Kybernetika

Jiří Kraus

Kódování a komprese psané češtiny

Kybernetika, Vol. 1 (1965), No. 1, (74)--84

Persistent URL: http://dml.cz/dmlcz/124841

KYBERNETIKA ČÍSLO 1, ROČNÍK 1/1965

Kódování a komprese psané češtiny

JIŘÍ KRAUS

Autor zkoumá (1) kódování znaků psané češtiny z hlediska optimálnosti, tj. takového přiřazení, kdy nejčetnějším grafémům odpovídá nejjednodušší znak přenosového systému a naopak, a (2) kompresi — zhuštění přenášeného textu zakódovaného různými abecedami.

Tato práce se zabývá účelným přenosem informace v psaném jazyce prostřednictvím různých kanálů využívaných ve sdělovací technice, při zápisu mluvené řeči atd. Předpokladem zpracování tématu jsou numerické výsledky grafematické statistiky, která byla provedena v oddělení matematické a aplikované lingvistiky Ústavu pro jazyk český [1] a která zároveň může být vhodným korektivem celé řady dalších studií, týkajících se ekonomie přenosu. Budu zde uvažovat o umělých transformacích z jedné grafické soustavy do druhé, jejichž cílem je prostorové zhuštění zprávy, a tedy i hospodárnější a rychlejší přenos informace.* Rozebíranými systémy jsou česká stenografická soustava Herouta-Mikulíka, Morseova abeceda a číselné kódy, jejichž zvláštním případem je dálnopis.

Výchozí operací každého přenosu je přiřazení znaku prvkům přenášeného textu. Toto přiřazení budeme nazývat překódováním. Charakter kódu je dán druhem sdělovacího zařízení, např. písmena abecedy se přenášejí pomocí soustavy čísel, nebo účelem použití – těsnopis umožňuje zápis mluvené řeči přiměřenou rychlostí. Hlavním rysem překódování je úplnost a jednoznačnost; každému prvku je totiž přiřazen jiný znak nebo kombinace znaků, takže při dekódování získáme výchozí posloupnost prvků původního textu. Překódovaný text je možno ještě komprimovat – zhušťovat – a tak snížit jeho nadbytečnost. Komprimovat lze nejrozmanitější soubory diskrétních prvků, které mají určitý stupeň uspořádanosti.

Základním požadavkem na překódování je optimálnost, to je takové přiřazení, při

* Ponechávám tedy stranou kondenzaci jazyka, jak o ní mluví např. V. Mathesius (viz [2] str. 171 n.) a J. Nosek (přednáška „Větné zhušťování v moderní angličtině"), protože ji považuji za objektivní jazykový jev.

kterém nejčetnějšímu písmenu odpovídá nejkratší znak a při uspořádání podle klesající četnosti písmen se složitost (délka) znaku zvyšuje.

Druhá úvodní poznámka se týká povahy přenášeného textu. Považuji v této souvislosti za vhodné pracovat s pojmy prvního a druhého členění, jak je zavádí M. Martinet [3] a jak je přejímají další lingvisté a komunikační inženýři [4]. Znaky prvního členění jsou ty, které mají vztah k sémantickému plánu jazyka – patří sem morfémy, slova, věty a syntagmata, příp. promluvy. Grafémy, fonémy a slabiky jsou pak prvky druhého členění. Rozlišení těchto dvou rovin je významné zvláště pro otázky účelného přenosu informací. Na úrovni prvního členění se může komprese opírat o sémantickou doplnitelnost, předpokládá se totiž – je-li dékódovačem někdo, kdo daný jazyk dobře zná – schopnost expandování zhuštěného textu na základě poměrně malého množství znaků. Naopak u druhého členění mimotextová opora neexistuje a zdrojem úspory je jedině snížení matematicky vyjádřené nadbytečnosti sdělení.

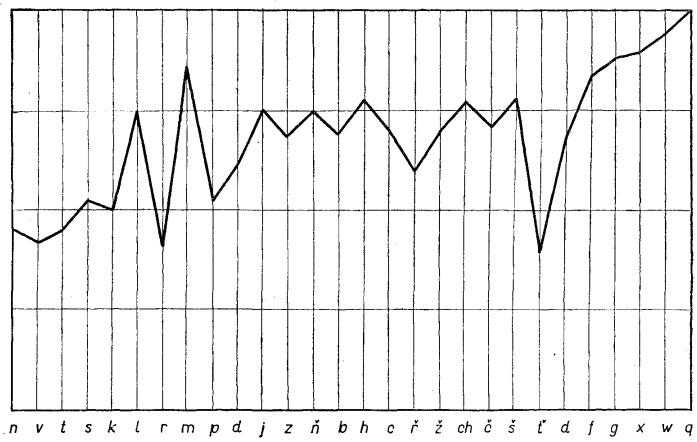
I. KÓDOVÁNÍ PRVKŮ DRUHÉHO ČLENĚNÍ

Nejdříve si povšimneme grafické soustavy, jejíž rozbor má tradici spojenou s rozvojem fonologie v Pražské škole. Je to těsnopis, který teoreticky prozkoumal Trnka v práci „Pokus o vědeckou metodu a praktickou reformu těsnopisu". Trnka zde vypracovává grafický analogon k fonologickému učení Pražské školy. Jeho soustava se opírá o systém fonologických protikladů. Bezpříznakovým členům jsou v ní přiděleny základní grafické znaky, příznakové členy opozic mají odpovídající grafémy s tvarovými modifikacemi:

část elipsy: přímka ((*p* : *b* retoretný neznělý: retoretný znělý
část elipsy) *f* zuboretný neznělý

(Příklad je zvolen z těsnopisné soustavy H – M, která někdy Trnkovu požadavku vyhovuje.)

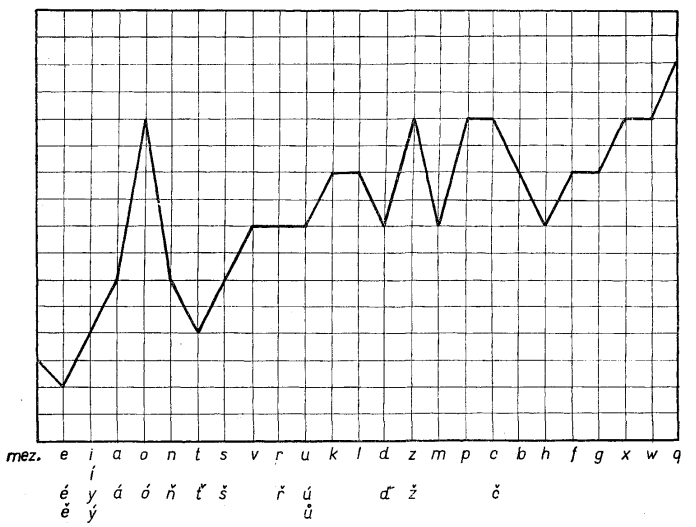
Grafémy tak vytvářejí jakýsi model souběžných protikladů (charakterizovaných různým způsobem tvoření) a souměrných protikladů (odlišených podle místa tvoření), takže zobrazují všechny vztahy hláskové soustavy. Srovnáním s jinými druhy písma a s jinými kódy není obtížné zjistit, že izomorfismus mezi zvukovou a grafickou stránkou je nejen těžko realizovatelný, ale i nadbytečný, i když teoretický význam Trnkova požadavku je lingvisticky poučný. Místo analogie se soustavou fonologických protikladů je nutno vyzdvihnout další kritéria, v jejichž vytýčení je – jak se domnívám – největší význam Trnkovy práce. Prvním z nich je vztah mezi obtížnostními hodnotami grafických elementů a relativními četnostmi grafémů, který odpovídá požadavku optimálnosti kódování, vymezenému na počátku této práce. Kvantitativní hodnoty grafické obtížnosti nebyly dosud pro český těsnopis zjištěny. Opírám se tedy o vlastní měření, pomůckou mi byla práce N. N. Sokolova „Teoretičeskije



Obr. 1. Grafické znázornění hodnot obtížnosti znaků při sestupném uspořádání četností souhlásek v těsnopisné soustavě H — M. Na svislé ose je číselně vyjádřena obtížnost zápisu těsnopisných znaků tvořených kombinací grafických elementů (obloučků, smyček, přímek). Grafické uspořádání je zde účelnější než výpočet korelačního koeficientu, protože ukazuje, které znaky se vymykají tendenci, od níž očekáváme, že při optimálním kódu bude mírně vzestupná.

osnovy gosudarstvennoj jedinoj sistemy stenografii", kde se autor podobným pokusem zabývá. Základní jednotkou měření je grafický element (oblouček, přímka, smyčka), k němuž přidávám poměrnou hodnotu spojnice, odstupňovanou, jde-li o spojnici v ostrém úhlu nebo v obloučku. Měření se opíralo jednak o experimentální údaje, jednak o výpočty skladnosti. Graf na obr. 1 ukazuje, že nejméně výhodný znak je přidělen písmenu *m*. Zdánlivě nevýhodné je i označení písmene *l*, ale o něm bude zmínka později. Ostatní znaky se významně neodchylují od intuitivně očekávané tendence, kód je tedy účelný, i když zatím nepočítáme s tím, že některá písmena mají jinou četnost na prvním místě ve slově, uprostřed a na konci slova; pro tento výpočet nejsou dosud soustavné údaje, které by patrně ovlivnily charakter grafu pouze u některých písmen.

Dalším rozebíraným kódem je Morseova abeceda, původně koncipovaná pro angličtinu. Proto zde často dochází k rozporu mezi jednotností mezinárodní normy a zvláštnostmi národních abeced (graf na obr. 2). Tento protiklad byl vyřešen v současné době pro některé indické jazyky vytvořením nové telegrafní abecedy [5]. Jsem přesvědčen, že i pro češtinu může mít zkoumání Morseovy abecedy svůj význam, i když mezinárodní závaznost a vyškolené kádry nedovolují zatím představu podstatnější reformy. Přiloženého grafu, který ukazuje, nakolik výhodná jsou označení



Obr. 2. Grafické znázornění obtížnostních hodnot při sestupném uspořádání četností grafémů v Morseově abecedě. Na svislé ose jsou prvky znaků Morseovy abecedy klasifikovány takto: tečka = 2, čárka = 4, mezera mezi znaky = 3, mezera mezi slovy = 6.

jednotlivých písmen, je ovšem možné využít pro předběžné optimální zakódování textu, takže se ukazuje, že při šifrování je výhodné se vyhýbat písmenům *l*, *c*, *f*, *q* nebo naopak častěji využívat *e*, *i*, *a*, *n*. Křivka grafu ukazuje stupeň optimálnosti Morseovy abecedy, to je souvislost mezi relativní četností písmen a časovou náročností znaků. Písmenové hodnoty znaků s háčky a čárkami jsou pro češtinu sumarizovány. Elementy klasifikuji ve shodě s R. Moreauem [6].

Srovnáme-li graf těsnopisného a Morseova kódu, ukáže se velmi zřetelně rozdíl mezi „českým" těsnopisem a mezinárodním charakterem Morseovy abecedy. Nemůžeme jí sice upřít očekávanou tendenci, ale výkyvy Morseových znaků jsou nadměrně vysoké.

Další možnost komprese poskytují abecedy klávesnicové, užívané na psacích strojích, dálnopisech a na samočinných počítačích. Do tohoto tématu lze zařadit několik úkolů:

a) které znaky do klaviatur zařadit;
b) jak je na klaviatuře umístit;
c) jak text optimálně zakódovat.

Ad a) První otázka je formulována neustále znovu s novou naléhavostí. Největší potíže představuje především požadavek jednotné normy mezinárodních klaviatur, které plně nevyhovují všem abecedám, ale rezervují pro ně pouze několik kláves. Zvláště omezené možnosti má dosavadní klaviatura dálnopisu o 32 prvcích. Proto je v současné době vypracováváno nové řešení, využívající šestimístných kódů o 64 prvcích (2^6). Statistickým šetřením a pozorováním zákonitostí vzniku chyb byla sestavena tabulka šestimístných kódů uspořádaná podle spolehlivosti přenosu [7]. Nejspolehlivějším šesticím jsou přiřazeny příkazní prvky (písmenová změna, číslicová změna, posun válce, návrat válce), další klávesy je nutno rezervovat pro interpunkci a zvláštní znaky, takže všechny zvláštnosti národních abeced, kterých je v češtině mnoho (15), se na klaviaturu nevejdou. Tato skutečnost nemusí vždy znamenat defektnost soustavy z několika důvodů:

1. význam každého slova je specifikován kontextem (u prvků prvního členění),
2. při kódování se těmto písmenům můžeme vyhnout vůbec,
3. v nutných případech je možné nahradit písmeno, pro něž nemáme označení, předem dohodnutou nepřípustnou, tj. v daném jazyce se nevyskytující kombinací, spřežkou atd.

Ad b) Umístění abecedních znaků na klávesnicích je otázkou spíše technickou. Chtěl bych jen stručně upozornit na neúčelnost klaviatur psacích strojů z hlediska psaní desetiprstovou metodou. Např. málo frekventované *f* je umístěno na optimálním místě, kdežto četné *a* je vysunuto na levém okraji a *e* je dokonce mimo základní řadu. Tento problém řešili na základě relativních četností grafémů a digrafů polští autoři při sestavování normalizované klaviatury psacích strojů [8].

Ad c) Otázkou optimálního dvojkového kódu pro uspořádanou množinu signálů se zabývali zvláště Shannon a Fano [9], [10], [11]. V jejich soustavě jsou písmena seřazena sestupně podle relativních četností a rozdělena na dvě skupiny o přibližně stejné pravděpodobnosti. První skupině znaků se přidělí znak 1, druhé 0. Každý nový sloupec opět rozdělíme a znovu označíme 1, 0. To provádíme tak dlouho, dokud každé písmeno není tímto binárním kódem jednoznačně určeno (viz tabulku). Protože nejčetnější grafém je vždy v horní skupině, dostane se mu nejkratšího znaku a nejméně četnému nejdelšího znaku, což je opět ve shodě s požadavkem optimálnosti.

Maximální počet 1 a 0, který připadne na jedno písmeno tohoto nestejnoměrného kódu, rovná se exponentu v mocnině dvou, který je nejblíže vyšší než počet uvažovaných písmen. Pro české grafémy, jichž je i s mezerou 42, rovná se maximální množství znaků 6 ($2^5 < 42 < 2^6$). Průměrné množství znaků podle Shannona a Fano je ovšem menší, pro češtinu podle tabulky 5,013. To odpovídá hodnotě nepatrně větší, než je minimální hodnota kódu $H_1/\log 2$, to je 4,96 dit. Účinnost kódu Shannona-Fano se ještě zvýší, uvažujeme-li, že se nemusí kódovat jen jednotlivá písmena, ale n-tice písmen, jichž je k^n. Z celkového počtu k^n ovšem můžeme počítat jen ty skupiny, které se v posloupnosti (u prvků druhého členění) nebo v textu přirozeného jazyka (v prvním členění) vyskytují.

Tabulka

Konstrukce šestimístného kódu pro psanou češtinu metodou Shannona — Fano

(1) $$2^5 < 42 < 2^6$$

(2) $$H_1 = 1{,}49 \text{ dit}$$

(3) $$\text{Minimální hodnota kódu} = \frac{H_1}{\log 2} = 4{,}96$$

Průměrný počet elementárních znaků

$$N = \sum rx = 5{,}013.$$

(*r* je počet míst binárního kódu, *x* je relativní četnost)

Znak	Relativní četnost	Kód	Znak	Relativní četnost	Kód
mezera	0,165	1111	r	0,029	010100
e	0,073	11101	p	0,028	010110
é	0,010	11100	m	0,028	010101
ě	0,006	11011	d	0,026	01001
a	0,054	11010	j	0,021	01000
á	0,022	11001	z	0,019	0100
i	0,033	11000	ň	0,014	00111
í	0,024	1011	b	0,013	001101
o	0,068	10101	h	0,011	001100
ó	0,001	10100	c	0,010	00101
u	0,030	10010	ch	0,009	001001
ú	0,005	10011	ž	0,009	001000
y	0,016	10001	ř	0,009	00011
ý	0,009	10000	š	0,007	000101
n	0,040	01111	ť	0,007	000100
v	0,040	011101	ď	0,004	000011
t	0,039	011100	f	0,002	000010
s	0,037	01101	g	0,002	000001
k	0,034	01100	x	0,001	000000
l	0,033	010111	q	neklasifikováno	

Pozn.: Hodnoty relativních četností jsou zaokrouhleny na 3 desetinná místa.

II. KOMPRESE PRVKŮ DRUHÉHO ČLENĚNÍ

Dosavadní přehled se týkal kódování všech prvků přenášené posloupnosti znaků. Z hlediska přenosu a rychlého zápisu je ovšem výhodnější vymezit prvky základní a eliminovat nadbytečné, to je ty, které jsou jednoznačně určeny strukturací textu. V dalším výkladu se vrátím k těm abecedám, jejichž primární zakódování již bylo objasněno.

V první etapě měření obtížnostních hodnot těsnopisných znaků jsem ponechal stranou samohlásky, pro něž má soustava H – M zvláštní řešení. Na první pohled by se mohlo zdát, že vydělení samohlásek je ústupkem požadavku na zavedení fonologických kritérií do grafiky, proti němuž jsem vznesl některé námitky v úvodu. Specifické vyznačování samohlásek má však svůj statistický smysl – v tabulce relativních četností grafémů zaujímají samohlásky vesměs první místa. Zvláště pozoruhodná je jejich vysoká valence (kombinovatelnost s jinými hláskami), takže jako komponenty tvoří značnou část digrafových skupin. Proto soustava H – M přesouvá vyznačování vokálů na ligatury (spojnice), které v běžném písmě nejsou funkčně zatíženy. To je také jeden ze zdrojů tahové úspory; ze vztahu mezi samohláskami a souhláskami (ve 12 písmenech je průměrně 7 souhlásek a 5 samohlásek) je možné vypočítat i zhuštění textu: při stenografickém přepisu jedné strojové stránky se symbolickou vokalizací ušetří 1250 znaků, tj. asi 45%.

Teoreticky ještě zůstává nedořešena otázka vyznačování délek, aktuální i mimo těsnopis zvláště v současné době, kdy je třeba navrhovat doplňování národních zvláštností jednotlivých abeced u klaviatur s mezinárodní normou. Náš těsnopis rozlišuje kvantitu jen u koncového *u*/*ů*, v ostatních případech nikoliv. Praxe tedy odmítla požadavek rozlišování délek a předpokládá doplnitelnost pomocí kontextu. Podobná situace je např. u nevyznačování ruského a bulharského přízvuku nebo u srbocharvátských intonací, které mají zhruba stejnou významovou platnost. Lze tedy předpokládat, že komprimovaný zápis bude požadavku vzájemné jednoznačnosti mezi hláskovou a grafickou soustavou poněkud vzdálenější, ale to nemusí vylučovat spolehlivost celé takové abecedy.

Dalším využitím valenčních vlastností písmen je v těsnopisu symbolizace (vyznačení následného písmena modifikací předcházejícího znaku) převzatá z grafických možností běžných abeced (znaky se mohou v liniatuře prodlužovat oběma směry). Tyto možnosti nutí hledat taková písmena, která jsou nejčastěji druhým komponentem souhláskových digrafů. Podle dosavadních výpočtů [1], [13] se v češtině na druhém místě v konsonantní dvojici nejčastěji vyskytuje *–l* (po 18 souhláskách, nikdy ne po 9), dále *–r* (17 – 10) a konečně *–n* (16 – 11). K tomu je možné poznamenat, že i zde četnostní vztahy ukazují na jistý rys fonologického systému, protože všechny tyto hlásky jsou sonory.

Pro nejčastější souhláskové dvojice (*st*, *sk*, *št*, *sv*) má těsnopis ještě další řešení – může je vyjádřit jedním znakem. Na podobný způsob komprese upozorňuje pro psanou němčinu Meyer-Eppler [12]. Nejčastější digrafy lze nahradit málo četnými grafémy, takže rozsah textu se znatelně sníží. Např. ve větě *PRAVIDELNÉ SPOŘENÍ – KRÁSNÝ DOMOV* místo dvojic *PR VI LN PO KR OV* použijeme grafémů *Q W X F G H*; vznikne nový text, kde místo 31 bude 25 písmen, úspora tedy dosahuje 20% v poměrně krátkém úseku. Tento způsob komprese je zvláště výhodný tam, kde je malé množství přenášených sdělení stereotypního rázu a kde lze tabulku substitucí konstantně vymezit.

Vztah mezi matematickým pojmem nadbytečnosti a eliminací znaků za účelem komprese nejlépe vysvětluje Shannonova formulace kapacity kanálů [14], [15], [16]. Mějme kanál s kapacitou dvou bitů, kterým se má přenést posloupnost znaků *A*, *B*, *C*, *D* s četnostmi 4 : 2 : 1 : 1. Entropie souboru je $1\frac{3}{4}$ na písmeno. Podle Shannona existuje takový kód, který umožňuje kompresi v poměru 2 : $1\frac{3}{4}$, tj. 8 : 7. U 64-místného kódu dálnopisu při 10 000 znacích představuje možná úspora až 50%. Shannon zde počítá s nezávislými prvky přenášené zprávy. Bereme-li v úvahu *n*-tice písmen, můžeme kompresi ještě zvětšit. Při entropii H_0 (entropii souboru grafémů bez pravděpodobnostního omezení), která se rovná v psané češtině 5,39, blíží se komprese hodnotě $1 - H_n/5{,}39$. Se stoupající hodnotou indexu *n* se snižuje čitatel zlomku a komprese se přibližuje jedné. Na základě dosavadních výpočtů entropie pro psanou češtinu je možné vyjádřit kompresi H_1 (uvažují se relativní četnosti písmen) hodnotou 13,3%, H_2 (digrafové četnosti) 28,2%. Jsou to čísla značně vyšší než ta, která uvádí Moreau pro francouzštinu [17] (při H_2 je úspora 10,6%), což je vysvětlitelné tím, že počítá pouze s 30-prvkovou telegrafní abecedou. Moreauovy hodnoty pro trigrafy, tetragrafy, pentagrafy jsou 18,5%, 27,9% a 37,1%. V češtině lze opět předpokládat čísla úměrně vyšší.

III. KÓDOVÁNÍ A KOMPRESE PRVKŮ PRVNÍHO ČLENĚNÍ

Nejelementárnějším postupem textové komprese na úrovni prvního členění je vynechávání doplnitelných slov nebo jejich redukování na menší počet písmen. Psycholingvistické pokusy [18] s rekonstrukcí porušeného textu celkem přesvědčivě ukázaly, že nejlépe jsou doplňována slova nejčetnější, která obvykle bývají i kontextově nejvíce určená. Nejhůře se rekonstruují slova neobvyklá s nízkou četností. Význam frekvenčních slovníků [19], [20], [21] pro zkoumání komprese spočívá převážně ve vymezování těch slov, která lze nejsilněji komprimovat, protože je to nejsnazší a nejefektivnější. Na základě frekvenčních výzkumů jsou budovány i soubory těsnopisných samoznaků a zkratek, které je možné považovat za typy textového zhušťování. Prvním typem jsou zkratky začátkem slova, vyznačující první písmena, jež většinou nesou největší množství informace. Pro češtinu jako jazyk flektivní jsou nejvhodnější zkratky dalšího typu – smíšené – vyznačující začátek a konec slova. Oba tyto základní způsoby krácení by bylo možné aplikovat na jiné formy přenosu sdělení, např. na kód dálnopisný nebo telegrafní.

Z roviny prvního členění vychází i dosavadní praxe kódování v zahraničním obchodě. *Unicode* např. stanoví písmenové znaky pro nejčastější věty z písemností v zahraničním obchodě, které vytvářejí určitou univerzální nomenklaturu, na niž navazují kódy podnikové. *Unicode* obsahuje 17 576 kódových slov lišících se mezi sebou třemi písmeny, tzn. že jsou zabezpečena proti možnému zkreslení. Lingvisticky je zajímavá předběžná redakce vět, při které dochází ke zvláštnímu typu umělé komprese, a to komprese syntaktické. Dostáváme se tak do blízkosti standardizovaného telegrafního stylu, který omezuje významově nadbytečná slova.

Rovněž Morseova abeceda se opírá o princip smyslové doplnitelnosti, zvláště u kontaktních signálů, tj. u těch frekventovaných sdělení, která jsou obvyklá při navazování spojení. Charakteristickým prvním písmenem kódu bývá obvykle písmeno q nebo x, v běžném textu na tomto místě řídké.

Zvláštním případem textové úspory jsou zkratky. Ve sdělovací technice mají největší význam tam, kde jde o ustálené relace, např. v jazyce kontrolních věží na letištích [22].

Významnou možnost účelného zhuštění textu poskytnou výpočty pentagrafových závislostí pro psanou češtinu. Protože průměrný počet grafémů na slovo je 5,6, nebude se entropie H_5 příliš podstatně lišit od průměrného množství informace na jedno slovo, zvláště uvažujeme-li kódování nejčetnějších slov, která jsou převážně kratší. Moreau [6] uvádí pro francouzský text zapsaný telegrafní abecedou úsporu 60,2% při kódování celých slov, pro ruštinu představuje hodnota komprese 71,6% [23].

Naznačená řešení ukazují jen ve velmi hrubých rysech význam otázek kódování a komprese psané češtiny pro oblast spojů, samočinných počítačů, pro administrativu a pro celou řadu dalších oborů. Lingvistika od nich může získat moho podnětů a svým teoretickým přístupem může přispět k hospodárnosti našich spojů i k účelnějšímu přenosu informace v přirozených i umělých jazycích.

(Došlo dne 7. května 1964.)

LITERATURA

[1] L. Doležel: Předběžný odhad entropie a redundance psané češtiny. SaS 3 (1963).
[2] V. Mathesius: Obsahový rozbor současné angličtiny na základě obecně lingvistickém. Praha 1961.
[3] M. Martinet: Eléments de linguistique générale. Paris 1960.
[4] G. Hérault, R. Moreau: Généralités sur les relations entre le plan de l'expression et le plan de contenu. La traduction automatique 3 (1962), 33—43.
[5] R. S. Ramakrishna: Information theory and some its applications. Current Science 27 (1958) č. 10.
[6] R. Moreau: Linguistique et Télécommunications. L'onde életrique (1962), No 426, 731—737.
[7] B. Kubín: nepubl. kand. práce.
[8] L. Zubrzycka: O dostosowywaniu klawiatury maszyny do pisania do struktury języka. Zastosowania matematyki VI, Warszawa-Wroclaw 1961—3, s. 419—439.
[9] R. M. Fano: The transmission of information. MIT Res. Lab. Electr., Techn. Rep. 65 (1 949).
[10] P. Fey: Informationstheorie. Berlin 1963.
[11] A. M. Jaglom, I. M. Jaglom: Pravděpodobnost a informace. Praha 1964.
[12] W. Meyer-Eppler: Grundlagen und Anwendungen der Informationstheorie. Berlin 1959.
[13] W. Appel: Energiebasis-Artikulationsbasis. Wiener Slavistisches Jahrbuch 6 (1957—59), 93—97.
[14] C. E. Shannon, W. Weaver: The mathematical theory of communication. Urbana 1949.
[15] W. R. Ashby: Kybernetika. Praha 1961.
[16] V. Dupač, J. Hájek: Pravděpodobnost ve vědě a technice, Praha 1962.
[17] R. Moreau: Quelques remarques en vue d'un codage automatique des télécommunications. Automatism 11 (1962).

[18] I. Pollack: Incorrect responses to unknown messages restricted in word frequency. Language and Speech 5 (1962), 125—157.
[19] J. Jelínek, J. V. Bečka, M. Těšitelová: Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha 1961.
[20] M. Matula, J. Čáp, F. Petrásek: Frekvence slov ve stenografické praxi. SÚT Praha 1961.
[21] M. Matula: Frekvence kořenů slov. SÚT Praha 1963.
[22] Control Tower Language. JASA 24 (1952), 595—596.
[23] G. G. Bělonogov, V. I. Grigorjev, R. G. Kotov: Avtomatičeskoje leksičeskoje kodirovanije soobščenij. VJaz 4 (1960), 107—111.

SUMMARY

Coding and Compression of Written Czech

JIŘÍ KRAUS

The paper treats some questions of effective transmission on the base of the results of letter frequencies in written Czech. The autor distinguishes (1) coding, i.e. assignment of one sign to one element of a message in accordance with the character of the communication system (telegraph alphabet, Morse code, shorthand etc.) and (2) compression of a message, i.e. decrease of redundancy by omitting strictly predictionable signs. The message can be transmitted in the form of natural language (formed by the elements of the 1st articulation – morphems, words etc.) or in the form of sets of graphems – elements of the 2nd articulation. On the level of the 1st articulation the knowledge of the language by the decoder is to be considered, and therefore, the message can be reduced more substantialy.

The most important feature of coding is optimality – a relation by which the simplest signs are assigned to the most frequent letters and vice versa.

In section I (Coding of the elements of the 2nd articulation) the optimality of Czech shorthand system is investigated by the means of the correlation between letter frequencies and graphical form of the signs; the optimality of Morse code and that of the binary code by method of Shannon and Fano (see graphs).

Section II (Compression of the elements of the 2nd articulation) examines the possibilities of the decrease of redundancy by way of substitution of letter combinations by simpler signs in shorthand, telegraph alphabet and binary code.

Section III (Coding and compression of the elements of the 1st articulation) shows the importance of word-count books (dictionnaries of word frequencies) for economical transmission of information in natural languages. It is assumed that the most frequent words could be strongly reduced.

From this point of view it has been shown that the shorthand is more optimal for Czech language than Morse code which is related to another statistical type of the language. The above mentioned criteria could be useful for research of economy of various codes in written language.

Jiří Kraus, prom. filolog, Ústav pro jazyk český ČSAV, Letenská 4, Praha 1 - Malá Strana.