ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

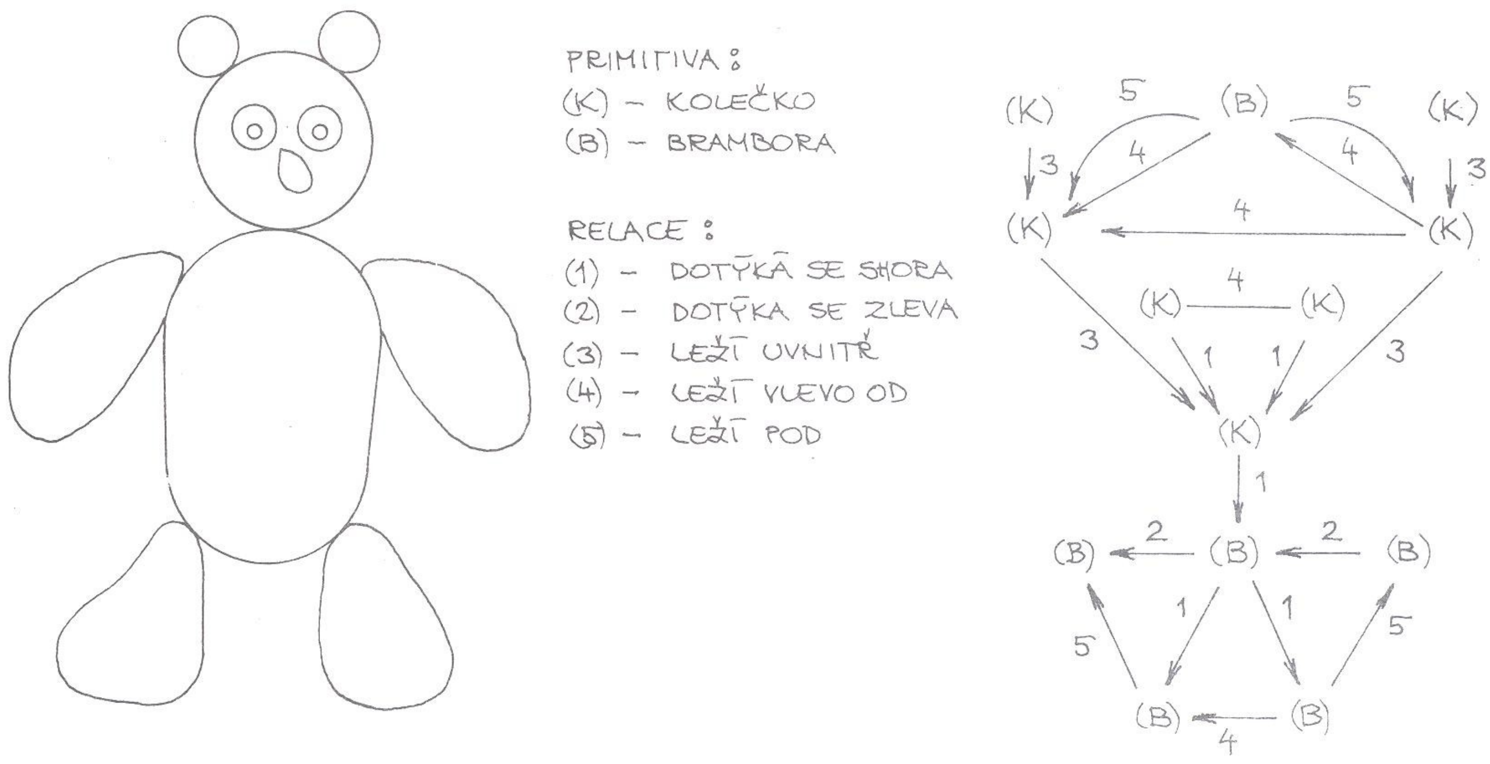
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

STRUKTURÁLNÍ KLASIFIKACE

STRUKTURÁLNÍ POPIS

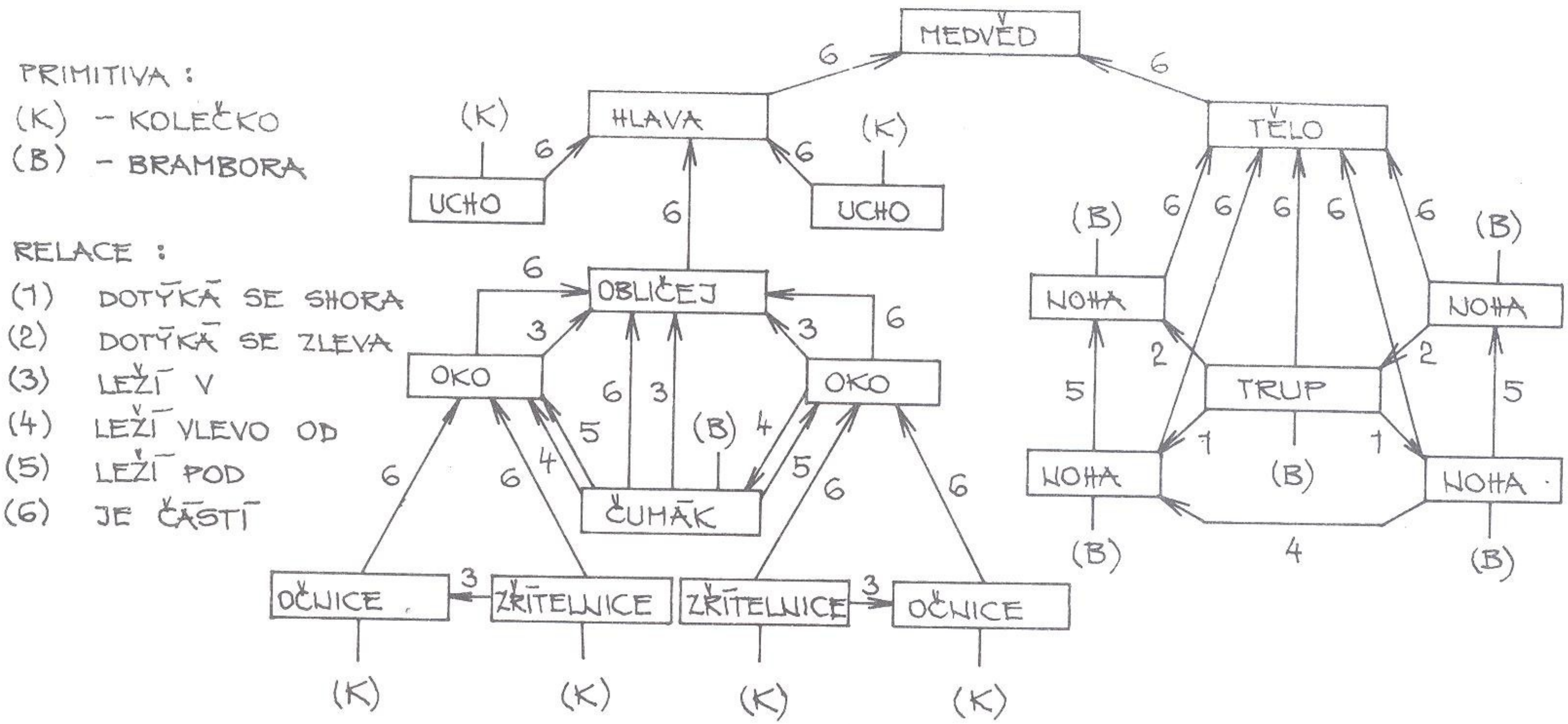
- relační struktura je vytvořena z určitých elementárních popisných částí dat, tzv. primitiv a vzájemných vztahů mezi nimi – relacemi;
- relační struktury zpravidla vyjadřujeme pomocí grafů;

STRUKTURÁLNÍ POPIS



Obr. 3.1 Primitiva, relace a relační struktura čarové kresby

STRUKTURÁLNÍ POPIS

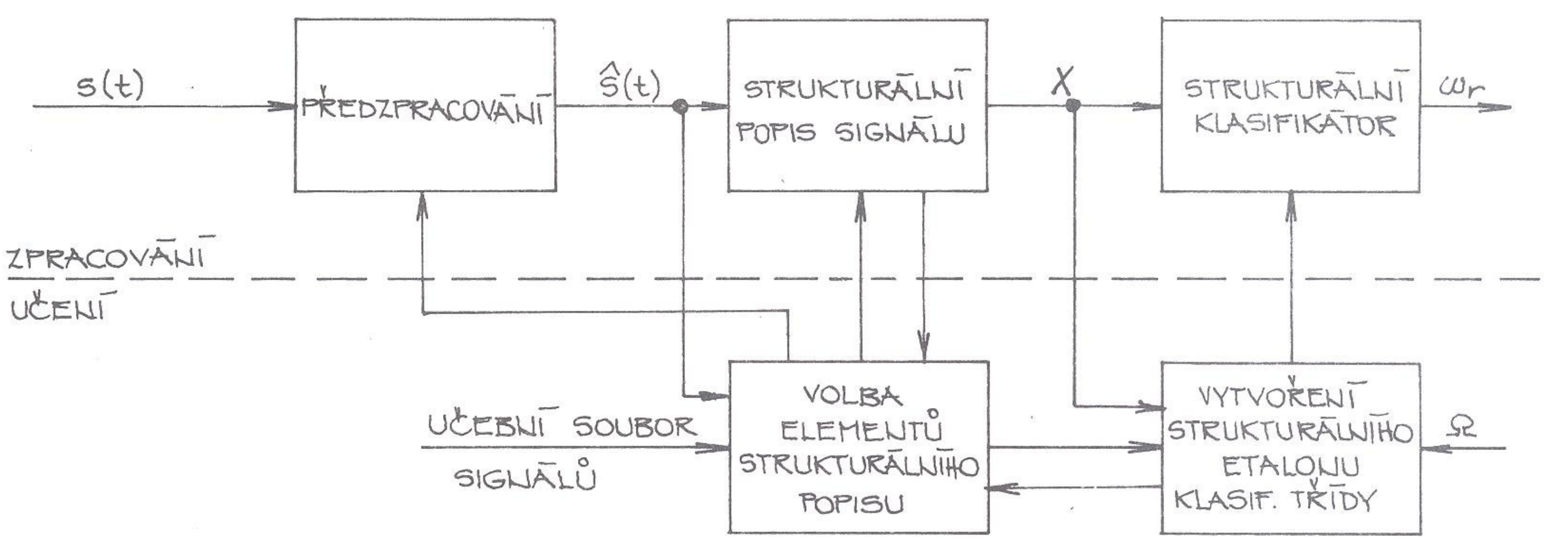


Obr. 3. 4 Hierarchická relační struktura kresby z obr. 3. 1

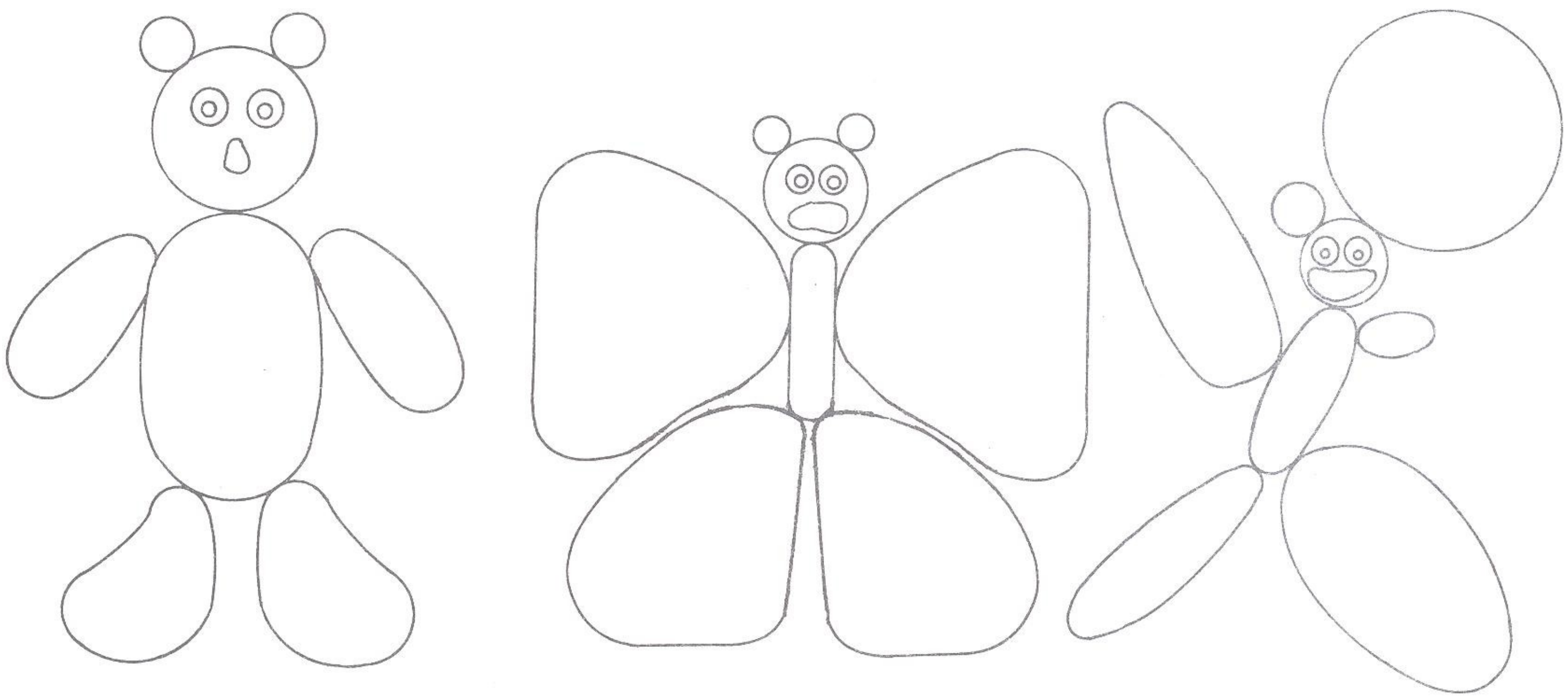
TYPY RELAČNÍCH STRUKTUR

- **řetězce** (uvuuyzuvw)
- **pole** (především pro reprezentaci 2D obrazů)
- **stromy** (relační struktura neobsahující cykly a paralelní cesty)
- **obecné grafy**

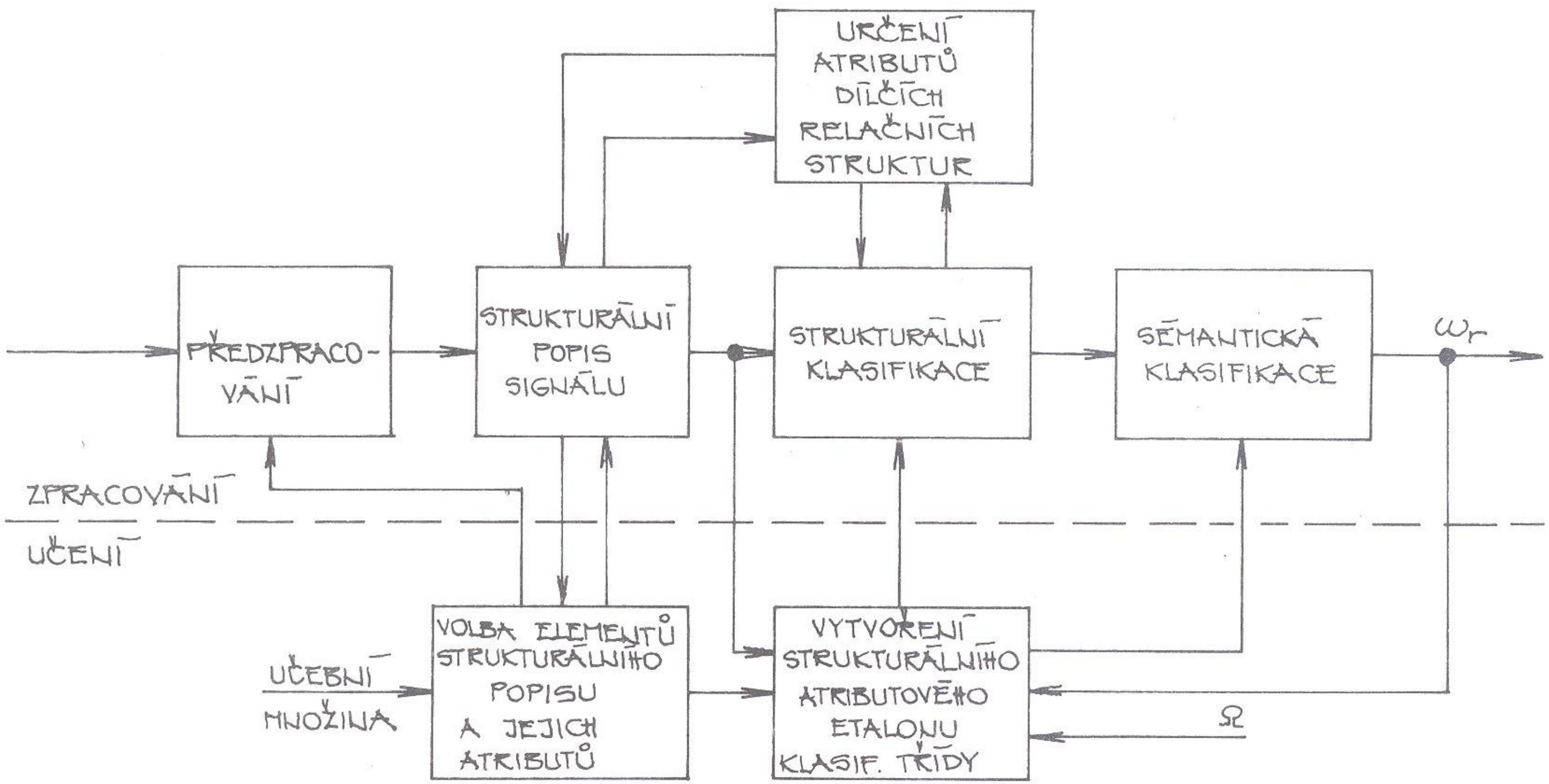
STRUKTURÁLNÍ POPIS



STRUKTURÁLNÍ POPIS



KOMBINOVANÝ STRUKTURÁLNÍ POPIS



Obr. 3. 5 Blokové schéma atributového strukturálního klasifikátoru

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur

  (může být bohužel velký až nekonečný)

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
  - ➔ Gramatika je čtveřice $G = (V_n, V_t, P, S)$, kde V_n a V_t jsou konečné disjunktní množiny (abecedy), přičemž prvky množiny V_n nazývají neterminální pomocné symboly a prvky V_t terminální symboly, $S \in V_n$ je tzv. axiom gramatiky nebo také počáteční symbol a P je množina substitučních pravidel tvaru $\alpha \to \beta$, které definují způsob náhrady dílčí relační struktury α novou strukturou β.

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur
  (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
  - ➔ Příklad gramatiky:
    G = ((A,B), (0,1), P, A), P = (A → 0B1, 0B→00B, B → „e“).

  Příklad generování řetezce:

  A → 0B1 → 00B1 → 000B1 → 0001

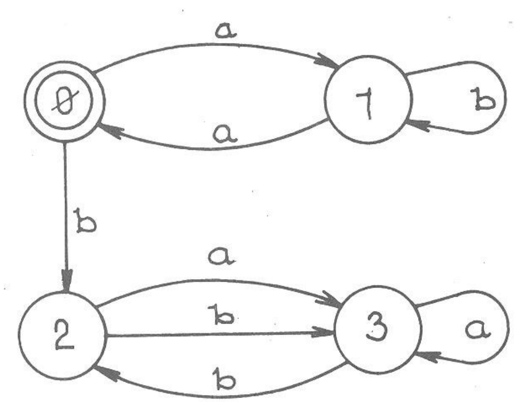
REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur
  (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
- ☑ příjemcem relačních struktur – automat
  - ➔ různé typy automatů podle charakteru relační struktury a substitučních pravidel – nejjednodušší konečný automat
  - ➔ Konečný stavový automat A je pětice A = (X, S, s_0, S_c, □), kde X je konečná vstupní abeceda, S je množina vnitřních stavů, s_0 je počáteční stav automatu, S_c je množina cílových stavů automatu a □: X x S → S je přechodová funkce.

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
- ☑ příjemcem relačních struktur – automat
  - ➔ Příklad konečného stavového automatu

δ	0	1	2	3
a	1	0	3	3
b	2	1	3	2



REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- výčtem relačních struktur
  (může být bohužel velký až nekonečný)
- generátorem relačních struktur – gramatika
- příjemcem relačních struktur – automat
- ekvivalence gramatiky a automatu – gramatika a automat jsou ekvivalentní, pokud množina relačních struktur generovaná gramatikou a množina akceptovaná automatem jsou stejné

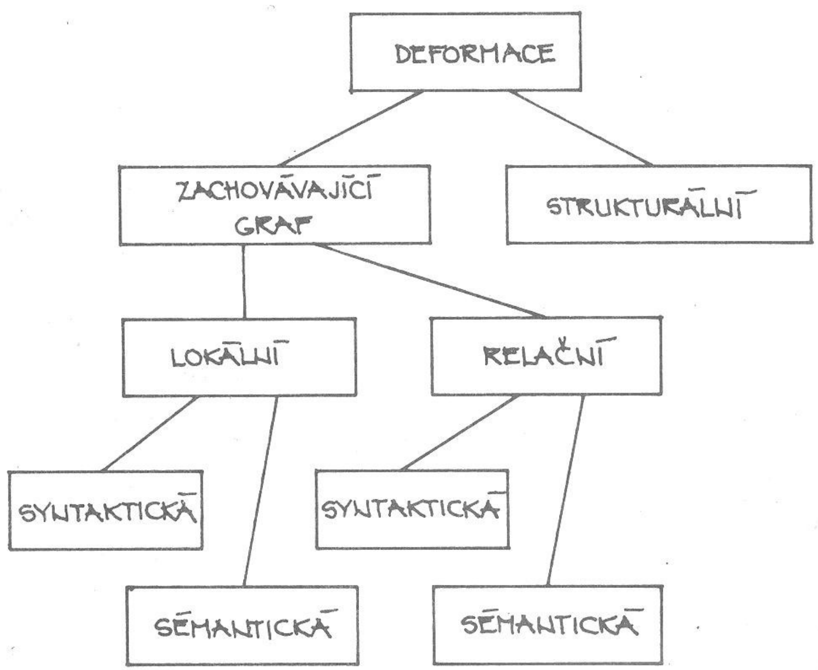
STRUKTURÁLNÍ KLASIFIKACE

nedeformované relační struktury

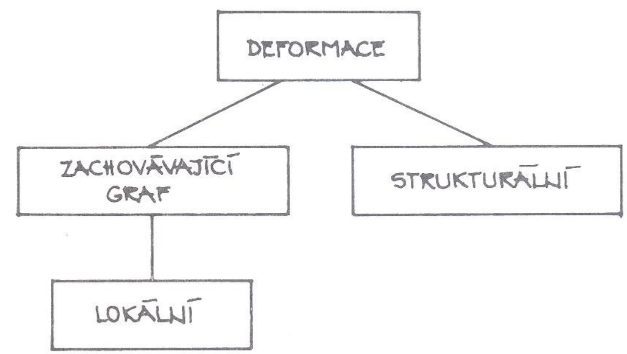
- ☑ ztotožnění s reprezentativními relačními strukturami;
- ☑ přijetí automatem

STRUKTURÁLNÍ KLASIFIKACE

deformované relační struktury

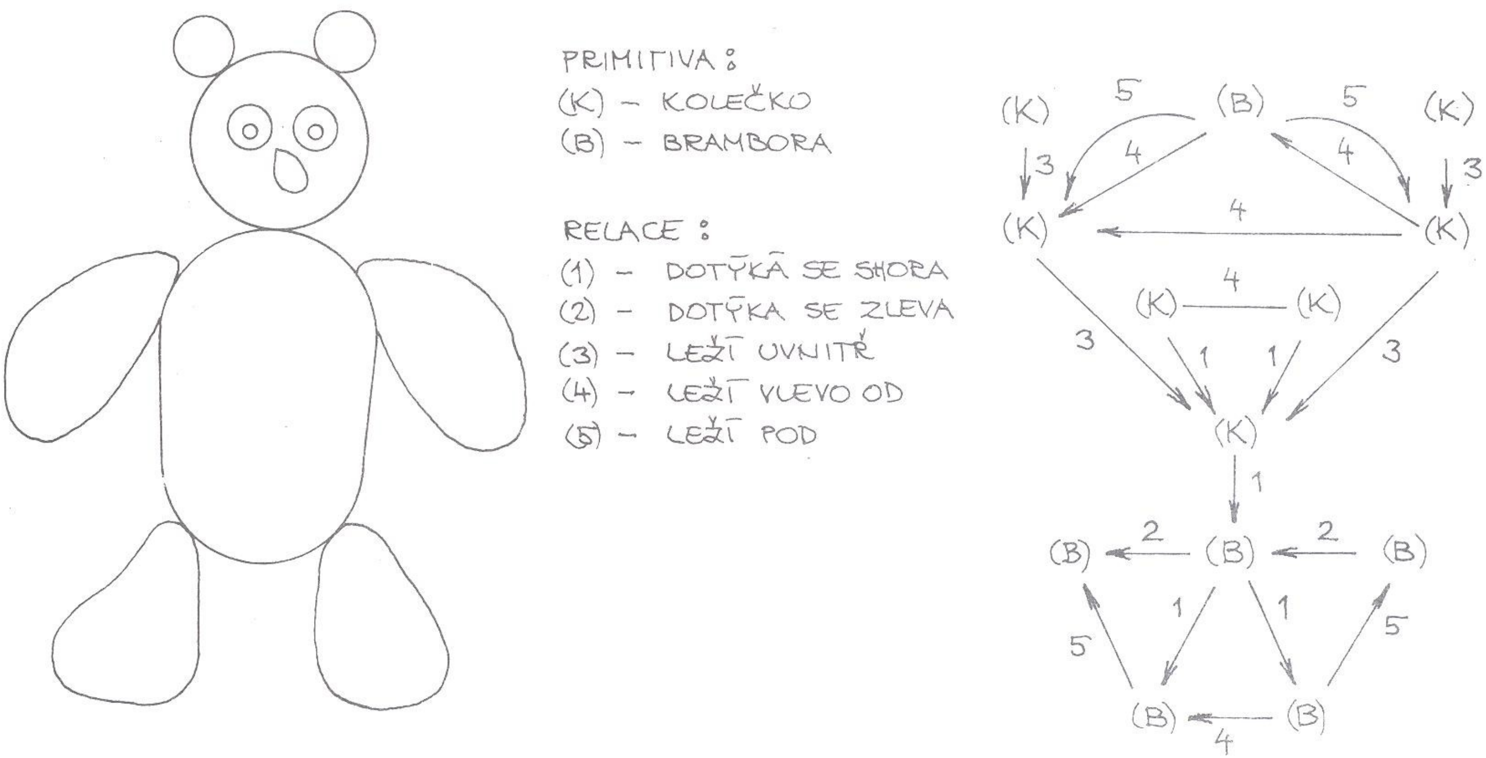


Obr. 3. 21 Obecné strukturální deformační schéma



Obr. 3. 22 Deformační schéma pro řetězce bez sémantické informace

STRUKTURÁLNÍ POPIS



Obr. 3.1 Primitiva, relace a relační struktura čarové kresby

STRUKTURÁLNÍ KLASIFIKACE

deformované relační struktury

- ☑ podle vzdálenosti od etalonu

jak vzdálenost určíme?

STRUKTURÁLNÍ KLASIFIKACE

deformované relační struktury

- ☑ podle vzdálenosti od etalonu

jak vzdálenost určíme?

vzdálenost vs. metrika

?

METRIKA - VZDÁLENOST

Metrika ρ na X je funkce ρ: X X $\rightarrow$ R, kde R je množina reálných čísel, taková, že:

$$\exists \rho_0 \in R: -\infty < \rho_0 \leq \rho(x,y) < +\infty, \forall x,y \in X$$

$$\rho(x,x) = \rho_0, \forall x \in X$$

a

$$\rho(x,y) = \rho(y,x), \forall x,y \in X.$$ *(symetrie)*

Když dále

$$\rho(x, y) = \rho_0$$ když a jen když $x = y$ *(totožnost)*

a $\rho(x, z) \leq \rho(x, y) + \rho(y, z), \forall x,y,z \in X.$ *(Δ nerovnost)*

Prostor X, ve kterém metrika ρ definována, nazýváme metrickým prostorem.

Vzdálenost je hodnota určená podle metriky.

METRIKA PODOBNOSTI - PODOBNOST

Metrická míra podobnosti s na X je funkce s: $X \quad X \to R$, kde R je množina reálných čísel, taková, že:

$$\exists s_0 \in R: -\infty < s(x,y) \leq s_0 < +\infty, \forall x,y \in X$$

$$s(x,x) = s_0, \forall x \in X$$

a

$$s(x,y) = s(y,x), \forall x,y \in X. \textit{ (symetrie)}$$

Když dále

$$s(x,y) = s_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

a $$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X.$$

(Δ nerovnost ?)

STRUKTURÁLNÍ VZDÁLENOST

V případě řetězců lze deformační vlivy vyjádřit (na úrovni primitiv) trojicí tzv. elementárních deformačních transformací – eliminací, substitucí a inzercí

a) eliminační deformační transformace

$$T_E: \quad \omega_1 a \omega_2 \xrightarrow{W_E(a)} \omega_1 \omega_2;$$

b) substituční deformační transformace

$$T_S: \quad \omega_1 a \omega_2 \xrightarrow{W_S(a,b)} \omega_1 b \omega_2;$$

c) inzerční deformační transformace

$$T_I: \quad \omega_1 \omega_2 \xrightarrow{W_I(b)} \omega_1 b \omega_2,$$

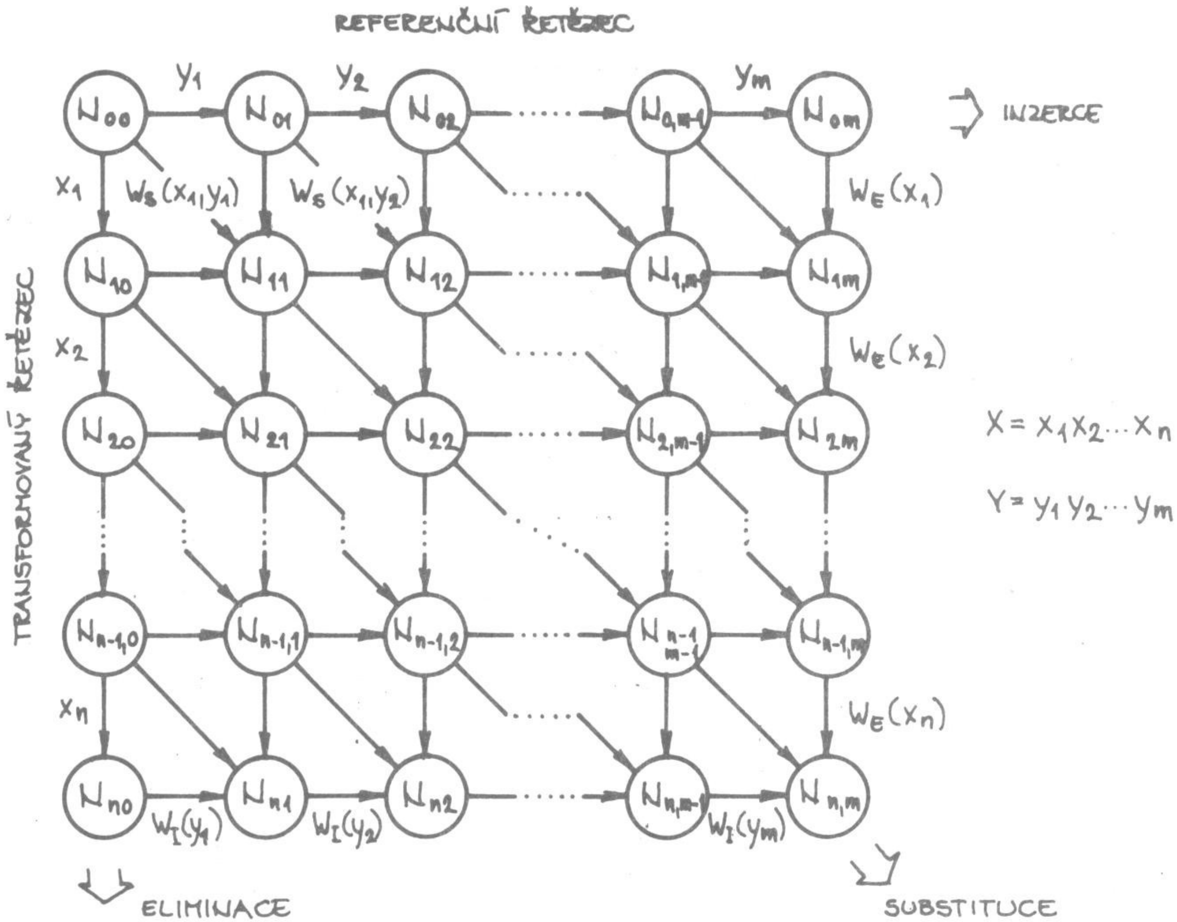
VÁHOVANÁ LEVENŠTEJNOVA METRIKA

Je-li $\mathcal{J} = \{T_1, T_2, \ldots, T_n\}$, $n \geq 0$, $T_i \in \{T_E, T_S, T_I\}$ posloupnost elementárních deformačních transformací taková, že pro libovolná konečná slova X,Y nad abecedou $\mathcal{V}_{te} = \mathcal{V}_t \cup \{"e"\}$ je $Y = \mathcal{J}(X)$, pak váhovaná Levenštejnova metrika je definována vztahem

$$d_{WL}(X,Y) = \min_{\mathcal{J}} \left\{ \sum_{\substack{\forall a \\ [T_E(a) \in \mathcal{J}]}} W_E(a) + \sum_{\substack{\forall a,b \\ [T_S(a,b) \in \mathcal{J}]}} W_S(a,b) + \sum_{\substack{\forall b \\ [T_I(b) \in \mathcal{J}]}} W_I(b) \right\} \quad (3\text{-}28)$$

Aby byly splněny všechny tři základní axiomy metrik (axiom totožnosti, symetričnosti a trojúhelníková nerovnost) je třeba, aby platilo $W_I(a) = W_E(a)$ a $W_S(a,b) = W_S(b,a)$ pro všechny terminální symboly a,b. Metrika splňující tyto požadavky je pravá metrika.

VÁHOVANÁ LEVENŠTEJNOVA METRIKA



Obr. 3. 23 Výpočet váhované Levenštejnovy vzdálenosti

DALŠÍ STRUKTURÁLNÍ METRIKY

- řetězce
  - prostá (neváhovaná) Levenštejnova metrika
  - Hammingova metrika
- stromy

:

:

KLASIFIKACE DEFORMOVANÝCH STRUKTUR

- ☑ výpočet vzdálenosti mezi reprezentativními strukturami (etalony) klasifikační třídy a klasifikovanou strukturou;
- ☑ začlenění deformačních pravidel do substitučních pravidel gramatiky (resp. automatu);

VYTVOŘENÍ STRUKTURÁLNÍHO ETALONU MNOŽINY PÍSMEN

TESTOVACÍ MNOŽINY ŘETĚZCŮ REPREZENTUJÍCÍCH PÍSMENA

a) testovací množiny a jejich charakteristické řetězce

D:
(bbb+ddcbaa)*
(bbb+dxddcbbaa)*
(bbb+dxddbbad)*
(bbbbb+dddcbaad)*
(bb+ddcbdd)*
bbb+ddaabcddd

(bbb+dddcbad)*

F:
bb+(b+dd)xd
bb+(bb+dd)xd
bbb+(dx(b+dd))xdd
bb+(bb+d)xdd
bb+(bb+dd)xdd
bb+(b+ddd)xd

---

bb+(bb+dd)xd

H:
b+bbxdddd+bbxb
b+bbbxd+bxdxb
bb+bbxddd+axbb
ba+abxdd+axbbb
b+dxabxdd+bxbb
ba+bxa+abxbbb

---

b+bbxdd+bxbb

K:
ba+bbxaaxcc
bb+bbxaaxcc
bb+bbbxaaxbcc
b+bbbxaaxbc
baa+bbxaaxcc
bb+bbxa+axcb

---

bb+bbxaaxcc

P:
bbb+(b+dddbdd)*
bb+(bbb+dddbaa)*
bbb+(bbadcbad)*
b+(bb+ddcbad)*
bb+(bb+dxddcaad)*

---

bb+(bb+dddbad)*

U:
cbbbxdabbbb
cbbxda+bbxb
bbbxddabb
cbbxdaabb
bbbxda+bbxb
cbbxdabbbb

---

cbbxdabbb

X:
aaa+cddxaaxcbb
aaa+cxaaxb
aa+cbxaaxcc
aa+ccxaxc
aa+ccxaaxcc
aa+cxaaxcc
baa+ccxbbxccc

---

aa+ccxaaxcc

Y:
bbbb+ccxbb
bbb+cxaa
bbb+bcxaa
bb+cbxaaa
ba+bcxaa
dab+cbxba
bbb+cxa

---

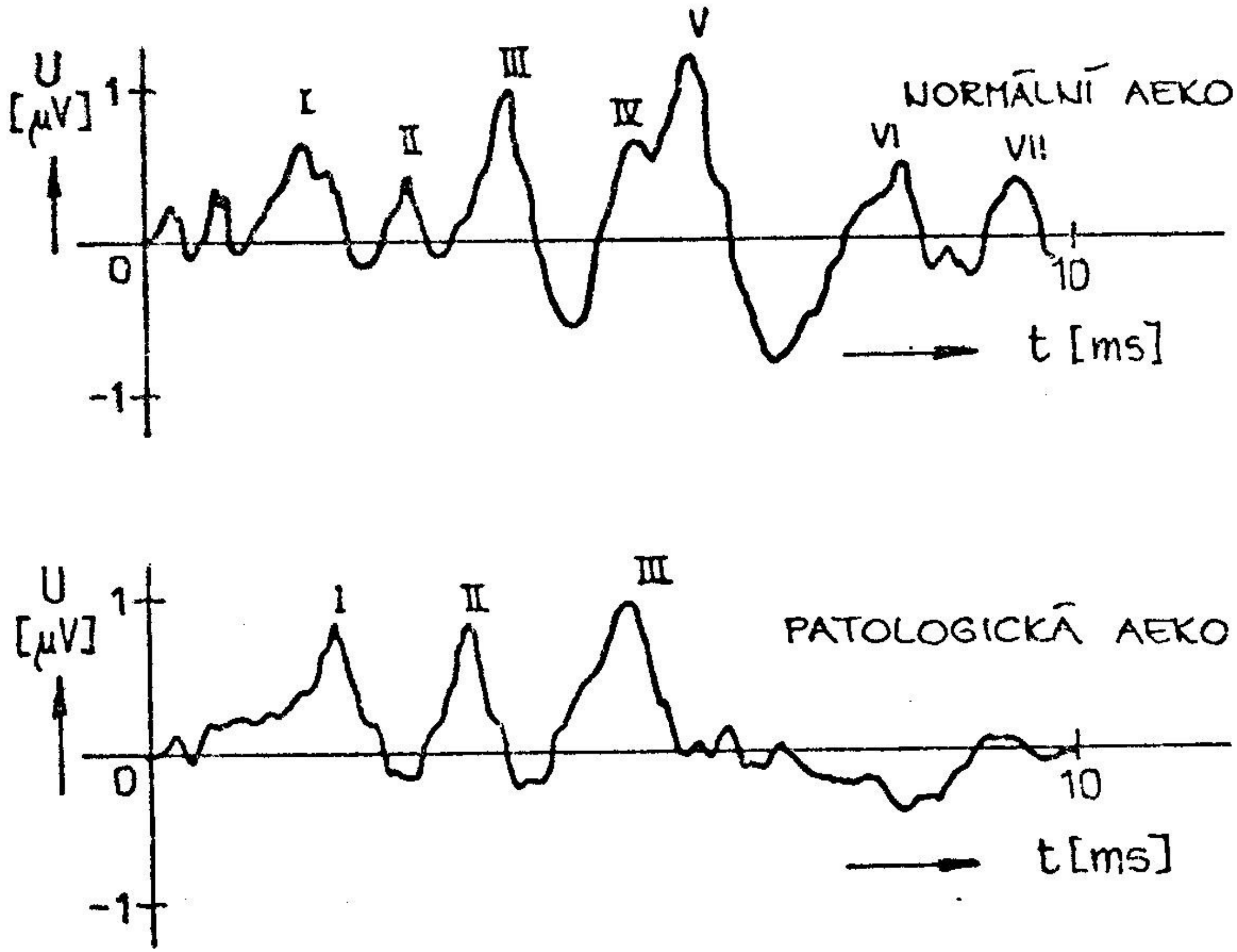
bbb+cxaa

VYTVOŘENÍ STRUKTURÁLNÍHO ETALONU MNOŽINY PÍSMEN

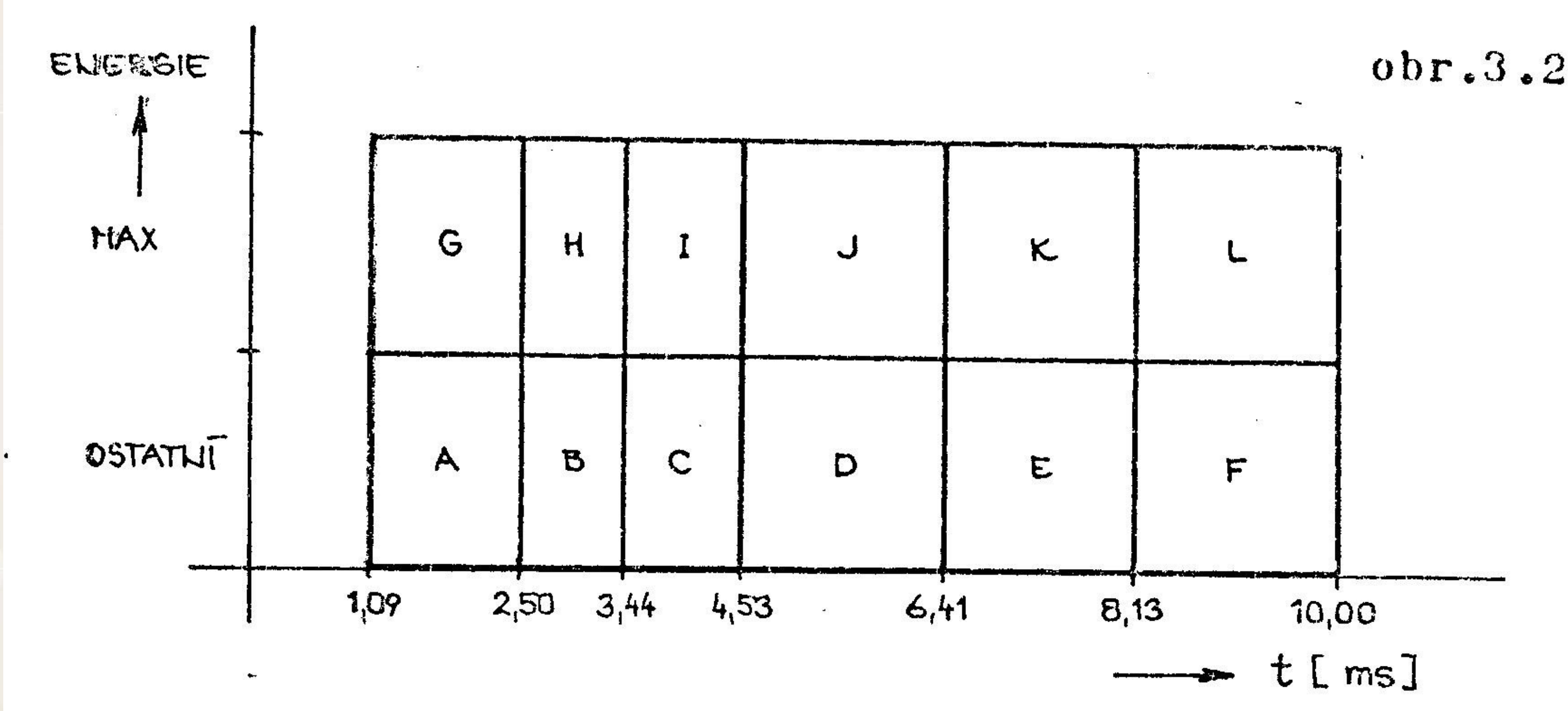
STRUKTURÁLNÍ KLASIFIKACE

DALŠÍ PŘÍKLAD ZE ŽIVOTA

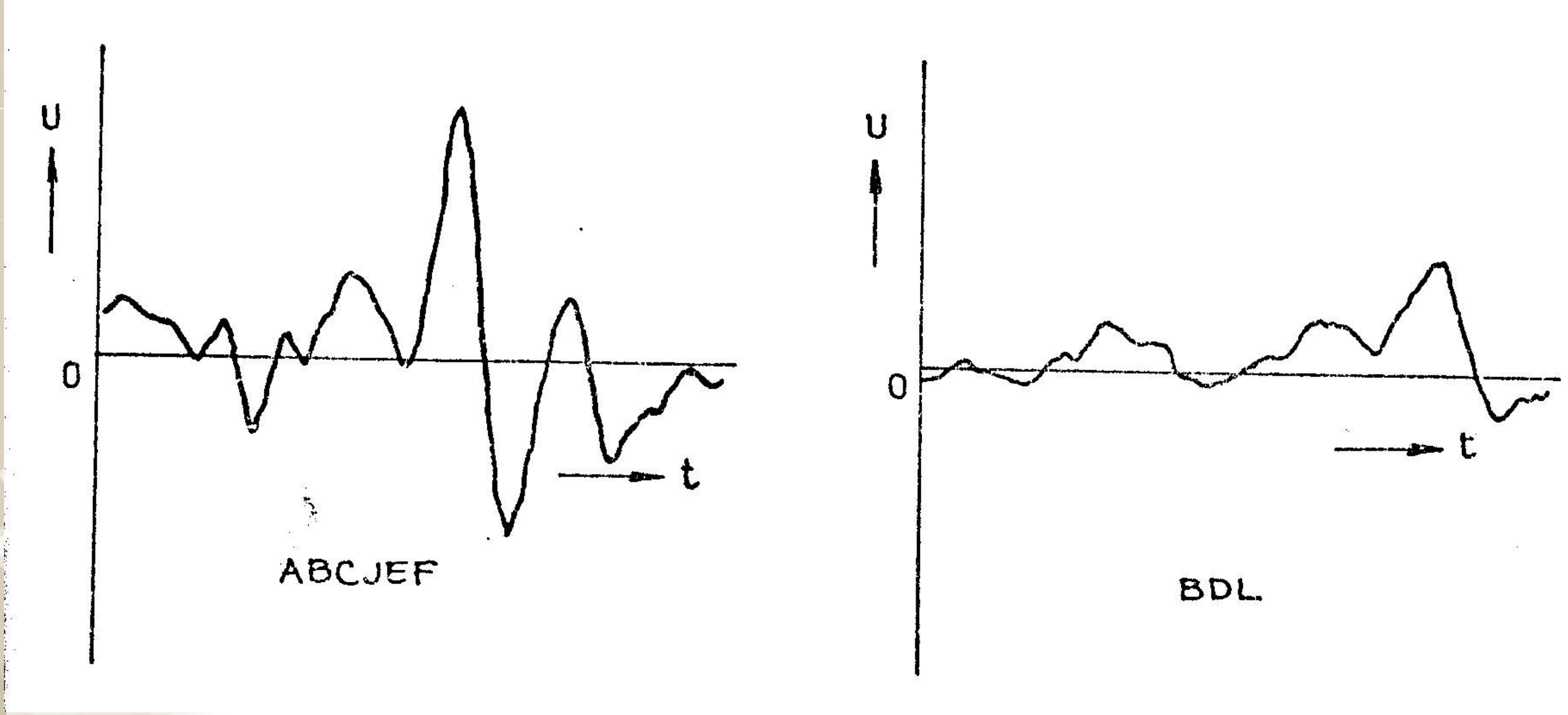
STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ



STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ



STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ



STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ

index	řetězec	četnost výskytu	index	řetězec	četnost výskytu
1	ABCJEF	6	10	ABCDDEL	1
2	ABCJE	3	11	ACJE	1
3	ACJEFF	1	12	ACDJE	1
4	GBCDDEF	3	13	ABCJFF	1
5	AIDDE	2	14	ABIDDF	2
6	GBCDDE	1	15	ACDDK	1
7	ABCJ	2	16	ABCDDKF	1
8	ACJEF	2	17		
9	ABCJF	1			

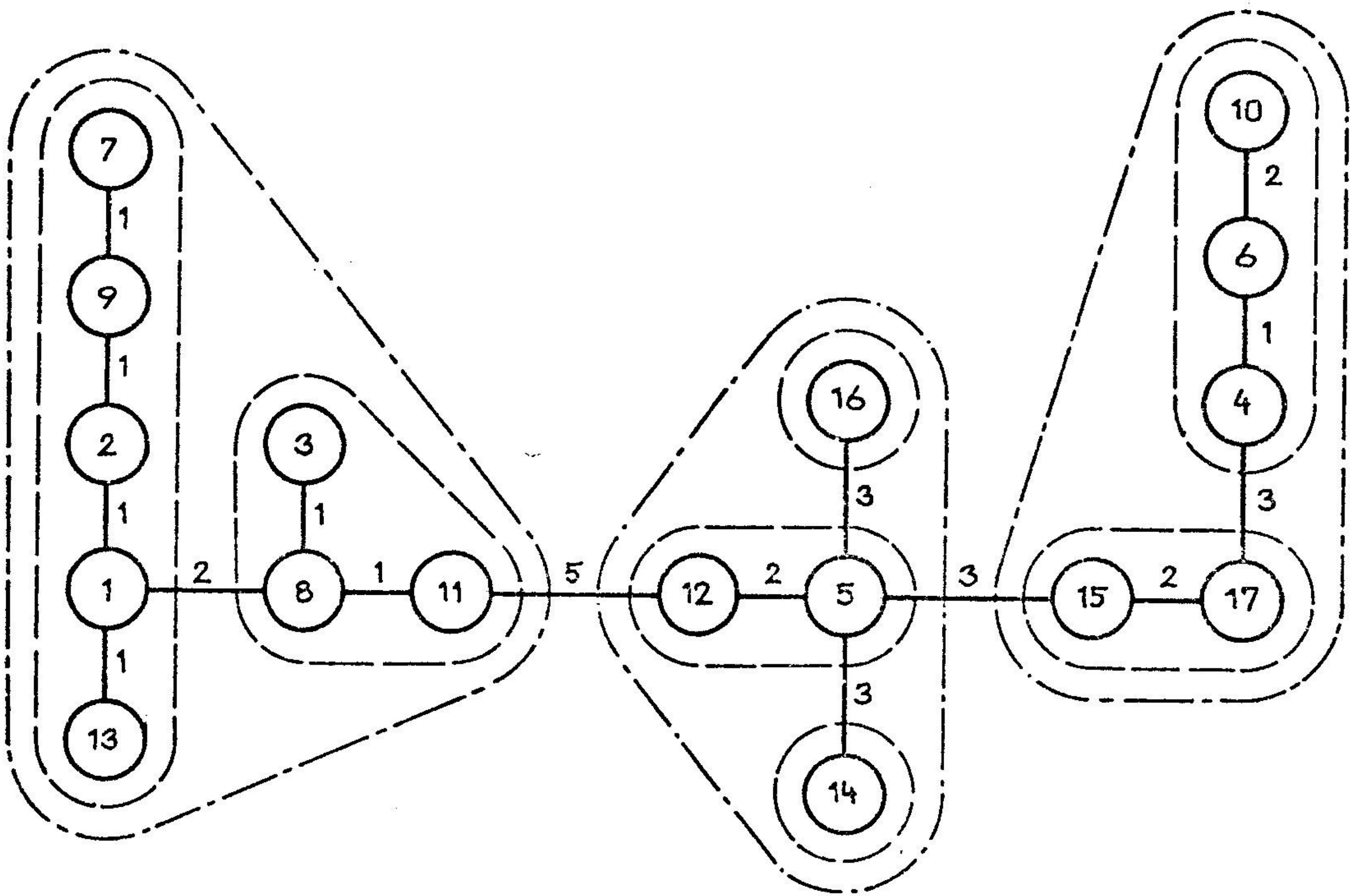
STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ

Symbol	"e"	A	B	C	D	E	F	G	H	I	J	K	L
"e"	0	20	2	15	5	2	1	20	5	15	10	1	1
A	20	0	10	100	100	100	100	1	100	100	100	100	100
B	2	10	0	10	100	100	100	100	5	100	100	100	100
C	15	100	10	0	10	100	100	100	100	1	100	100	100
D	5	100	100	10	0	10	100	100	100	100	1	100	100
E	2	100	100	100	10	0	1	100	100	100	100	2	6
F	1	100	100	100	100	1	0	100	100	100	100	6	2
G	20	1	100	100	100	100	100	0	20	100	100	100	100
H	5	100	5	100	100	100	100	20	0	20	100	100	100
I	15	100	100	1	100	100	100	100	20	0	20	100	100
J	10	100	100	100	1	100	100	100	100	20	0	20	100
K	1	100	100	100	100	2	6	100	100	100	20	0	20
L	1	100	100	100	100	6	2	100	100	100	100	20	0

STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ

Index	Řetězec	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	ABCJEF	0	1	3	7	10	8	3	2	2	8	3	8	1	11	9	11	8
2	ABCJE	1	0	4	8	9	7	2	3	1	7	2	7	2	10	8	10	8
3	ACJEFF	3	4	0	10	9	11	6	1	5	11	2	7	4	10	12	10	11
4	GBCDDEF	7	8	10	0	5	1	10	9	9	3	10	5	8	4	4	6	3
5	AIDDE	10	9	9	5	0	4	11	8	10	4	7	2	11	3	3	3	5
6	GBCDDE	8	7	11	1	4	0	9	10	8	2	9	4	9	3	3	5	3
7	ABCJ	3	2	6	10	11	9	0	5	1	9	4	9	2	10	8	9	8
8	ACJEF	2	3	1	9	8	10	5	0	4	10	1	6	3	9	11	9	10
9	ABCJF	2	1	5	9	10	8	1	4	0	8	3	8	1	9	7	10	7
10	ABCDDEL	8	7	11	3	4	2	9	10	8	0	9	4	9	5	3	5	3
11	ACJE	3	2	2	10	7	9	4	1	3	9	0	5	4	8	10	8	10
12	ACDJE	8	7	7	5	2	4	9	6	8	4	5	0	9	3	5	3	5
13	ABCJFF	1	2	4	8	11	9	2	3	1	9	4	9	0	10	8	11	8
14	GCDDF	11	10	10	4	3	3	10	9	9	5	8	3	10	0	4	3	4
15	ABIDDF	9	8	12	4	3	3	8	11	7	3	10	5	8	4	0	5	2
16	ACDDK	11	10	10	6	3	5	9	9	10	5	8	3	11	3	5	0	3
17	ABCDDKF	8	8	11	3	5	3	8	10	7	3	10	5	8	4	2	3	0

STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ



STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ

Minimální kostra (*minimum spanning tree*) je graf, který spojuje všechny objekty tak, že se zde nevyskytují žádné smyčky a zároveň součet délky spojnic mezi uzly (objekty) je minimální

STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ

Řetězec	Četnost	1. skupina shluků	2. skupina shluků
ABCJEF	6	Ia	Ib
ABCJE	3		
ABCJ	2		
ABCJF	1		
ABCJFF	1		
ACJEFF	1	IIa	
ACJEF	2		
ACJE	1		
AIDDE	2	IIIa	IIb
ACDJE	1		
GCDDF	1	IVa	
ACDDK	1	Va	
ABIDDF	2	VIa	IIIb
ABCDDKF	1		
GBCDDEF	3	VIIa	
GBCDDE	1		
ABCDDEL	1		

STRUKTURÁLNÍ KLASIFIKACE AKUSTICKY EVOKOVANÝCH KMENOVÝCH POTENCIÁLŮ

- ☑ etalonové řetězce pro jednotlivé shluky určené iteračními algoritmy jsou:

Ia	ABCJEF	**Va**	ACDDK	**Ib**	ABCJEF
IIa	ACJEF	**VIa**	ABIDDF	**IIb**	ACDDE
IIIa	AIDDE	**VIIa**	GBCDDEF	**IIIb**	ABCDDE
IVa	GCDDF				

Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE"

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ