Fonologický lexikální korpus češtiny a slabičná struktura českého slova[1]

Aleš Bičan

V prvé části tohoto článku představíme Fonologický lexikální korpus češtiny a v části druhé naznačíme možnosti, jak jej lze využít pro popis fonologie současné češtiny. Fonologický lexikální korpus podává nejen údaje o frekvenci jednotlivých fonémů a jejich kombinací v českých slovech (jako celku či v konkrétních slovních druzích), ale zároveň poskytuje nové poznatky o fonologické struktuře českých slov. Zde se zaměříme na slabičnou strukturu slov, které dosud nebyla věnována dostatečná pozornost. Zajímat nás bude souvýskyt slabik podle tří částí slabiky, a to podle typu slabičného jádra, podle přítomnosti oproti nepřítomnosti slabičné kódy (otevřenost slabiky) a podle komplexnosti slabičné iniciály. Ukážeme, že ve všech případech čeština preferuje určité konstelace slabik ve slovech.

1. Fonologický lexikální korpus češtiny

1.1 O užitečnosti korpusů pro analýzu a pochopení jazyka nelze pochybovat. Lingvisty jsou využívány ke gramatickým, lexikálním, sociolingvistickým a jiným analýzám. Mají však své opodstatnění i pro zkoumání zvukové stránky jazyků, o čemž svědčí četné fonetické korpusy (Durand a kol. 2014). Pro češtinu existuje např. Pražský fonetický korpus (Skarnitzl 2010), ale též korpusy mluveného jazyka (je jich několik, např. Olomoucký mluvený korpus). Na rozdíl od fonetických korpusů neexistuje mnoho korpusů fonologických, byť se tyto dva typy korpusů mnohdy překrývají. Zatímco se fonetické korpusy využívají pro popis skutečných zvukových realizací, záměrem fonologických korpusů je podat informaci o využití zvukových prostředků v daném jazyce a o formální stavbě jeho slov, vět či jiných významových jednotek. Pro češtinu nebyl dosud žádný fonologický korpus zveřejněn, přestože jich předchozí lingvisté, v omezené míře dané technickými možnostmi své doby, využívali (Mathesius 1929; Vachek 1940; Kučera – Monroe 1968; Trnka 1966; Ludvíková in Těšitelová 1985).

Korpusy zpravidla sestávají ze skutečných textů, ale za korpus lze považovat i lexikální databáze, tedy soubor slov uvedených například ve slovnících. Pro různé jazyky existuje několik lexikálních korpusů (např. databáze CELEX pro angličtinu, němčinu a nizozemštinu). Také jich lze využít pro zkoumání

[1] Příspěvek vznikl za podpory grantu 13-15361P Problémy ve fonologii slova v češtině (GAČR).

fonologických vlastností určitého jazyka (srov. např. Frisch 2012), popř. pro porovnání typologických fonologických vlastností různých jazyků (Rousset 2004).

Poněvadž se fonologické struktuře českého lexika dosud nikdo ve větším měřítku nevěnoval, rozhodli jsme se vytvořit Fonologický lexikální korpus češtiny (FLK).[2] FLK je soubor lexikální zásoby moderní češtiny převedený do fonologické transkripce. V současné době obsahuje 275 805 položek zaznamenaných v nejdůležitějších slovnících češtiny z 20. a 21. století (viz dále oddíl 1.4). Jeho samostatnou součástí je několik subkorpusů se seznamem názvů českých obcí a jejich částí, seznamem nejčastějších křestních jmen a jejich domácích variant a seznamem českých botanických a zoologických názvů.

Data ve FLK jsou uložena v prostém textu v souborech ve formátu CSV (Comma-Separated Value). Díky tomuto formátu lze data prohlížet, editovat a filtrovat jako tabulky v editorech podporujících uvedený formát (např. Microsoft Excel). Každá lexikální položka, odpovídající jednomu řádku v tabulce, se skládá ze své ortografické podoby a z fonologické interpretace jiné předpokládané zvukové realizace (viz oddíl 1.2). Dále je ke každé položce přiřazena informace o jejich fonologických vlastnostech (1.3) a informace o slovním druhu a výskytu ve slovnících (1.4). Tyto informace je možné následně využít k vyhledávání v korpusu a k jeho třídění, takže lze například zjistit, zda ve *Slovníku spisovné češtiny* existují přídavná jména o pěti slabikách obsahující znělou velární okluzívu, slabičné /r/ a vysoké přední vokály.

Podrobný popis FLK je uveden na webových stránkách citovaných v poznámce 2 (Bičan, ms.).

1.2 Ve FLK je každé lexikální položce přiřazena fonologická interpretace její zvukové realizace, která odpovídá předpokládané ortoepické výslovnosti podle pravidel formulovaných v publikaci *Výslovnost spisovné češtiny 1* (Hála 1968). Fonologická transkripce byla v první fázi získána automatickým převodem z ortografické podoby pomocí počítačového programu.[3] Výsledek byl poté ručně překontrolován.

Pro mnoho slov, především cizího původu, jsou v publikovaných slovnících uvedeny různé způsoby zápisu, přestože je jejich výslovnost stejná, např. *filosofie* × *filozofie*, tj. [filozofiɟɛ]. Takové lexémy se sice ve FLK vyskytují jako dvě samostatné položky, ale jejich fonologická transkripce je stejná. Vzhledem k velikosti korpusu jsme se rozhodli duplicitní položky nesjednocovat. Vzniklý problém se při vyhodnocování dat částečně řeší rozlišováním mezi tokeny a typy (jako token se /filozofije/ v korpusu objevuje dvakrát, jako typ jen jednou). Tento postup má

[2] <http://www.ujc.cas.cz/phword>.

[3] Program prozatím slouží k interním účelům, avšak autor nevylučuje možnost jeho pozdějšího zveřejnění.

ovšem nevýhodu v tom, že např. homofona *stát* (podstatné jméno) a *stát* (sloveso) jsou nutně hodnocena jako jeden typ.

Fonologická transkripce slov je vizualizace fonologické analýzy daných jazykových faktů a jako každá analýza se řídí principy a metodologií určité teorie. Tou je pro nás teorie funkční fonologie formulovaná pražskou školou, především Nikolajem Trubeckým (1939), a dále rozvinutá André Martinetem (Martinet 2011) a Janem Mulderem (Mulder 1989). Detaily jsou podrobně popsány v Bičan (2013) a nejsou pro účely tohoto článku podstatné.

Základní podobou fonologické transkripce je posloupnost fonémů, která je rozdělena do fonologických slov. Fonologické slovo je fonologicky definovaná jednotka, která není nutně totožná s gramaticky (tj. nefonologicky) definovaným slovem, a to ani v češtině (Bičan 2014), ani v mnoha jiných jazycích (Dixon – Aikhenvald 2003). Pomocí speciálních značek je ovšem v transkripci naznačena i hranice gramatických slov.

V češtině je fonologické slovo vymezeno především přízvukem. Není menší než přízvukový takt, ale nemusí se s ním nutně shodovat. Za signály jeho hranic totiž považujeme i výskyt tzv. rázu a dalších zvukových prostředků (viz Bičan 2014). Ráz se v češtině vyslovuje na začátku gramatických slov před vokálem (např. *oko*) nebo na hranicích předpony a základu slova, popř. v kompozitech, začíná-li základ slova, resp. druhá část kompozita vokálem. Proto se například gramatická slova jako *naučit*, *velkoobchod*, která v izolaci odpovídají jednomu přízvukovému taktu, skládají ze dvou fonologických slov. Ráz (či minimálně jeho potenciální výskyt) signalizuje jejich fonologickou složenost.[4]

Ve zvláštním sloupci je dále z praktických důvodů u každé položky naznačeno, jak ji slabikovat, tedy dělit do fonologicky definovaných slabik. Pravidla slabikování, jež byla v korpusu využita, vycházejí z pravidel navržených Kučerou a Monroem (1968), jsou však uzpůsobena zvolené teorii a metodologii. Detaily viz Bičan (ms.).

1.3 Z fonologické transkripce jsou odvozeny fonologické vlastnosti, které jsou paralelou gramatické anotace v nefonologických korpusech. Kromě informace o délce slova podle počtu fonémů a slabik je u hesel naznačen jejich konsonanticko-vokalický vzorec. Každý foném je v češtině buď neslabičný (tj. konsonant – C), nebo slabičný (tj. vokál nebo slabičná sonanta /r/ a /l/ – V), takže např. formy /voda/ a /vlna/ mají oba vzorec CVCV.

Dále heslo obsahuje informace o distinktivních rysech fonémů, z nichž se skládá. Distinktivní rysy zde chápeme jako analytické vlastnosti fonémů, jež jsou modelem (či projekcí) zvukových vlastností hlásek (Mulder 1989). Pouze

[4] Budeme-li dále bez jakéhokoliv přívlastku hovořit o slovech, budeme tím mínit slova fonologická.

ty vlastnosti, které jsou relevantní pro komunikaci, mohou být distinktivními rysy. Toto chápání se nutně neshoduje s univerzálně pojatými teoriemi Romana Jakobsona (Jakobson – Halle 1956) či Noama Chomského a jejich následovníků (Chomsky – Halle 1968). V pojetí funkční fonologie nejsou distinktivní rysy nutně binární a univerzální pro všechny jazyky. Různé teorie nicméně spojuje fakt, že distinktivní rysy odlišují jeden foném od fonémů jiných. V češtině u konsonantů proto rozlišujeme distinktivní rysy, které odpovídají místu a způsobu artikulace a znělosti jejich alofonů, u vokálů pak rysy, které odpovídají horizontální a vertikální poloze jazyka a kvantitě jejich alofonů.

1.4 Dalším oddílem FLK je informace o slovních druzích konkrétních hesel. Slovní druh byl přiřazen automaticky pomocí desambiguačního nástroje Desamb[5] a následně manuálně zkontrolován. Podle slovních druhů lze nejen korpus třídit, ale především zjišťovat, zda se fonologická struktura jednotlivých slovních druhů navzájem liší. Výsledky ukazují, že takové rozdíly skutečně existují, a to například v tom, že u neohebných slovních druhů je struktura slabiky jednodušší než u slovních druhů ohebných (viz Bičan 2015).

Konečně posledním oddílem každé lexikální položky je informace o slovnících, v nichž se položka objevuje, což nabízí další možnost, jak fonologickou strukturu slov srovnávat. Korpus zahrnuje slovo (lemmata) z publikovaných slovníků moderní češtiny obsažených v Databázi heslářů.[6] Jedná se o *Slovník spisovné češtiny pro školu a veřejnost* (2003), *Slovník spisovného jazyka českého I–IV* (1960–1971), *Příruční slovník jazyka českého I–VIII* (1935–1957), *Co v slovnících nenajdete: novinky v současné slovní zásobě* (1994), *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves* (1997), *Nová slova v češtině. Slovník neologizmů 1* (1998), *Nová slova v češtině. Slovník neologizmů 2* (2004), *Slovník slovesných, substantivních a adjektivních vazeb a spojení* (2005), *Frekvenční slovník češtiny* (2010). Dále jsme korpus doplnili o seznam slov cizího původu z výslovnostního slovníku *Výslovnost spisovné češtiny* (Romportl 1978). Slovník je sice dnes již poněkud zastaralý, stále však nabízí nejucelenější přehled výslovnosti slov, která byla do češtiny přejata z různých jazyků. Díky němu lze porovnat fonologickou strukturu lexémů domácího původu a lexémů cizího původu.

2. Slabičná struktura českých slov

2.1 FLK jako celek nebo jakoukoliv jeho část je možné automaticky vyhodnotit pomocí počítačového programu, jenž byl pro tento úkol vyvinut. Takto získáme přesné informace o frekvenci jednotlivých fonémů, kombinacích fonémů, slabik, fonologických slov, popř. jiných definovatelných kategorií včetně gramatic-

[5] <http://nlp.fi.muni.cz/projekty/rule_ind/>.

[6] Viz <http://lexiko.ujc.cas.cz/heslare/>.

kých slov. Podobné údaje se doposud získávaly jen z omezeného množství dat. Např. statistické údaje o fonologii češtiny v knize *Kvantitativní charakteristiky současné češtiny* (Těšitelová a kol. 1985) jsou založeny jen na vzorku 5 000 slabik.[7] Kvůli technickým omezením se předchozí popisy některými aspekty fonologické struktury češtiny vůbec nezabývaly. Mezi ně patří například otázka souvýskytu slabik různé struktury uvnitř českých slov. Slabika v češtině má bohatou, avšak omezenou strukturu (Bičan 2013). Legitimní otázkou potom je, zda mohou být česká slova tvořena jakoukoliv posloupností povolených slabik. A pokud ano, jsou některé posloupnosti preferované, zatímco jiné jsou vzácné? Abychom ukázali užitečnost našeho korpusu, pokusíme se na tyto otázky odpovědět. Zaměříme se na souvýskyt slabik podle typu slabičného jádra (2.2), na souvýskyt slabik podle jejich otevřenosti či zavřenosti (2.3) a na souvýskyt slabik podle komplexnosti slabičné iniciály (2.4). Případnou podrobnější interpretaci nebo vysvětlení zjištěných faktů musíme s ohledem na délku článku ponechat do samostatné studie.

Jelikož FLK vzniká v rámci grantu, jenž je v současnosti stále řešen, bude v celé své šíři zveřejněn po skončení grantového projektu v roce 2016. Následující analýza je založena na lexikální zásobě ze *Slovníku spisovné češtiny* (SSČ) a tato část korpusu je na webových stránkách korpusu volně přístupná spolu s jejím vyhodnocením. Čtenář si tak může nezávisle ověřit naše závěry. SSČ byl zvolen, protože ze všech slovníků obsažených ve FLK nejlépe zaznamenává slovní zásobu současné češtiny.[8] Zvolený vzorek čítá 49 365 lexikálních položek a je tedy dostatečně velký, aby nabídl užitečné informace. Lexikální položky v SSČ odpovídají 45 978 fonologickým slovům (dále jen slovům) jako typům, jež se dělí na 146 703 slabik. Poněvadž nás zajímá souvýskyt slabik, nebudeme se dále zabývat jednoslabičnými slovy (3,44 % všech slov). Hlavní pozornost bude věnována slovům o 2–5 slabikách, která tvoří 94,09 % slov (43 263 slov). Slova o více slabikách než pět tvoří jen 2,47 % (nejdelší jsou v našem vzorku slova o deseti slabikách).

2.2 Slabičnou strukturu slov můžeme v prvé řadě popsat podle typu slabičného jádra. V češtině ho tvoří vokál (krátký, dlouhý, diftongální), nebo slabičná sonanta (/r/, /l/).[9] FLK ukazuje, že slabiky s vokály jasně převažují nad slabikami

[7] Poznamenejme, že byť fonologickou část má i kniha *Statistiky češtiny* (Bartoň a kol. 2009), jež uvádí různé kvantitativní vlastnosti Českého národního korpusu (SYN2005), její autoři se zabývají frekvencí *hlásek* a *hláskovou* strukturou slov. Proto například autoři berou v potaz existenci rázu (glotální okluzívy), kterážto hláska však není v češtině alofonem žádného fonému.

[8] Na rozdíl od *Slovníku spisovného jazyka českého* nebo *Příručního slovníku jazyka českého*. Ostatní slovníky zařazené do FLK musíme chápat jako doplňky či dodatky k těmto třem nejdůležitějším slovníkům češtiny.

[9] Se slabičným /m/ nepracujeme, poněvadž slova *sedm* a *osm* přepisujeme jako /sedum/ a /osum/.

se slabičnými sonantami (98,72 % oproti 1,28 %). U vokálů pak jednoznačně převažují slabiky tvořené krátkým vokálem (77,62 %). Na druhém místě jsou slabiky tvořené dlouhým vokálem (19,1 %) a pak slabiky s diftongálním vokálem (2 %).

Celkově tedy krátké vokály převažují nad nekrátkými vokály (kam pro zjednodušení řadíme i slabičné sonanty). Tabulka 1 ukazuje, jak se na dvoj- až pětislabičných slovech projevuje preference krátkých vokálů oproti ostatním typům slabičných jader (= nekrátké vokály). Šedým podkladem jsou zvýrazněny nejčastější typy slov. Poslední řádek uvádí celkový počet slov podle počtu slabik; stejné počty platí i pro ostatní tabulky.

Počet slabik s nekrátkým vokálem	**Počet slabik ve slově**			
	2	3	4	5
0	48,39 %	40,2 %	39,08 %	36,42 %
1	46,6 %	46,67 %	42,55 %	44 %
2	5,02 %	12,72 %	16,91 %	17,58 %
3	–	0,41 %	1,43 %	1,98 %
4	–	–	0,03 %	0,03 %
5	–	–	0 %	0 %
Celkem slov	10 269	18 013	11 499	3 482

Tabulka 1: Procentuální zastoupení slov o daném počtu nekrátkých vokálů (tj. počtu slabičných jader jiných než krátkých vokálů)

S výjimkou dvojslabičných slov jsou nejvíce doložena slova, která kromě krátkého vokálu obsahují právě jeden vokál nekrátký, a teprve na druhém místě jsou slova se všemi jádry tvořenými krátkými vokály. Pak už shodně platí, že se zvětšujícím se počtem nekrátkých vokálů klesá i četnost takových slov. Nejméně častá jsou slova se čtyřmi nekrátkými vokály a slova s pěti nekrátkými vokály nejsou doložena vůbec. Za pozornost stojí fakt, že opět s výjimkou dvojslabičných slov je procentuální poměr slov o určitém počtu nekrátkých vokálů srovnatelný u slov různé slabičné délky. Teoreticky by totiž dvojslabičná slova mohla být tvořena jen nekrátkými vokály a i potom by celkově mohly být krátké vokály častější než vokály nekrátké.

Z preference krátkých vokálů tedy nutně nevyplývá, že nejčastější budou jen slova s krátkými vokály. Tabulka 1 rozlišovala pouze mezi dvěma typy slabičných jader, tj. krátké vokály a ostatní typy slabičných jader. Rozdělíme-li však druhou skupinu do jednotlivých tříd (dlouhé vokály, diftongální vokály a slabičné /r/ a /l/), pak zjistíme, že celkově nejčastějšími jsou přece jenom slova, jež obsahují vokály krátké. Označíme-li slabiky podle typu jádra písmeny K (krátký vokál ve slabice), D (dlouhý vokál), F (diftongální vokál) a R (sonanta), můžeme rozlišovat různé druhy kvantitativních vzorců slov (např. /rolāda/ má vzorec KDK).

Tabulka 2 nabízí hierarchii pěti nejčastějších kvantitativních vzorců. Ve všech případech jsou nejčastější slova obsahující jen krátké vokály (tj. vzorce KK, KKK atd.). U dvojslabičných slov se k tomuto typu řadí téměř polovina slov; u ostatních typů zastoupení vždy přesahuje 35 %.[10]

Pořadí	**Počet slabik ve slově**			
	2	3	4	5
1	KK (48,39 %)	KKK (40,2 %)	KKKK (39,08 %)	KKKKK (36,42 %)
2	KD (20,95 %)	KKD (20,42 %)	KKKD (24,02 %)	KKKKD (30,62 %)
3	DK (16,14 %)	KDK (13,22 %)	KKDK (8,4 %)	KKKDD (6,61 %)
4	KF (3,59 %)	DKK (6,41 %)	KLKD (5,08 %)	KKKDK (5,4 %)
5	FK (2,86 %)	DKD (4,99 %)	KKDD (4,54 %)	KKDKD (4,68 %)

Tabulka 2: Pět nejčastějších kvantitativních vzorců a jejich procentuální zastoupení (K = krátký vokál ve slabice, D = dlouhý vokál ve slabice, F = diftongální vokál ve slabice)

2.3 Druhou možností, jak popsat podle typu slabik celkovou strukturu českých slov, je porovnat v nich výskyt otevřených a zavřených slabik, tedy zaměřit se na slabičnou kódu. Otevřená je slabika končící na vokál (69,99 % slabik je v SSČ tohoto typu), zatímco zavřená slabika končí na konsonant (30,01 % slabik). Otevřených slabik je více než dvě třetiny z celkového počtu slabik, avšak z toho opět nutně nevyplývá, že nejčastější budou slova obsahující pouze otevřené slabiky. To potvrzuje tabulka 3, v níž uvádíme procentuální zastoupení dvoj- a pětislabičných slov s určitým počtem zavřených slabik. Šedým podkladem jsou opět zvýrazněny nejčastější typy slov.

Počet zavřených slabik	**Počet slabik ve slově**			
	2	3	4	5
0	22,05 %	25,08 %	29,52 %	29,32 %
1	62,72 %	55,36 %	48,28 %	45,72 %
2	15,23 %	18,4 %	19,67 %	20,36 %
3	–	1,17 %	2,45 %	4,19 %
4	–	–	0,07 %	0,37 %
5	–	–	–	0,03 %

Tabulka 3: Procentuální zastoupení slov o daném počtu zavřených slabik

[10] Preference slov se všemi slabikami tvořenými krátkými vokály není v rozporu s faktem, že nejčastější jsou slova s jedním nekrátkým vokálem. Nekrátký vokál totiž může být v jakékoliv pozici slova a v součtu je takových slov více než slov se všemi krátkými vokály.

Zcela jednotně jsou nejvíce zastoupena nikoliv slova se všemi slabikami otevřenými, nýbrž slova s právě jednou slabikou zavřenou a ostatními otevřenými (vždy více než 45 %). Dalším pozoruhodným faktem, v němž se různoslabičná slova shodují, je preference slov o určitém počtu zavřených slabik. Ve všech případech jsou druhým nejčastějším typem slova neobsahující ani jednu zavřenou slabiku a dále slova se dvěma zavřenými slabikami. Konečně pro všechna slova platí, že se vzrůstajícím počtem zavřených slabik klesá i četnost slov. Slova s pěti zavřenými slabikami jsou velmi vzácná – v SSČ jsou doložena jen dvě (*interkontinentální*, *pluskvamperfektum*). Slova s vyšším počtem zavřených slabik nejsou doložena vůbec, a to ani v celém FLK.

Stejně jako u slabičných jader, tak i v případě otevřenosti můžeme rozlišovat různé vzorce slov podle typu slabik. Označme si otevřené slabiky písmenem O a zavřené slabiky písmenem Z. Hierarchie pěti nejčastějších vzorců podle otevřenosti slabiky a jejich procentuální zastoupení nabízí tabulka 4. U čtyř- a pětislabičných slov jsou shodně nejčastější slova, v nichž jsou všechny slabiky otevřené. U dvoj- a trojslabičných slov jsou taková slova až na druhém místě; u nich jsou nejčastější slova, v nichž je koncová slabika zavřená. Slova s jednou slabikou zavřenou jsou velmi častá i v případě čtyř- a pětislabičných slov. Preferenci slov s jednou zavřenou slabikou ostatně potvrzuje již tabulka 3.

Pořadí	**Počet slabik ve slově**			
	2	3	4	5
1	OZ (47,99 %)	OOZ (31,93 %)	OOOO (29,52 %)	OOOOO (29,32 %)
2	OO (22,05 %)	OOO (25,08 %)	OOOZ (21,01 %)	OOOZO (16,05 %)
3	ZZ (15,23 %)	OZO (15,31 %)	OOZO (12,9 %)	OOOOZ (12,67 %)
4	ZO (14,73 %)	ZOZ (9,21 %)	ZOOO (7,97 %)	ZOOOO (8,73 %)
5	–	ZOO (8,12 %)	ZOOZ (6,76 %)	ZOOZO (4,54 %)

Tabulka 4: Pět nejčastějších vzorců slov podle otevřenosti slabik (O = otevřená slabika, Z = zavřená slabika)

2.4 Další možnost k popsání slabičné struktury slov nabízí slabičná iniciála. Ta může buď chybět (4,52 % slabik je v SSČ tohoto typu), nebo může být jednoduchá, tj. tvořena jedním konsonantem (69,44 %), nebo je komplexní, tj. tvořena kombinací konsonantů (26,04 %). Je zjevné, že slabiky s jednoduchou iniciálou převládají, ale ani zde nelze tvrdit, že právě proto budou v češtině nejčastější slova tvořená slabikami tohoto typu. Tabulka 5 shrnuje procentuální zastoupení slov podle počtu slabik a podle počtu v nich obsažených slabik s nejednoduchou iniciálou, tedy iniciálou jinou než tvořenou jedním konsonantem.

Počet slabik s nejednoduchou inic.	Počet slabik ve slově			
	2	3	4	5
0	35,33 %	29,69 %	28,72 %	25,59 %
1	48,09 %	45,28 %	42 %	41,84 %
2	16,58 %	21,33 %	23,02 %	25,27 %
3	–	3,69 %	5,77 %	6,2 %
4	–	–	0,48 %	1,03 %
5	–	–	–	0,06 %

Tabulka 5: Procentuální zastoupení slov o daném počtu slabik s nejednoduchou iniciálou (tj. jinou iniciálou než tvořenou jedním konsonantem)

Pozoruhodný až překvapivý je fakt, že bez ohledu na počet slabik ve slově jsou vždy nejčastější slova obsahující jednu slabiku s nejednoduchou iniciálou a ostatní slabiky s jednoduchou iniciálou. Takových slov je vždy více než 41 %. Na druhém místě jsou shodně slova s žádnou nejednoduchou iniciálou a pak platí, že se vzrůstajícím počtem slabik s nejednoduchou iniciálou klesá i četnost takových slov. Slova s pěti slabikami s nejednoduchou iniciálou jsou vzácná a slova se šesti a více takovými slabikami nejsou doložena v celém FLK.

Konečně se zastavme u vzorců slov podle komplexnosti iniciály slabiky. Označme si písmenem S slabiku s jednoduchou iniciálou, M slabiku s komplexní iniciálou a B slabiku bez iniciály. Pořadí pěti nejčastějších vzorců slov zobrazuje tabulka 6. Z ní je patrno, že nejčastější jsou vždy slova se slabikami, jejichž iniciálu tvoří jeden konsonant (více než čtvrtina slov). Tento závěr ovšem není v rozporu s výše uvedeným konstatováním, že v češtině jsou nejčastější slova s jednou slabikou mající nejednoduchou iniciálu. Je třeba si uvědomit, že nejednoduchá iniciála může být v jakékoliv pozici ve slově. V součtu je takových slov více než slov se všemi slabikami typu SS, SSS, SSSS a SSSSS.

Pořadí	Počet slabik ve slově			
	2	3	4	5
1	SS (35,33 %)	SSS (29,69 %)	SSSS (28,72 %)	SSSSS (25,59 %)
2	MS (28,79 %)	MSS (16,64 %)	MSSS (9,96 %)	SSSSM (10,86 %)
3	SM (15,74 %)	SSM (12,18 %)	SSSM (9,55 %)	BSSSS (9,08 %)
4	MM (14,22 %)	SMS (10,88 %)	SMSS (9,06 %)	MSSSS (7,32 %)
5	BS (3,5 %)	MSM (5,5 %)	BSSS (7,46 %)	SMSSS (5,6 %)

Tabulka 6: Pět nejčastějších vzorců slov podle komplexnosti iniciály (S = slabika s jednoduchou iniciálou, M = slabika s komplexní iniciálou, B = slabika bez iniciály)

3. Shrnutí a závěr

V našem příspěvku jsme pro češtinu popsali slabičnou strukturu fonologických slov na základě dat z Fonologického lexikálního korpusu češtiny (FLK). Součástí korpusu je slovní zásoba zaznamenaná v nejdůležitějších slovnících češtiny vydaných v průběhu 20. a 21. století (viz oddíl 1.4), avšak pro naši analýzu jsme se rozhodli omezit jen na zásobu ze *Slovníku spisovné češtiny* (SSČ). Ze všech slovníků v korpusu obsažených totiž nejlépe zachycuje slovní zásobu současné češtiny, byť pochopitelně netvrdíme, že je v něm odražen aktuální stav české slovní zásoby. Základní jednotkou analýzy bylo fonologicky definované fonologické slovo, které není nutně totožné s nefonologicky definovaným gramatickým slovem. V češtině fonologické slovo vymezuje především přízvuk a výskyt hraničních signálů jako ráz. Analýzu jsme omezili na dvoj- až pětislabičná slova, jež tvoří 94,09 % zvoleného vzorku, tj. celkem 43 263 fonologických slov. Věnovali jsme se souvýskytu slabik podle jejich struktury jádra, kódy a iniciály. V prvém případě nás zajímal výskyt slabik podle toho, který foném fungoval jako slabičné jádro (krátký, dlouhý, diftongální vokál či slabičné /r/, /l/). V druhém případě jsme se soustředili na distribuci otevřených a zavřených slabik. Ve třetím případě jsme popsali výskyt slabik podle komplexnosti slabičné iniciály (jeden konsonant, konsonantická kombinace nebo nulová iniciála).

Tím jsme samozřejmě nevyčerpali všechny možnosti, jak v češtině popsat slabičnou strukturu slov. Místo otevřenosti slabiky by bylo možné slova rovněž popsat podle komplexnosti slabičné kódy, tj. podle toho, zda kóda končí na jeden konsonant, konsonantickou kombinaci nebo zda kóda chybí. Poněvadž na konsonantickou kombinaci končí jen 4,46 % slabik, nebyly by výsledky tak zajímavé, a proto jsme se zabývali pouze přítomností a nepřítomností slabičné kódy. U iniciál jsme však komplexnost zohlednili, jelikož na konsonantickou kombinaci začíná 26,04 % slabik. Konečně další možností, jak slabičnou strukturu slov popsat, je soustředit se na tzv. rým slabiky, čímž se myslí jádro + kóda. Podle rýmu lze například rozlišovat tzv. lehké a těžké slabiky (Hyman 1985). Uvedený rozdíl se v češtině uplatňuje především v časoměrném verši (Ibrahim a kol. 2013).

Těmito a dalšími aspekty slabičné struktury slov se musí zabývat jiné studie. Zvolené pohledy nám přesto poskytly dostatečné podklady pro učinění několika důležitých závěrů. Za prvé, slova v češtině nemají zcela libovolnou slabičnou strukturu v tom smyslu, že by se mohla skládat z jakékoliv posloupnosti povolených slabik. Za druhé, je možné pozorovat určité tendence či preference ve stavbě slov.

Jak data ukázala, slova určité slabičné struktury nejsou možná nebo alespoň v SSČ nedoložená. V prvé řadě nenajdeme doklad na slova obsahující více než čtyři slabiky, jejichž jádro je tvořeno jiným vokálem než krátkým. Stejně tak nejsou doložena slova s více než pěti zavřenými slabikami a konečně v SSČ není doklad na slova s více než pěti slabikami, jejichž iniciála je jiná než tvořená jedním konsonantem. Otázkou zůstává, zda zmíněné limity vyplývají jen z omezenosti našich dat nebo zda v češtině existuje horní hranice pro výskyt slabik určitého

typu, které nejsou ve slovech překročeny. Máme za to, že obě možnosti platí. Ve FLK nejsou nahrnuty různé slovní tvary gramatických slov, což zkresluje celkový obraz, poněvadž v případě skloňování a časování musíme přinejmenším očekávat nárůst zavřených slabik. Jak jsme zmínili výše, je doloženo slovo *interkontinentální* s pěti zavřenými slabikami a jeho instrumentál singuláru *interkontinentálním* obsahuje šest zavřených slabik. Podobně můžeme očekávat i nárůst slov se slabikami tvořenými nejednoduchými iniciálami.

Na druhé straně je pozoruhodné, že v našem vzorku chybí slova s více než čtyřmi nekrátkými vokály. Mezi nekrátké vokály řadíme vokály dlouhé a diftongální, ale také pro zjednodušení slabičné sonanty /r/ a /l/. Lingvisté popisující češtinu obvykle předpokládají, že česká slova mohou obsahovat libovolný počet dlouhých vokálů (např. Horálek 1985: 128–129, Ibrahim a kol. 2013: 14), avšak FLK toto tvrzení vyvrací. Slova s více než čtyřmi dlouhými vokály (a obecně slova s více než čtyřmi nekrátkými vokály) se totiž nevyskytují v celém FLK čítajícím 288 132 fonologických slov. Je sice pravda, že náš korpus obsahuje jen slova bez rozličných morfologických tvarů, nicméně je málo pravděpodobné, že bychom z nich díky flexi získali tvary, které by obsahovaly více než pět nekrátkých vokálů. V celém FLK je jen 49 slov se čtyřmi nekrátkými vokály a tato slova odpovídají buď substantivům středního rodu zakončeným na *í* (např. *zaříkávání*), nebo adjektivům zakončeným na *ý* (např. *králíkářský*). Jejich flexí však žádný nekrátký vokál nepřibude. Existenci takových slov nepotvrzuje ani předběžně zpracovaný textový korpus o téměř 390 tisících slovních tvarů (typů). To nás ospravedlňuje k tvrzení, že počet nekrátkých vokálů je v českých slovem omezen na počet čtyř.

Kromě absolutních limitů slabičné struktury českých slov lze pozorovat i jisté preference slov určité struktury. Jedna ze zjevných preferencí se týká výskytu vokálů jako slabičných jader. To se projevuje dvěma vzájemně souvisejícími způsoby. Za prvé, převážná většina slabičných jader je tvořena právě krátkými vokály (77,62 %). Dlouhé a diftongální vokály odpovídají jen 21,1 % slabičných jader. Poměr mezi nimi je tedy zhruba 4 ku 1. To znamená, že ačkoliv čeština rozlišuje mezi krátkými a dlouhými/diftongálními vokály, 73 % krátkých vokálů není v opozici s vokálem dlouhým či diftongálním. Konečně slabičné sonanty /r/ a /l/, jež bývají zmiňovány jako typický rys češtiny, jsou nejméně frekventovaným slabičným jádrem (1,28 %). V konkrétních textech může být díky skloňování a časování jejich frekvence jiná, ale dostupné popisy ukazují, že je ještě nižší (Ludvíková 1976).

Druhá preference týkající se slabičných jader se projevuje souvýskytem slabik s krátkými vokály v rámci celého slova. Naše data dokazují, že nejfrekventovanější jsou slova, která obsahují pouze slabiky s krátkým vokálem (více než třetina všech slov). Přesto rozlišíme-li jen krátké vokály na jedné straně a na straně druhé nekrátké vokály (ostatní vokály plus slabičné sonanty), pak platí, že nejvíce jsou doložena slova s právě jedním nekrátkým vokálem. Výjimkou jsou dvojslabičná slova, kde jsou nejčastější slova s žádným nekrátkým vokálem. Dále je jasně

pozorovatelná tendence ke snižujícímu se zastoupení slov se vzrůstajícím počtem nekrátkých vokálů.

Důvody pro preferenci slov s krátkými vokály mohou být jak historické, tak fonetické. Podle Sukače (2011) na češtinu v minulosti zřejmě působil podobný rytmický zákon jako na slovenštinu, byť nikoliv v takovém rozsahu a tak pravidelně. Z fonetického hlediska jsou dlouhé vokály náročnější na výslovnost (a čas) a v delších slovech může být obtížnější je rozlišit od krátkých, poněvadž se výslovnost často redukuje.

Jak již bylo řečeno, podle frekvence výskytu jsou v českých slovech častější krátké vokály než vokály dlouhé a diftongální. Stejně tak jsou častější zavřené slabiky oproti slabikám otevřeným. 69,99 % slabik je otevřených a 30,01 % slabik je zavřených. Zatímco v případě vokálů jsou nejfrekventovanější slova se všemi krátkými vokály, v případě otevřených a zavřených slabik nelze konstatovat, že by nejfrekventovanější byla slova, v nichž jsou všechny slabiky otevřené. To platí pouze u troj- a čtyřslabičných slov, kdežto u dvoj- a trojslabičných slov jsou taková slova až na druhém místě. Celkově převládají slova s jednou slabikou zavřenou: 55,48 % všech slov obsahuje jednu zavřenou slabiku a ostatní slabiky otevřené. 25,23 % slov neobsahuje žádnou zavřenou slabiku a jen 19,29 % obsahuje více než dvě zavřené slabiky.

Kromě kódy je druhým svahem iniciála. U ní nás zajímala její komplexnost, nikoliv pouze přítomnost či nepřítomnost konsonantů. Z celkového hlediska jsou nejvíce zastoupena slova, v nichž všechny slabiky začínají na jeden konsonant (tj. slova typu CV, CVCV, CVCVCV atd.). Rozlišíme-li však na jedné straně slabiky s iniciálou tvořenou jedním konsonantem a na straně druhé zbývající typy slabik, mají potom největší převahu slova, v nichž alespoň jedna a právě jedna slabika začíná buď na konsonantickou kombinaci, nebo žádnou iniciálu nemá. Necháme-li stranou slabiky bez iniciály, jichž není mnoho (4,52 %), pak v 38,65 % všech dvoj- až pětislabičných slov je právě jedna slabika začínající na konsonantickou kombinaci, zatímco všechny ostatní začínají na jeden konsonant.

Shrneme-li závěry našeho výzkumu, pak čeština preferuje slova 1) tvořená slabikami jen s krátkými vokály, 2) obsahující právě jednu zavřenou slabiku a 3) obsahující právě jednu slabiku s iniciálou tvořenou jedním konsonantem. Tato zjištění lze spojit s dalšími fakty, které se týkají slabičné struktury jako takové (viz Bičan 2015: 4) podle typu jádra jsou nejfrekventovanější slabiky s krátkými vokály (77,62 % všech slabik v SSČ), 5) podle typu kódy jsou celkově nejfrekventovanější slabiky bez kódy (69,99 % všech slabik), 6) podle typu iniciály jsou celkově nejfrekventovanější slabiky začínající na jeden konsonant (69,44 %). Konečně ze všech typů slabik je v češtině nejfrekventovanější typ CV (konsonant-vokál, např. /ta/), jemuž odpovídá 48,05 % z celkového množství 146 703 slabik. Tato zjištění však platí především pro lexikální zásobu češtiny, a proto bude nutné v budoucnu naše závěry ověřit na materiálu ze skutečných textů.

Literatura

BIČAN, Aleš
2013 *Phonotactics of Czech* (Frankfurt am Main: Peter Lang)
2014 „K pojmu fonologické slovo v češtině"; in Boček, Vít – Vykypěl, Bohumil (eds.): *Sophia Slavica* (Brno: Tribun), s. 13–23
ms. „Description of the Phonological Corpora of Czech", k dispozici zde: <http://www.ujc.cas.cz/phword>
2015 „Kvantitativní analýza slabiky v českém lexikonu"; *Linguistica Brunensia* 63/2 , s. 87–107

DIXON, R. M. W. – AIKHENVALD, Alexandra Y.
2003 „Word: A Typological Framework"; in Dixon, R. M. W. – Aikhenvald, Alexandra Y. (eds.): *Word: A Cross-linguistic Typology* (Cambridge: Cambridge University Press), s. 1–41

DURAND, Jacques a kol. (eds.)
2014 *The Oxford Handbook of Corpus Phonology* (Oxford: Oxford University Press)

FRISCH, Stefan A.
2012 „Phonotactic Patterns in Lexical Corpora"; in Cohn, Abigail a kol. (eds.): *The Oxford Handbook of Laboratory Phonology* (Oxford: Oxford University Press), s. 458–470

HÁLA, Bohuslav
1968 *Výslovnost spisovné češtiny 1* (Praha: Academia)

HORÁLEK, Karel
1986 „Fonologie spisovné češtiny"; in Petr, Jan a kol. (eds.): *Mluvnice češtiny 1* (Praha: Academia), s. 122–156

HYMAN, Larry
1985 *A Theory of Phonological Weight* (Dordrecht: Foris)

CHOMSKY, Noam – HALLE, Morris
1968 *The Sound Pattern of English* (New York: Harper & Row)

IBRAHIM, Robert a kol.
2013 *Úvod do teorie verše* (Praha: Akropolis)

JAKOBSON, Roman – HALLE, Morris
1956 *Fundamentals of Language* ('S-Gravenhage: Mouton & Co.)

KUČERA, Henry – MONROE, George K.
1968 *A Comparative Quantitative Phonology of Russian, Czech, and German* (New York: Elsevier)

LUDVÍKOVÁ, Marie
1976 „On Some Statistical Differences in Two Spoken Texts on the Syllabic Level"; in Horecký, Ján – Sgall, Petr – Těšitelová, Marie (eds.): *Prague Studies in Mathematical Linguistics 5* (Praha: Academia), s. 91–104

TĚŠITELOVÁ, Marie (ed.)
1985 *Kvantitativní charakteristiky současné češtiny* (Praha: Academia)

MARTINET, André
2011 *Éléments de linguistique générale* (Paris: Armand Colin)

MATHESIUS, Vilém
1929 „La structure phonologique du lexique du tchèque moderne"; *Travaux du Cercle Linguistique de Prague* 1, s. 67–84

MULDER, Jan
1989 *Foundations of Axiomatic Linguistics* (Berlin – New York: Mouton de Gruyter)

ROMPORTL, Milan
1978 *Výslovnost spisovné češtiny* (Praha: Academia)

ROUSSET, Isabelle
2004 *Structures syllabiques et lexicales des langues du monde* (Grenoble, Université Stendhal, Ph.D. práce)

SKARNITZL, Radek
2010 „Prague Phonetic Corpus: Status Report"; *Phonetica Pragensia* 12, s. 65–67

SUKAČ, Roman
2013 „Fish and its Fisherman. Paradigmatic and Derivative Length in Czech"; *Zeitschrift für Slawistik* 58, č. 1, s. 72–101

TRNKA, Bohumil
1966 „The Distribution of Vowel Length and its Frequency in Czech"; in Doležel, Lubomír – Sgall, Petr – Vachek, Josef (eds.): *Prague Studies in Mathematical Linguistics 1* (Praha: Academia), s. 11–16

TRUBECKOJ, Nikolaj
1939 *Grundzüge der Phonologie* (Praha: Jednota československých matematiků a fysiků)

VACHEK, Josef
1940 „Poznámky k fonologii českého lexika"; *Listy filologické* 67, č. 3–4, s. 395–402

Resumé
Phonological Lexical Corpus of Czech and the Syllabic Structure of Czech Words

The paper describes the Phonological Lexical Corpus of Czech (http://www.ujc.cas.cz/phword) and presents a sample of its analysis. The corpus is a phonologically transcribed database of lexical items from published dictionaries of Czech. Every item contains information about its length (in terms of phonemes and syllables), its syllabification, and phonological properties of the constituent phonemes. Information about the word's part of speech and its presence in various dictionaries is also included. As an example of the usefulness of the corpus, an analysis of the syllabic structure of Czech words is presented on the basis of 49 365 lexical items recorded in *Slovník spisovné češtiny*. Attention is paid to the co-occurrence of syllables according to the quality of the syllable nucleus, the presence vs. absence of the syllable coda, and the complexity of the syllable onset. It is demonstrated that there are certain tendencies in the distribution of syllables within words. First, the frequency of words decreases with the increase of non-short vowels within them. Second, words where one syllable is closed and the others are open are preferred to possible configurations. Third, Czech furthermore prefers words where exactly one syllable onset is complex and the others are simple.

Keywords: Phonological Lexical Corpus of Czech; syllabic structure of Czech words; distribution of syllables in Czech

PhDr. Aleš Bičan, Ph.D.
Ústav pro jazyk český AV ČR, v. v. i.
Etymologické oddělení
Veveří 97
602 00 Brno
bican@phil.muni.cz