Jiří Š e b e k (Praha)

SESKUPENÍ PŘECHÝLENÝCH A PODOBNÝCH PŘÍJMENÍ UŽÍVANÝCH V ČESKÉ REPUBLICE POMOCÍ POČÍTAČOVÉHO ZPRACOVÁNÍ DAT Z REGISTRU OBYVATEL

GROUPING OF FEMALE FORMS OF SURNAMES AND SIMILAR SURNAMES IN THE CZECH REPUBLIC USING COMPUTER DATA PROCESSING FROM POPULATION REGISTER

Most strings of female surnames registered in the Czech Republic are lexically different from related male surnames. This article provides a method of grouping surnames by similarity and computing surname frequencies for these grouped surnames. The method reduces the 251,723 registered surname variants to 142,586 groups. Grouped surname frequencies can be used for linguistic research of similar surnames, determining geographic distribution of surnames, or by researchers which require surname frequencies irrespective of gender.

Keywords
surnames, frequency, Czech Republic

Při zkoumání příjmení v České republice se při strojovém zpracování naráží na nemožnost slučovat četnosti mužských a příslušných přechýlených ženských podob příjmení. Dosud sloužila k identifikaci ženské formy příjmení při takovém zpracování zejména přítomnost grafému *á* na konci jména. Taková jména byla pak vylučována z analýz četnosti příjmení,[1)] neboť nemohla být automaticky sloučena s příslušnými mužskými formami, ani se nedalo strojově rozhodnout, zda-li se jedná pouze o ženská příjmení či také o mužská příjmení. Navíc je zřejmé, že ne všechna ženská příjmení končí grafémem „*á*“. Tento článek má za cíl dokumentovat metodu pro vytvoření převodníku pro seskupování přechýlených a podobných příjmení ze seznamu všech příjmení empiricky pozorovaných v České republice. Takový převodník umožňuje návazné analýzy například geografického rozdělení příjmení či jiných zkoumání jejich četnosti. Součástí článku jsou protokoly v prostředí R[2)] a konfigurační soubory pro

[1)] J. Novotný – J. A. Cheshire (2012), The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-Occurrence of Surnames, PloS ONE 7 (10) (leden): e48568. doi:10.1371/journal.pone.0048568. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3485322&tool=pmcentrez&rendertype=abstract>. Viz s. e48568:3.

[2)] R CORE TEAM (2013) [software], R: A Language and Environment for Statistical Computing, Vienna, R Foundation for Statistical Computing. <http://www.r-project.org/>.

počítačové prostředí MetaCentrum,[3] což umožní ověřit závěry, vytvořit převodník, a dále měnit navržený algoritmus podle potřeby výzkumníka.

V České republice bylo k 3. srpnu 2013 Ministerstvem vnitra zaznamenáno v evidenci obyvatel 270 172 různých příjmení v ženském přechýlení nebo mužské podobě.[4] Tato databáze obsahovala příjmení 10 244 357 registrovaných osob, neboť součástí seznamu příjmení byla také informace o četnosti výskytu příjmení v registru. Po importování dat (Protokol R, část 1) je vhodné před dalším zpracováním provést korekce zjevně chybných zápisů (Protokol R, část 2). Jedná se například o dvě mezery vedle sebe nebo další překlepy. Ty nejsou vždy jednoznačné,[5] přesto je vhodné ze souboru přinejmenším odstranit znaky, které se obvykle ve jménech samostatně nevyskytují (Tab. 1). Záměny snížily počet různých příjmení v databázi z původních 270 172 na 270 131.

Tab. 1: Záměny

	z	na
1	␣␣	␣
2	O␣´	O’
3	D´␣	D’
4	´␣	’
5	´	’
6	`	’
7	D’␣	D’
8	’␣	’

Celkem 22 094 položek s příjmením obsahovalo mezery. Jednalo se buď o mezery, které jsou součástí příjmení, například „Abu Al [...]“, „De La [...]“, „Van der [...]“, nebo se jednalo o mezery, které označují rozdělení položky na několik různých příjmení, např. získaných po manželovi či manželce. Před rozdělením podle příjmení bylo nutno nahradit mezeru, jež je legitimní součástí příjmení, například podtržítkem, který se v souboru nevyskytuje. V souboru byly provedeny náhrady zachycené v tabulce č. 2 a tabulce č. 3. Po těchto transformacích zbylo v souboru 20 687 položek příjmení, které obsahují dvě až pět příjmení (či složek příjmení) rozdělených mezerami. Tato příjmení byla před analýzou rozdělena na své konstitutivní části, jde například o příjmení v tabulce č. 4.

Tab. 4: Příklad příjmení rozdělených mezerou

	příjmení
1	BAČOVÁ SOUČKOVÁ
2	LISÁ VITNEROVÁ
3	RABECH BRABCOVÁ
4	DRÁŽKOVÁ DRENGUBÁKOVÁ
5	PALACIOS MIRANDA

[3] CESNET (2013) [výpočetní infrastruktura], Národní gridová infrastruktura MetaCentrum. Praha, <http://www.metacentrum.cz/>.

[4] Ministerstvo vnitra (2013) [online], Četnost jmen a příjmení (Frequency of Names and Surnames), Praha, <http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx> [cit. 3. srpna 2013].

[5] Například není zřejmé, zdali v databázi uvedená skupina znaků L´ (písmeno L a vedle samostatně stojící čárka) má být Ľ nebo Ĺ, v databázi bývá znak ’ (apostrof) mylně zapsán znakem ´ (samostatně stojící čárkou).

Tab. 2: Záměny na jakékoli pozici v poli

	z	na
1	ABU AL␣	ABU_AL_
2	␣AL␣	␣AL_
3	ABA S	ABA_S
4	ABD EL␣	ABD_EL_
5	ABD␣	ABD_
6	ABI␣	ABI_
7	ABOU EL␣	ABOU_EL_
8	ABOU␣	ABOU_
9	ABU␣	ABU_
10	AIT EL␣	AIT_EL_
11	BEN␣	BEN_
12	DE LA␣	DE_LA_
13	␣DA␣	␣DA_
14	␣DI␣	␣DI_
15	DE LOS␣	DE_LOS_
16	␣DOS␣	␣DOS_
17	␣DEL	␣DEL_
18	VAN DE␣	VAN_DE_
19	VAN DER␣	VAN_DER_
20	VAN DEN␣	VAN_DEN_
21	␣VON␣	␣VON_
22	␣UND␣	_UND_
23	␣DE␣	␣DE_
24	␣VAN␣	␣VAN_
25	␣EL␣	␣EL_
26	MAC␣	MAC_
27	␣Y␣	␣
28	␣E␣	␣
29	␣-␣	␣
30	␣DAL B	␣DAL_B
31	␣LA␣	␣LA_
32	␣LE␣	␣LE_
33	␣DIT␣	-DIT-
34	␣MÁC AN␣	␣MÁC_AN_
35	␣MC␣	␣MC_
36	␣SAN␣	␣SAN_
37	␣OP HET␣	␣OP_HET_
38	␣DO ESPIRITO SANTO	␣DO_ESPIRITO_SANTO
39	␣DES␣	␣DES_

Tab. 3: Záměny ze začátku pole

	z	na
1	VON␣	VON_
2	VAN␣	VAN_
3	DELA␣	DELA_
4	DEL␣	DEL_
5	DOS␣	DOS_
6	DI␣	DI_
7	DA␣	DA_
8	AIT␣	AIT_
9	EL␣	EL_
10	AL␣	AL_
11	DE␣	DE_
12	DAL B	DAL_B
13	LA␣	LA_
14	LE␣	LE_
15	MÁC AN␣	MÁC_AN_
16	MC␣	MC_
17	OP HET␣	OP_HET_
18	DO ESPIRITO SANTO	DO_ESPIRITO_SANTO
19	O␣	O_
20	O'␣	O'_
21	LO␣	LO_
22	LI␣	LI_
23	DES␣	DES_
24	DO MONTE	DO_MONTE
25	Ó␣	Ó_
26	ZA␣	ZA_

Je zřejmé, že výše uvedené transformace neřeší všechny varianty chybných zápisů v rejstříku, neboť jejich pravopisná varianta nemusí být zřejmá. Například v tabulce č. 5 mohlo na první řádce jít například o paní „Alberovou hraběnku von Glanstätte-novou“, přičemž „Freifrau“ značí „hraběnka“. Příjmení „*Alberfreifrau*“ zřejmě vůbec nemusí existovat. Na druhé řádce mohlo jít například o příjmení „*Sukeľová*“, které je slovenského původu, tj. o háček na L, tj. Ľ, a nikoli o dlouhé Ĺ, které se sice vyskytuje ve slovenštině, ale ne v tomto příjmení. Na třetí řádce můžeme prakticky vyloučit, že jde o legitimní příjmení, s největší pravděpodobností jde o příjmení spojené spojovníkem „*Šebek-Loubalová*“ nebo o dvě příjmení, která měla být rozdělena mezerou, např. „*Šebek Loubalová*“. Na čtvrtém řádku se pak nachází vzácné příjmení, které se ale vyskytuje na internetu v této podobě. A u příjmení na pátém řádku Ministerstvo vnitra (2013) uvádí četnost 4, a tudíž jde také zřejmě o legitimní příjmení.

Tab. 5: Příklad neobvyklých záznamů příjmení

	příjmení
1	ALBERFREIFRAU VON_GLANSTÄTTENOVÁ
2	SUKEĹOVÁ
3	ŠEBEKLOUBALOVÁ
4	L'HELGOUALC'H
5	ŠÉ

Nejčastěji se v seznamu příjmení vyskytují taková, která mají pouze jeden člen (92,84 % ze všech příjmení, tj. 99,79 % populace). Jen sedm procent ze seznamu příjmení se skládá ze dvou až pěti prvků, viz tabulka č. 6. Tato vícenásobná příjmení sdílí 0,21 % populace.

Tab. 6: Počet prvků v příjmení

	četnost	v %	populace	v %
1 prvek	270 131	92,84	10 222 792	99,79
2 prvky	20 687	7,11	21 399	0,21
3 prvky	130	0,04	157	0,00
4 prvky	8	0,00	8	0,00
5 prvků	1	0,00	1	0,00

Po rozdělení složených příjmení na členy a jejich sloučením vznikl zúžením původních 270 136 příjmení a jejich četností seznam 251 723 unikátních příjmení (Protokol R, část 3. a 4.). Zároveň byla pomocná podtržítka zaměněna zpět za mezery, které tvoří součást příjmení. V tabulce č. 7 je znázorněn začátek a konec datového souboru. Sloupec četnost obsahuje původní četnost příjmení v rejstříku Ministerstva vnitra s tím, že členové složených příjmení přenesli četnost celého příjmení ke svým původcům. Sloupec vážená četnost obsahuje četnost příjmení, která vznikla rozdělením četnosti složeného příjmení mezi jeho konstitutivní části s tím, že suma tohoto sloupce odpovídá počtu osob registrovaných v rejstříku Ministerstva vnitra. Jinými slovy, sloupec četnost vykazuje počet příjmení v populaci, zatímco sloupec vážená četnost vykazuje počet osob, které mají ve svém jméně konkrétní příjmení.

Tab. 7: Začátek a konec datového souboru

řádek č.	příjmení	četnost	vážená čet.
1	AADI	1	1
2	AAFJES	3	3
3	AALBREGT	1	1
4	AALDERS	1	1
...	...	...	...
251720	ŻYWCZOKOVÁ	3	3
251721	ŽYWIAK	5	5
251722	ŽYWIAKOVÁ	4	4
251723	ZYZEN	1	1
suma		10 266 098	10 244 357

V tabulce č. 8 je znázorněn výčet deseti příjmení s nejčastějším výskytem (četnost) v evidenci obyvatel České republiky s tím, že v souboru jsou zvlášť zaznamenána mužská příjmení a příjmení v ženském přechýlení.

Tab. 8: Deset nejběžnějších příjmení

řádek č.	příjmení	četnost	vážená čet.
151993	NOVÁKOVÁ	35 314	35 182,3
151977	NOVÁK	33 954	33 951,0
216530	SVOBODOVÁ	26 479	26 396,5
152114	NOVOTNÁ	25 308	25 216,8
216522	SVOBODA	25 241	25 235,0
152126	NOVOTNÝ	24 267	24 265,0
46182	DVOŘÁKOVÁ	23 337	23 264,5
46177	DVOŘÁK	22 277	22 273,0
27972	ČERNÁ	18 427	18 391,5
28242	ČERNÝ	17 754	17 753,5

Vytvoření skupin možných přechýlených variant příjmení

Základní statistická informace o četnosti příjmení je zobrazena v tabulce č. 9. Cílem seskupení příjmení podle podobnosti je sloučit četnosti podobných příjmení a přiblížit se tak informaci o skutečném zastoupení sloučeného příjmení v populaci České republiky bez ohledu na přechylování.

Tab. 9: Statistika před sloučením

	četnost	vážená čet.
Min. :	1,00	0,25
1st Qu.:	2,00	2,00
Median :	4,00	4,00
Mean :	40,78	40,70
3rd Qu.:	15,00	15,00
Max. :	35 314,00	35 182,33

Triviální metoda

Tento algoritmus vyhledávání přechýlených příjmení je založen na následujících předpokladech ohledně podobnosti příjmení jako textových řetězců.

1. Přípona *-ová*: varianta mužská se od ženského přechýlení liší tím, že na konci příjmení ženských příjmení je přípona *-ová*, například pro příjmení *Novák* je užitá přechýlená varianta *Nováková*.

2. Samohláska *e*: varianta mužská se od ženské přechýlené liší četností samohlásky *e*. Jde například o rozdíl mezi příjmeními *Vellek* a *Vellková*.

3. Výskyt samohlásek *ý*, *y*, *a*, *á*: varianta mužská se od ženské přechýlení liší četností výskytu samohlásek *ý*, *y*, *a*, *á* na konci jména, například pár Novotný a Novotná.

4. Specifické pořadí znaků, ve které jsou příjmení o tyto znaky redukována při přípravě příjmení před seskupením: „*OVÁ*“, „*E*“, „*Ý*“, „*Y*“, „*Á*“ a „*A*“. Nejdříve je odstraněna přípona „*OVÁ*“ a naposledy grafém „A” na konci zbývajícího řetězce. Pořadí záměn (syntax) ovlivňuje výsledek, neboť všechny znaky kromě „*E*“ jsou odstraňovány z konce zbývajícího textového řetězce.

Identifikace podobných příjmení pak probíhá tak, že po odstranění *-ová* na konci, grafému *e* a také grafémů *ý*, *y*, *a*, *á* na konci proběhne identifikace shody ve skupině příjmení, která mají stejné dva počáteční grafémy. V tabulce č. 10 je výčet osmnácti nejčastěji se vyskytujících příjmení v evidenci (dle četnosti) spolu s příslušnými skupinami klíčových grafémů sloužící k identifikaci přechýlených párů (Protokol R, část 5.). Třetí sloupec (schéma) obsahuje redukci textového řetězce podle algoritmu. Ve druhém sloupci (skupina) jsou výsledné redukované řetězce, které slouží k seskupování příjmení. Z toho je zřejmé, že algoritmus identifikuje všechny zřejmé páry: *Nováková – Novák* (identické novák), *Svobodová – Svoboda* (identické svobod), *Novotná – Novotný* (identické novotn), *Dvořáková – Dvořák* (identické dvořák), *Černá – Černý* (identické čern).

Tab. 10: Triviální seskupování: příklad nejčetnějších příjmení

	příjmení	skupina	schéma	četnost	vážená čet.
1	NOVÁKOVÁ	NOVÁK	NOVÁK␣␣␣	35 314	35 182,33
2	NOVÁK	NOVÁK	NOVÁK	33 954	33 951,00
3	SVOBODOVÁ	SVOBOD	SVOBOD␣␣␣	26 479	26 396,50
4	SVOBODA	SVOBOD	SVOBOD␣	25 241	25 235,00
5	NOVOTNÁ	NOVOTN	NOVOTN␣	25 308	25 216,83
6	NOVOTNÝ	NOVOTN	NOVOTN␣	24 267	24 265,00
7	DVOŘÁKOVÁ	DVOŘÁK	DVOŘÁK␣␣␣	23 337	23 264,50
8	DVOŘÁK	DVOŘÁK	DVOŘÁK	22 277	22 273,00
9	ČERNÁ	ČRN	Č␣RN␣	18 427	18 391,50
10	ČERNÝ	ČRN	Č␣RN␣	17 754	17 753,50
11	PROCHÁZKOVÁ	PROCHÁZK	PROCHÁZK␣␣␣	16 654	16 604,83
12	PROCHÁZKA	PROCHÁZK	PROCHÁZK␣	16 006	16 001,50
13	KUČEROVÁ	KUČR	KUČ␣R␣␣␣	15 845	15 792,00
14	KUČERA	KUČR	KUČ␣R␣	15 155	15 153,00
15	VESELÁ	VSL	V␣S␣L␣	13 599	13 542,00
16	VESELÝ	VSL	V␣S␣L␣	12 912	12 910,50
17	HORÁKOVÁ	HORÁK	HORÁK␣␣␣	12 680	12 638,00
18	HORÁK	HORÁK	HORÁK	12 208	12 208,00

Přílohou tohoto článku je počítačový program v prostředí R, který umožňuje provést tento algoritmus na celé databázi (Protokol R, část 6.). Tento algoritmus může být spuštěn na obyčejné pracovní stanici, přičemž byly jeho různé varianty vzhledem k velkému počtu zkoumaných kombinací spuštěny a testovány v Českém národním počítačovém centrum MetaCentrum.[6)] Celkem algoritmus navrhuje pro zkoumaný datový soubor vytvořit 142 586 skupin přechýlených a podobných příjmení a do nich sloučit celkových 251 723 příjmení. Jde o snížení počtu zkoumaných položek zhruba o polovinu (43,36 %).

Tabulka č. 12 obsahuje seznam 15 skupin nejčetnějších příjmení sloučených podle podobnosti. Je zřejmé, že algoritmus dobře seskupuje přechýlená příjmení s mužskými formami a občas zahrne do skupin méně podobná příjmení, viz příjmení *Marke*, *Markey* ve skupině „*Mark*“ dominované formami *Marek* a *Marková*. Tato příjmení méně podobná dominantním příjmením mají četnost 1 až 9, což je bez vlivu na pořadí

[6)] Zde patří poděkování MetaCentrum, které je dle zvyklostí v anglickém jazyce: I am indebted to the Czech national computing grid infrastructure “MetaCentrum” (www.metacentrum.cz) for allowing me to access to the shared supercomputing and large storage facilities across the contributing parties and projects under the joint programme “Projects of Large Infrastructure for Research, Development, and Innovations” (LM2010005).

skupin. Výjimkou je zřejmě mužské příjmení *Mark*, která má četnost 68. To se však může vázat k přechýlenému „*Marková*“, stejně jako „*Marek*“.

Tab. 12: Nejčetnější příjmení podle skupin, alg. D2

	supina	četnost	příjmení
1	NOVÁK	69 268	NOVÁK, NOVÁKOVÁ
2	SVOBOD	51 720	SVOBODA, SVOBODOVÁ
3	NOVOTN	49 583	NOVOTNA, NOVOTNÁ, NOVOTNY, NOVOTNÝ
4	DVOŘÁK	45 614	DVOŘÁK, DVOŘÁKOVÁ
5	ČRN	36 191	ČERNA, ČERNÁ, ČERNAY, ČERNOVÁ, ČERNÝ
6	PROCHÁZK	32 660	PROCHÁZKA, PROCHÁZKOVÁ
7	KUČR	31 000	KUČERA, KUČEROVÁ
8	VSL	26 519	VESELA, VESELÁ, VESELY, VESELÝ
9	HORÁK	24 888	HORÁK, HORÁKOVÁ
10	NĚMC	22 781	NĚMCOVÁ, NĚMEC
11	MARK	22 622	MAREK, MAREKOVÁ, MARK, MARKA, MARKE, MARKEY, MARKOVÁ
12	POSPÍŠIL	21 949	POSPÍŠIL, POSPÍŠILOVÁ
13	POKORN	21 827	POKORNA, POKORNÁ, POKORNY, POKORNÝ
14	HÁJK	21 019	HÁJEK, HÁJEKOVÁ, HÁJKA, HÁJKOVÁ
15	KRÁL	20 430	KRÁL, KRÁLOVÁ

Tabulka č. 11 obsahuje statistiku četnosti příjmení po sloučení do skupin. Z ní je v porovnání s tabulkou č. 9 zřejmé, že došlo ke změně rozdělení zkoumaných četností.

Tab. 11: Statistika po sloučení, alg. D2

	četnost	vážená čet.
Min. :	1	0,25
1st Qu.:	2	2,00
Median :	6	6,00
Mean :	72	71,85
3rd Qu.:	23	23,00
Max. :	69 268	69 133,33

Například průměrná četnost příjmení stoupla ze 40,78 na 72, přičemž medián souboru se zvýšil ze 4 na 6. Měřeno váženou četností devět z deseti obyvatel České republiky sdílí některé ze 22 594 nejběžnějších seskupených příjmení, tj. některé z 15,85 procent seskupených příjmení. Vzácná seskupená příjmení, která sdílí méně než 10 procent obyvatel České republiky, mají četnost výskytu menší než 51. Velmi vzácná příjmení, která sdílí méně než 1 procento obyvatel České republiky, mají četnost výskytu menší než 4 osoby na příjmení. Těchto skupin velmi vzácných příjmení je 56 728, tj. 39,79 procent z celkového počtu skupin.

Další metoda seskupení může být například pomocí měření Levenshteinovy vzdálenosti (procedura adist v R); ta však nebyla při psaní tohoto článku zohledněna.

Přílohu článku tvoří soubory vystavené na adrese http://sebekj.blogspot.cz/2014/06/priloha-k-seskupeni-prechylenych.html. Jedná se o tyto soubory:

1. PDF: Protokoly v prostředí R (zpracováno pomocí *R Core Team* 2014)
2. CSV: Datový soubor "Acta.Onomastica.2014.source.master.csv" (získáno z *Ministerstvo vnitra* 2013) ke zpracování pomocí protokolu "Protokoly v prostředí R." Tento datový soubor již nahrazuje části "Protokolů v prostředí R" od řádky č. 1 do řádky 14.
3. CSV: Výsledný datový soubor "Acta.Onomastica.2014.unique.csv" byl získán úpravou datového souboru "source.master.csv". Datový soubor je výsledkem implementace "Protokolů v prostředí R" v celé délce.
4. PDF: Podklad pro tisk: Tabulky a grafy, R a LaTeX

Literatura

CESNET (2013) [výpočetní infrastruktura], Národní gridová infrastruktura MetaCentrum, Praha, <http://www.metacentrum.cz/>.

D. B. Dahl (2013), [software], Xtable: Export Tables to LaTeX or HTML. <http://cran.r-project.org/package=xtable>.

M. Hlaváč (2013), [software], Stargazer: LaTeX Code and ASCII Text for Well--Formatted Regression and Summary Statistics Tables, <http://cran.r-project.org/package=stargazer>.

Ministerstvo vnitra (2013) [online], Četnost jmen a příjmení (Frequency of Names and Surnames), Praha, <http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx> [cit. 3. srpna 2013].

J. Novotný – J. A. Cheshire (2012), The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-Occurrence of Surnames. PloS ONE 7 (10) (leden), e48568. doi:10.1371/journal.pone.0048568, <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3485322&tool=pmcentrez&rendertype=abstract>.

R CORE TEAM (2013) [software], R: A Language and Environment for Statistical Computing, Vienna, R Foundation for Statistical Computing, <http://www.r-project.org/>.

sebekj@gmail.com
Filozofická fakulta Univerzity Karlovy v Praze
nám. Jana Palacha 2
116 38 Praha 1