en cz

Statistika českých grafémů s využitím moderní výpočetní techniky

Jan Králík

(pdf)

Статистика чешских графем при помощи вычислительной техники / Statistics of the Czech graphemes with the aid of modern computational technique

Moderní výpočetní technika, bez níž je současnost i budoucnost kvantitativní lingvistiky dnes téměř nemyslitelná, umožňuje získat poměrně snadno statisticky reprezentativní data i v oblastech, kde shromáždění a utřídění většího materiálu bylo ještě donedávna značným problémem. Tuto přednost využití výpočetní techniky potvrdila i všestranná kvantitativní analýza současné spisovné češtiny, která se provádí v úseku matematické lingvistiky Ústavu pro jazyk český ČSAV za vedení M. Těšitelové. Vedle automatické analýzy textů na rovině lexikální, morfologické a syntaktické (Linguistica II—III; Těšitelová, 1983) bylo možno provést i statistickou analýzu grafematickou, založenou na dosud nejrozsáhlejším korpusu, jaký kdy byl u nás pro tento účel zkoumán — na korpusu o rozsahu více než tři milióny grafémů.

> Jako materiál (srov. Linguistica II—III) posloužil soubor 180 textů (každý po 3000 slovech) věcného stylu, tedy korpus o rozsahu 540 000 slov (slovoforem), tj. 3 139 926 grafémů. Výběr věcného stylu (VS) přitom zahrnul styl odborný, publicistický a administrativní, a tím zaručil také jazykovou a tematickou reprezentativnost tohoto souboru. Uvedený materiál byl pomocí děrných štítků převeden do externí paměti počítače Tesla 200 (na magnetické pásky) ve výpočetních střediscích ÚTZCHT a ÚFPL (FÚ) ČSAV a zde podrobně automaticky zpracován pomocí několika typových řad speciálních programů (Králík, 1982; Těšitelová, zde, s. 275).

Ke starším statistickým výzkumům českého grafematického systému (srov. zejm. Ludvíková - Königová, 1967) tak nyní můžeme připojit nová data, získaná za pomoci moderní výpočetní techniky — a proto velmi rychle a snadno — ze soudobého jazykového materiálu (srov. pro slovenštinu Mistrík, 1979).

1. Do grafematického systému češtiny se, jak známo (Šmilauer, 1972), obvykle řadí 24 základních latinských grafických znaků (písmen), z nichž 13 (více než polovina) přijímá diakritická znaménka, rozlišující jednak fonetickou kvantitu (čárkou, kroužkem), jednak kvalitu (háčkem, nahrazovaným u *d* a *t* někdy z výtvarných důvodů jednodušším apostrofem: *ď*, *ť*); písmena *c* a *h* se jako jediná dvojice spojují ve spřežku *ch*, a to vždy, stojí-li za sebou v tomto pořadí. Úplný grafematický systém češtiny tak tvoří 39 samostatných znaků (včetně spřežky *ch*); vedle toho se k nim přiřazují grafémy *q*, *w*, a *x*, které se vyskytují ve slovech cizích nebo přejatých. S nimi dosahuje počet grafémů užívaných v českých textech čísla 42. V našem výběru se dále vyskytly tyto cizí grafémy: *à*, *ä*, *å*, *ć*, *è*, *ë*, *ê*, *ï*, *ĺ*, *ł*, *ľ*, *ń*, *ô*, *ö*, *ŕ*, *ś*, *ü*, *ÿ*, *ż*. Celkem na ně připadá 119 výskytů.

Pokud jde o diakritická znaménka, využívají se v češtině takto:

Znaménko	Počet užití	Frekvence v textech VS	
	v systému písmen	absolutní četnost	%
čárka	6	252 658	8,0466

kroužek	1	21 820	0,6949
háček	8	180 784	5,7576

Diakritika se vyskytují celkem u 13 písmen v systému (tj. u 31 %), v textu má diakritikum v průměru každé sedmé písmeno (14,5 %). Z cizích diakritik se nejčastěji (přesto velmi zřídka) uplatňuje přehláska (47 případů v našem materiálu, tj. 0,0015 %).

[296]Statistika grafémů ve věcném stylu (VS)

Grafém	Frekvence	%	Grafém	Frekvence	%
a	195 282	6,2193	o	272 119	8,6664
á	70 193	2,2355	ó	982	0,0313
b	48 927	1,5582	p	107 157	3,4127
c	50 449	1,6067	q	41	0,0013
č	29 799	0,9490	r	116 083	3,6970
d	113 097	3,6019	ř	38 201	1,2166
ď	697	0,0222	s	141 798	4,5160
e	241 622	7,6952	š	25 283	0,8052
é	41 906	1,3346	t	179 817	5,7268
ě	51 662	1,6453	ť	1 338	0,0426
f	8 577	0,2732	u	98 730	3,1443
g	8 568	0,2729	ú	3 236	0,1031
h	39 914	1,2712	ů	21 816	0,6948
ch	36 766	1,1709	v	146 369	4,6616
i	136 673	4,3528	w	276	0,0088
í	102 673	3,2699	x	2 370	0,0755
j	66 549	2,1194	y	59 950	1,9093
k	117 329	3,7367	ý	33 662	1,0721
l	120 649	3,8424	z	69 038	2,1987
m	101 317	3,2267	ž	31 250	0,9952
n	205 204	6,5353			
ň	2 557	0,0814	*Σ*	3 139 926	100,0000

Tab. č. 1

Protože ani strojový vstup z děrných štítků, ani strojová tiskárna počítače nejsou vybaveny písmeny s diakritickými znaménky, bylo třeba řešit záznam a tisk (vstup a výstup) zvláštním postupem.

Při projektu záznamu (vstupu) textu do externí paměti počítače bylo třeba brát v úvahu nejen grafický problém diakritik, ale také specifiku české normy abecedního řazení (zakotvené v ČSN 01 0181), podle které se písmena *č*, *ř*, *š*, *ž* řadí za obdobná písmena bez háčků, zatímco písmena *ď*, *ě*, *ň*, *ť* stejně jako písmena s čárkami *á*, *é*, *í*, *ó*, *ú*, *ý* se řadí v zásadě tak, jako by diakritická znaménka neměla. Pouze pro případy, kdy je třeba určit pořadí

bezprostředně sousedících slov rozlišených pouze diakritickým znaménkem (typ: *krásné* - *krásně*, *užití* - *užíti*), určuje se pravidlo, podle kterého následuje písmeno s diakritickým znaménkem po písmenu bez diakritika (*pan* - *pán*, *vedro* - *vědro*) a pro znaménka se určuje pořadí čárka, háček, popř. čárka, kroužek (*kúra* - *kůra*). Spřežka *ch* má místo mezi písmeny *h* a *i*.

Aby bylo možno tyto zvláštnosti respektovat i při automatickém strojovém zpracování, bylo třeba především zachovat všechna diakritika a nově pojmout spřežku *ch*. Pro záznam na vstupní médium (děrné štítky) jsme proto jednak vypracovali zvláštní kód pro písmena s diakritiky (na běžných děrovačích ve výpočetních střediscích ČSAV jsme např. namísto *á* děrovali "-", namísto *ě* znak "/" atd.), jednak jsme využili zvláštní úpravy klávesnic na děrovačích Bull a popisovače ve výpočetní laboratoři úseku matematické lingvistiky ÚJČ ČSAV (klávesnice i popisovač zde byly již dříve vybaveny písmeny s diakritiky a samostatným znakem pro *ch*; jediným omezením tu byl pouze jeden znak pro dlouhé *u* — *ú*). Pro štítky

[297]Frekvenční seznam grafémů ve věcném stylu (VS)

Grafém	Frekvence	%	Grafém	Frekvence	%
o	272 119	8,6664	b	48 927	1,5582
e	241 622	7,6952	é	41 906	1,3346
n	205 204	6,5353	h	39 914	1,2712
a	195 282	6,2193	ř	38 201	1,2166
t	179 817	5,7268	ch	36 766	1,1709
v	146 369	4,6616	ý	33 662	1,0721
s	141 798	4,5160	ž	31 250	0,9952
i	136 673	4,3528	č	29 799	0,9490
l	120 649	3,8424	š	25 283	0,8052
k	117 329	3,7367	ů	21 816	0,6948
r	116 083	3,6970	f	8 577	0,2732
d	113 097	3,6019	g	8 568	0,2729
p	107 157	3,4127	ú	3 236	0,1031
í	102 673	3,2699	ň	2 557	0,0814
m	101 317	3,2267	x	2 370	0,0755
u	98 730	3,1443	ť	1 338	0,0426
á	70 193	2,2355	ó	982	0,0313
z	69 038	2,1987	ď	697	0,0222
j	66 549	2,1194	w	276	0,0088
y	59 950	1,9093	q	41	0,0013
ě	51 662	1,6453			
c	50 449	1,6067	*Σ*	3 139 926	100,0000

Tab. č. 2

s těmito dvěma různými způsoby záznamu pak byly připraveny dva různé překladové podprogramy, kterými byly záznamy alfabetických znaků převáděny do nové jednotné verze a v té ukládány na magnetickou pásku (tzv. grafémová verze). Vedle toho byl pro každý grafém pořízen zvláštní strojový překlad, který pomohl zaručit respektování normy českého abecedního řazení (např. grafémům *e*, *é*, *ě* řazeným podle normy tak, jako by

nebyly nijak rozlišeny, tu přísluší týž jednoznačně uspořádatelný symbol). Z tohoto důvodu se délka alfabetického záznamu každého slova zdvojnásobila. Aby byla zachována možnost kontroly tzv. přímým znakovým výpisem z magnetické pásky, ponechali jsme v překladu písmena bez diakritik beze změny. Pro vlastní třídicí program bylo proto třeba definovat zvláštní sekvenci znaků.

Při retrográdním třídění se naopak ukázala potřeba řadit grafémy důsledně mechanicky (srov. Štindlová, 1966), tj. každému grafému (i písmenům *ď*, *ě*, *ň*, *ť* a grafémům dlouhých vokálů) určit v abecedním uspořádání pevné místo na základě pravidla, podle něhož písmena s diakritiky následují za příslušnými písmeny bez diakritik, a pravidla určujícího pro diakritika pořadí čárka, háček, kroužek (viz výše). Znovu se tu osvědčilo strojové překládání spřežky *ch* jednomístným symbolem se samostatným abecedním zařazením mezi *h* a *i* (dvoumístně chápané *ch* by jinak stroj automaticky řadil mezi *cg* a *ci*, tedy před *č*, *d*, *e*, *f*, *g*, *h*). Retrográdní třídění si tak vyžádalo jiný systém jednoznačně uspořádatelných překladových sym-

[298]Frekvence grafémů v odborné češtině

Grafém	Frekvence	%	Grafém	Frekvence	%
o	155 142	8,7828	b	27 057	1,5317
e	138 840	7,8599	é	24 450	1,3841
n	114 470	6,4803	h	22 461	1,2715
a	107 517	6,0866	ch	21 015	1,1897
t	106 011	6,0014	ř	20 665	1,1699
v	80 057	4,5321	ý	19 549	1,1067
s	78 717	4,4562	ž	17 356	0,9825
i	77 590	4,3924	č	15 596	0,8829
k	67 058	3,7962	š	12 624	0,7146
l	66 036	3,7384	ů	11 770	0,6663
r	65 052	3,6827	f	5 740	0,3249
d	60 804	3,4422	g	5 451	0,3086
p	59 311	3,3577	x	1 714	0,0970
m	58 926	3,3359	ú	1 575	0,0892
í	56 883	3,2202	ň	1 411	0,0799
u	56 001	3,1703	ť	773	0,0438
z	39 307	2,2252	ó	591	0,0335
j	39 211	2,2198	ď	406	0,0230
á	38 066	2,1550	w	160	0,0091
y	35 076	1,9857	q	37	0,0021
ě	28 252	1,5994			
c	27 709	1,5686	*Σ*	1 766 437	100,0000

Tab. č. 3

bolů, a tedy jiný překlad a jinou přípravu strojového třídění. V tomto případě byl překlad alfabetických dat do jednoznačně uspořádatelných symbolů izomorfním zobrazením mezi dvěma množinami znaků, a nebylo proto třeba zachovávat původní grafémovou verzi. Záznam mohl být kratší

a třídění jednorázové, bez definování zvláštní sekvence znaků (využívalo se interně definovaného pořadí). Pouze pro přípravu tisku (výstupu) bylo třeba vypracovat podprogram automatické inverze použitého překladu.

Problém tisku (výstupu) písmen s diakritickými znaménky na řádkové tiskárně počítače jsme řešili užitím dvouřádkového systému, tedy kompozicí písmen s diakritiky složením z šikmé čáry v horním řádku a písmene v dolním řádku, resp. ze znaku „V“ (na místě háčku) a písmene pod ním. Zmíněné technické omezení záznamu (a tím i tisku) dlouhého *u* (*ú*) nečinilo při čtení slov i celých textů žádné obtíže. (S tím související statistický problém kvantitativního poměru *ú*/*ů* jsme řešili zvláštní statistickou sondou.)

2. Základem vlastního statistického šetření byl speciální (stavebně však jednoduchý) sčítací program, kterým se automaticky analyzovaly a statisticky postupně zpracovávaly všechny zkoumané texty. Pro zjištění frekvencí grafémů v koncové pozici (viz dále) bylo použito pracovního retrográdního obrácení grafických slov (tvarů) a grafické podoby slovníku (lexémů).

[299]Frekvence grafémů v české publicistice

Grafém	Frekvence	%	Grafém	Frekvence	%
o	85 547	8,4419	b	16 368	1,6152
e	76 508	7,5499	é	13 224	1,3050
a	65 248	6,4388	h	12 885	1,2715
n	65 176	6,4317	ř	12 788	1,2619
t	55 899	5,5162	ch	11 890	1,1733
s	48 131	4,7496	ž	10 952	1,0808
v	47 609	4,6981	ý	10 425	1,0288
i	44 859	4,4268	č	9 886	0,9756
l	41 551	4,1003	š	9 765	0,9636
k	37 933	3,7433	ů	6 984	0,6892
d	37 499	3,7005	g	2 201	0,2172
r	36 745	3,6261	f	2 042	0,2015
p	33 254	3,2816	ú	1 162	0,1147
í	32 677	3,2246	ň	725	0,0715
m	32 457	3,2029	ť	426	0,0420
u	30 882	3,0475	x	406	0,0401
á	23 057	2,2753	ó	334	0,0329
j	21 363	2,1081	ď	234	0,0231
z	21 273	2,0992	w	103	0,0102
y	18 564	1,8319	q	3	0,0003
ě	17 873	1,7637			
c	16 453	1,6236	*Σ*	1 013 361	100,0000

Tab. č. 4

Souhrnný výsledek každé grafematické statistiky — po sečtení všech grafémů a všech slov — poskytuje zároveň data o průměrné délce grafického slova, která se měří v počtu grafémů:

Počet slov	Věcný styl		**Texty**	
	celkem	odborné	public.	admin.
	540 000	300 000	180 000	60 000
grafických slov	568 634	318 211	188 117	62 406
grafických znaků	3 140 590	1 766 854	1 013 437	360 299
grafémů	3 139 926	1 766 437	1 013 361	360 128
Průměrná délka grafického slova	5,5219	5,5512	5,3869	5,7707

Poznámka: Grafické slovo je zde důsledně písmeno nebo skupina písmen mezi dvěma mezerami; grafémem rozumíme písmeno; grafickým znakem rozumíme buď písmeno, nebo samostatný grafický znak uvnitř slova (takové znaky se vyskytly pouze dva: spojovník — s frekvencí 660 (0,0210 %) a apostrof ’ — s frekvencí 4 (0,0001 %)).

[300]Frekvence grafémů v textech administrativních

Grafém	Frekvence	%	Grafém	Frekvence	%
o	31 430	8,7275	b	5 502	1,5278
e	26 274	7,2958	ř	4 748	1,3184
n	25 558	7,0969	h	4 568	1,2684
a	22 517	6,2525	č	4 317	1,1987
v	18 703	5,1934	é	4 232	1,1751
t	17 907	4,9724	ch	3 861	1,0721
s	14 950	4,1513	ý	3 688	1,0241
d	14 794	4,1080	ů	3 062	0,8503
p	14 592	4,0519	ž	2 942	0,8169
r	14 286	3,9669	š	2 894	0,8036
i	14 224	3,9497	g	916	0,2544
í	13 113	3,6412	f	795	0,2208
l	13 062	3,6270	ú	499	0,1386
k	12 338	3,4260	ň	421	0,1169
u	11 847	3,2897	x	250	0,0694
m	9 934	2,7585	ť	139	0,0386
á	9 070	2,5185	ď	57	0,0158
z	8 458	2,3486	ó	57	0,0158

y	6 310	1,7522	w	13	0,0036
c	6 287	1,7458	q	1	0,0003
j	5 975	1,6591			
ě	5 537	1,5375	*Σ*	360 128	100,0000

Tab. č. 5

V rámci odborných textů kolísá délka slova měřená v grafémech od nízké hodnoty 5,2626 (u textů mluvených) k maximu 5,8593 (u textů psaných); průměrně se udržuje nad hranicí 5,5. Podobně výrazný pokles délky slova lze pozorovat u publicistických textů (u mluvených pod hranicí 5,3, u psaných kolem 5,5). Pro administrativní texty (i mluvené) jsou typická delší slova (asi 5,8).

3. Některé základní výsledky statistiky grafémů v češtině uvádíme v tabulkách č. 1—5. V tab. č. 6 (na s. 301) pro větší názornost konfrontujeme naše data s dalším materiálem. Ukazují se tu zajímavé skutečnosti.

Frekvence grafémů (tab. č. 1—2) např. potvrzuje silnou stabilitu pořadí nejfrekventovanějších grafémů i numerických hodnot jejich relativních četností. U většiny grafémů jsou výkyvy ve frekvenci uvnitř korpusu statisticky nevýznamné, v hodnotách frekvence grafémů obvykle nepřekročí 0,6 %. I při této stabilitě však lze dobře sledovat, jak se z hlediska statistiky grafémů výrazně odlišují texty administrativního stylu od publicistiky a odborné češtiny (tab. č. 3—5). Např. neosobní úřední (administrativní) vyjadřování, které v češtině neužívá mj. slovesných tvarů 1. os. sg. a pl. pomocného slovesa *být* (tj. tvarů *jsem*, *jsme*), snižuje frekvenci grafémů *e*, *j*, *m*, *s* (tab. č. 5). Naopak výrazně vyšší relativní frekvenci má v administrativních textech grafém *č*, vyskytující se ve slovech *český*, *československý*, *Československo* a ve zkratkách *čs.*, *ČSR*, *ČSSR* apod. (ostatní grafémy v těchto slovech a zkratkách jsou

[301]Frekvenční seznam počátečních písmen ve slovníku
(% slov začínajících daným grafémem)

SSČ		VS		SSČ		VS	
p	14,5973	p	15,3422	c	0,9250	č	1,0877
z	13,8160	s	8,9895	f	0,9240	c	1,0397
s	9,1152	v	8,0136	ú	0,8685	g	0,9357
v	8,4380	z	7,1979	ch	0,8464	ch	0,8584
n	5,6097	n	6,4301	i	0,7422	ú	0,7332
k	5,5733	k	6,1102	e	0,6784	ž	0,6105
o	5,1019	o	5,7210	ř	0,5208	ř	0,3919
r	4,2711	d	4,9079	g	0,4297	e	0,3572
d	3,8023	m	4,3614	w	0,0260	w	0,1759
m	3,6461	r	4,0655	x	0,0260	y	0,0400
t	3,4507	t	3,7189	á	0,0156	x	0,0240
b	2,8648	b	3,4017	ť	0,0136	q	0,0213
u	2,2697	h	2,9831	ď	0,0130	á	0,0160
h	2,6108	u	2,4100	q	0,0130	í	0,0133
l	1,8491	l	2,3220	y	0,0130	é	0,0079

š	1,7189	a	1,9061	é	0,0065	ť	0,0079
j	1,5365	j	1,6369	í	0,0026	ď	0,0026
a	1,4324	f	1,4796	ň	0,0026		
ž	1,1719	i	1,4502	ó	0,0026		
č	1,0548	š	1,2290				

Tab. č. 6

přitom v zásadě frekventovanější, přírůstek výskytu je u nich proto relativně méně patrný než u jinak řídkého *č*).

Podobnou distinktivní vlastnost má ve zkoumaných textech dlouhé *u* (*ú*, *ů*). Statistické sondy ukázaly několikanásobnou (v průměru sedmināsobnou převahu *ů* nad *ú* (v rozpětí od 6,0103krát v publicistice do 7,4740krát v odborné češtině). V odborných textech je tento poměr značně ovlivněn relativně vyšší frekvencí *ů* v genitivu pl. maskulin, v publicistice naopak zřejmě vyšší frekvencí *ú* ve slovech jako *úkol*, *úspěch*, *účast*, *úsek*, *ústřední* (i ve zkratce *ÚV*). (K příčinám vyšší frekvence některých grafémů viz Králík, v tisku.)

Úprava sčítacího programu omezením pouze na první pozici grafému ve slově umožnila získat přehled o frekvenci počátečních písmen ve slovech. Aby mohla i tato statistika sloužit praktickým účelům, pořídili jsme ji z materiálu slovníku celého korpusu VS (tab. č. 6) a její výsledky konfrontujeme se statistickým šetřením na materiálu jednosvazkového Slovníku spisovné češtiny, 1978 (SSČ). Také u počátečních písmen je frekvence pořadí poměrně stabilní (srov. Těšitelová, 1965).

Po strojovém retrográdním přeskupení (obrácení) magnetického záznamu hesel slovníku (lexémů) bylo možno užitím téhož sčítacího programu získat kvantitativní údaje o rozložení frekvencí grafémů v koncové pozici ve slovníku, tedy obvyklou součtovou statistiku uváděnou v retrográdních slovnících (frekvence posledního písmene v lexémech).

O významu studia frekvence grafémů se zřetelem k jejich pozici ve slově svědčí tab. č. 7. Invariantní vzhledem k pozici ve slově není žádný grafém. Určitý náznak vy-

[302]Statistika grafémů v závislosti na jejich pozici ve slově
(na materiálu VS)

Grafém	Na počátku slova	V textu bez ohledu na pozici	Na konci slova	
			ve slovníku	v repertoáru tvarů
a	1,9061	6,2193	9,1360	5,2464
á	0,0160	2,2355	1,1010	2,5167
b	3,4017	1,5582	0,1893	0,1274
c	1,0397	1,6067	0,8744	0,3384
č	1,0877	0,9490	0,5518	0,0969
d	4,9079	3,6019	1,0397	0,4619
ď	0,0026	0,0222	0,0373	0,0247
e	0,3572	7,6952	5,0705	5,9953
é	0,0079	1,3346	0,5572	6,7309
ě	—	1,6453	3,1404	2,2506
f	1,4796	0,2732	0,1546	0,0304

g	0,9357	0,2729	0,2559	0,0494
h	2,9831	1,2712	0,2879	0,1388
ch	0,8584	1,1709	0,3252	6,1778
i	1,4502	4,3528	16,9763	6,8050
í	0,0133	3,2699	10,8940	14,2944
j	1,6369	2,1194	0,3279	0,2034
k	6,1102	3,7367	3,6149	1,2831
l	2,3220	3,8424	1,4742	2,4958
m	4,3614	3,2267	1,2876	7,1871
n	6,4301	6,5353	2,1727	1,2869
ň	—	0,0814	0,1839	0,0874
o	5,7210	8,6664	2,3540	5,2007
ó	—	0,0313	0,0080	0,0057
p	15,3422	3,4127	0,4052	0,1464
q	0,0213	0,0013	0,0080	0,0019
r	4,0655	3,6970	1,7115	0,5817
ř	0,3919	1,2166	0,4772	0,0836
s	8,9895	4,5160	1,9034	0,5037
š	1,2290	0,8052	0,2399	0,0969
t	3,7189	5,7268	4,5347	4,0754
ť	0,0079	0,0426	0,1413	0,0931
u	2,4100	3,1443	0,1999	8,9454
ú	0,7332	0,1031	0,0213	—
ů	—	0,6948	0,0506	2,0606
v	8,0136	4,6616	1,8421	0,4619
w	0,1759	0,0088	0,0293	0,0076
x	0,0240	0,0755	0,1173	0,0304
y	0,0400	1,9093	1,3863	11,1143
ý	—	1,0721	19,1570	2,1232
[303]z	7,1979	2,1987	0,3572	0,1730
ž	0,6105	0,9952	0,4025	0,4657
Σ	100,0000 %	100,0000 %	100,0000 %	100,0000 %
počet různých grafémů	37	42	42	41

Tab. č. 7

rovnání frekvencí v různých pozicích (ovšem se zásadně odlišnými motivacemi) by bylo možno nalézt u grafémů *t* a *l*, méně u grafémů *o* a *k*.U všech ostatních grafémů se frekvence v různých pozicích navzájem mnohonásobně liší (u grafému *í* až o tři dekadické řády) (srov. Konečná - Hronek, 1962).

Z hlediska morfologického a slovotvorného, ale i z hlediska morfematické analýzy je užitečné sledovat frekvenci grafémů na konci slov nejen pouze vzhledem k pozici posledního grafému, ale i u posledních dvou grafémů (digramů), ev. trigramů atd.; pro nedostatek místa zde uvádíme jen část seznamu nejfrekventovanějších digramů

Nejfrekventovanější koncové digramy
(v repertoáru tvarů slov — VS)

ní	3841	ku	542	ký	259	tů	130
ky	2877	je	534	ech	254	dí	127
ou	1886	tí	515	ry	249	te	119
ých	1623	ém	498	nu	248	ám	119
ho	1558	ci	495	ům	246	án	119
né	1378	mu	489	ků	233	ne	118
ích	1058	ný	486	vy	231	et	117
ím	1017	me	480	ru	225	ře	116
em	973	lo	477	le	212	ji	116
mi	939	it	448	va	202	ni	114
ké	917	ná	434	el	201	in	113
ce	891	al	426	vý	199	lů	112
la	844	st	419	lé	188	rů	112
li	829	ka	395	vě	188	ží	110
ny	769	ty	358	ách	184	hu	104
ým	729	ek	352	ta	181	čí	101
cí	708	il	341	to	178	or	99
jí	699	vá	326	ie	176	my	99
at	662	ví	290	té	171	su	97
vé	658	tu	285	tě	163	ze	94
ně	649	no	281	lu	156	ii	94
ti	599	ká	270	ra	145	ík	94
ší	593	du	265	by	136	zí	92
ly	572	en	259	lí	133	át	92
na	551	dy	259	vu	133	da	91

Tab. č. 8

[304]z materiálu všech tvarů slov v korpusu VS (tab. č. 8) (srov. Korvasová - Palek, 1962; Ludvíková - Königová, 1967).

Předkládaná data o frekvenci grafémů v češtině jsou výběrem z výsledků širší analýzy, opírající se o výpisy a další tabulky, které jsou uloženy v úseku matematické lingvistiky ÚJČ ČSAV.

LITERATURA

KONEČNÁ, D. - HRONEK, J.: Morfologická analýza podle posledního písmena. Slavica Pragensia, 4. AUC. Praha 1962, s. 259—266.
KÖNIGOVÁ, M.: K otázce statistického výběru v lingvistice. SaS, *26*, 1965, s. 161—168.
KORVASOVÁ, K. - PALEK, B.: Některé kvantitativní charakteristiky kombinací písmen v českém slovníku. Slavica Pragensia, 4. AUC. Praha 1962, s. 89—95.
KRÁLÍK, J.: Technika zpracování hromadných dat. In: Linguistica II, s. 72—80.
KRÁLÍK, J.: Kvantitativní charakteristiky českého systému grafematického. In: Kvantitativní charakteristiky současné češtiny. M. Těšitelová a kol. (V tisku.)
KRÁLÍK, J.: Kvantitativní charakteristiky grafémů v psaných a mluvených odborných projevech. In: Psaná a mluvená odborná čeština z kvantitativního hlediska. Ed. M. Těšitelová. Linguistica IV. ÚJČ ČSAV, Praha 1983, s. 121—127.
KRAUS, J.: K některým otázkám pravopisu z hlediska grafematické soustavy. SaS, *26*, 1965, s. 51—54.
KVANTITATIVNÍ CHARAKTERISTIKY SOUČASNÉ ČESKÉ PUBLICISTIKY. Linguistica II—III. Ed. M. Těšitelová. Praha 1982.
LUDVÍKOVÁ, M. - KÖNIGOVÁ, M.: Quantitative research of graphemes and phonemes in Czech. PBML, *7*, 1967, s. 15—29.
LUDVÍKOVÁ, M. - KRAUS, J.: Kvantitativní vlastnosti soustavy českých fonémů. SaS, *27*, 1966, s. 334—344.
MISTRÍK, J.: Frekvencija grafém v slovenčine. SlR, *44*, 1979, s. 193—204.
ŠMILAUER, V.: Nauka o českém jazyku. Praha 1972, s. 241—245.
ŠTINDLOVÁ, J.: Podruhé o retrográdních slovnících. SaS, *27*, 1966, s. 370—374.
TĚŠITELOVÁ, M.: O entropii počátečních písmen v češtině. Informační bulletin pro otázky jazykovědné. Kvantitativní lingvistika, *6*, 1965, s. 31—37.
TĚŠITELOVÁ, M.: Some quantitative characteristics of non-fiction texts in present-day Czech. PSML, *8*, 1983 (v tisku).

R É S U M É

Statistics of the Czech graphemes with the aid of modern computational technique

The author describes the procedure of an automatic processing of a set of texts performed with the purpose to obtain frequencies of the Czech graphemes. The analysis is based on the material of non-fiction style (VS) including as components newspaper texts, texts of administration and texts of science and technology. The statistical results are presented in tables 1—8.

Slovo a slovesnost, ročník 44 (1983), číslo 4, s. 295-304

Předchozí Ludmila Uhlířová: Aktuální členění a styl jazykových projevů (na materiále z publicistických textů)
Následující Petr Sgall: Teoretická lingvistika ve věku počítačů