, and received his M.Sc. from the same University in 1981, and Sc.D. degree in informatics from the Faculty of Philosophy, University of Zagreb in 1997. He is currently a research assistant at the Institute for Anthropological Research in Zagreb. His research interest is mathematical and statistical modelling of the structures of human populations.

ANITA SUJOLDŽIĆ is currently senior research associate and head of the Department for Anthropological Linguistics and Sociocultural Research at the Institute for Anthropological Research in Zagreb. She graduated from the Faculty of Arts, University of Zagreb in 1975, and received her M.Sc. in ethnology in 1981 and Ph.D. degree in anthropology in 1985, from the University of Belgrade. Her research interests include anthropological demography and linguistics, with particular focus on linguistic change and differentiation.

MARTINA PRIMORAC graduated from the Faculty of Science, University of Zagreb in 1996. She is assistant at the Faculty of Philosophy, University of Zagreb. Her research interests include mathematical and statistical modelling.