Nástroj pro slovotvornou analýzu jazykového korpusu[1]

Václav Cvrček – Pavel Vondřička
ÚČNK FF UK
vaclav.cvrcek@ff.cuni.cz – pavel.vondricka@ff.cuni.cz

ABSTRACT: Abstract: In this paper we would like to present a new software tool for corpus-based study of word formation in corpora of written Czech called Morfio (see http://morfio.korpus.cz). Its purpose is to allow linguists to search the corpus by series of parallel queries which specify chosen derivational model. It also analyses obtained results for morphological productivity of affixes and estimates the completeness of the derivational model.

1 Úvod

Česká slovotvorba se od zakladatelského počinu M. Dokulila (1962, 1967) věnovala převážně onomaziologickému pohledu na problematiku tvoření slov v češtině. Práce, které na jeho pionýrské úsilí navazují, se s větším či menším úspěchem snaží vyrovnat s problémem, který představuje sémantická klasifikace jako východisko pro popis jiných slovních druhů než substantiv (Karlík et al., 1995; Čermák, 2011; Cvrček et al., 2010). Přitom zjevným limitem těchto od významu vycházejících přístupů je rozsah popsaného materiálu.

V době rozsáhlých a reprezentativních korpusů je materiálová základna, kterou poskytují, jistě vhodným začátkem slovotvorného zkoumání. Problém vyvstává ovšem s tím, že korpusy neposkytují potřebnou sémantickou informaci o jednotkách (nebo ne v takové míře a s takovou spolehlivostí), abychom mohli slovotvornou analýzu založit na jiném než sémaziologickém pohledu.

Pro účely takto chápaného popisu tvoření slov byla vytvořena aplikace **Morfio**, která slouží k odhadování rozsahu a produktivity slovotvorných modelů v češtině na základě korpusových dat. Obecně je v rámci sémaziologického přístupu každý slovotvorný vztah – vedle složky sémantické, kterou lze jen obtížně automaticky uchopit - vytvářen 1) formální shodou/podobností v určitých částech slova, tzv. báze (např. *dřev-* je část společná pro slova *dřevo* i *dřevěný*) a 2) formálními odlišnostmi v částech specifických, tzv. formantech (morfy *-o* a *-ěný* v předchozím příkladu). Cílem aplikace je najít všechny dvojice, resp. trojice nebo čtveřice, jednotek v korpusu, které se shodují v bázi a liší se pouze specifikovanými formanty.

Semaziologický přístup má samozřejmě také svoje slabiny, problémy nastávají např. při zpracování homonymie; jejichž řešení přesahuje možnosti takto koncipovaného nástroje. Výstupem aplikace Morfio proto není a nemůže být bezchybný a bez jakýchkoli úprav, revize a lingvistické manipulace publikovatelný výstup, spíš se jedná o pomůcku, která množství dat dokáže pro lingvistické účely předzpracovat tak, aby analýza byla rychlejší, výtěžnější a celkově pro badatele jednodušší. Stejně jako u jiných korpusových vyhledávačů je tedy cílem pouze snadné dosažení 100% úspěšnosti hledání daného typu (*recall*) a přehledné setřídění výsledků, zatímco jejich relevantnost (*precision*) je zcela ponechána na úsudku uživatele: tj. samotné formulaci dotazu a následném vyhodnocení nálezů.

[1] Tento článek vznikl při realizaci projektu Český národní korpus (LM2011023) financovaného Ministerstvem školství, mládeže a tělovýchovy v rámci aktivity Projekty velkých infrastruktur pro VaVaI.

2 Technická stránka aplikace Morfio

Stejně jako jiné moderní aplikace pro práci s korpusem i Morfio je nástrojem, ke kterému uživatel přistupuje pomocí webového prohlížeče. Dostupný je na adrese http://morfio.korpus.cz nebo ze stránek http://www.korpus.cz. Výhodami webových aplikaci je především to, že přesunují všechny náročné výpočty na výkonný server, zatímco s výsledky může uživatel pracovat v pohodlí svého počítače bez ohledu na operační systém a bez nutnosti instalace jakéhokoli software. Výsledky, k nimž uživatel dospěje během práce, jsou kdykoli dostupné pro kohokoli pomocí odkazu ve spodní části zadávacího formuláře. Pomocí tohoto odkazu, který je tvořen unikátním klíčem, je možné výsledky též citovat v odborných pracích nebo sdílet s jinými uživateli.

Pro účely aplikace Morfio jsou data z korpusů ČNK získávána pomocí programu Manatee. Výsledky, které poskytuje, zpracovává série skriptů v jazyce Perl. Uživatelské rozhraní je pak generováno PHP skripty a využívá technologie JavaScriptu pro zlepšení přehlednosti prezentace výsledků (knihovny jQuery a D3js)

3 Zadání dotazu

V rámci aplikace Morfio je slovotvorný model, jehož produktivitu chceme s jeho pomocí odhadnout a demonstrovat, chápán jako vztah mezi dvěma nebo více skupinami jednotek (tradičně: slova základová a odvozená). Vzhledem k tomu, že analýza je založena na formální podobnosti, odhaluje vztahy i tam, kde by tradiční slovotvorná teorie viděla vztah komplikovanější (např. zahrnutím třetí jednotky, s jejíž pomocí je první slovo od druhého odvozeno). Procesuální charakter slovotvorby (někdy hraničící až s etymologií) je v tomto přístupu nahrazen přístupem striktně synchronním a relačním; slovotvorba je ve světle těchto premis chápána jako hledání a analyzování pravidelně se opakujících languových vztahů mezi jednotkami lexikonu, které se realizují formální odlišností, jíž odpovídá ustálený lexikálně-sémantický rozdíl (jde tedy o nacházení takových formálních změn, které mají pravidelně se opakující sémantický korelát).

Pro ilustraci si můžeme uvést jednoduchý příklad. Představme si, že nás zajímá rozsah a produktivita slovotvorného modelu reprezentovaného dvojicí slov lovit a úlovek. Nejprve identifikujeme společné části (báze, B) a odlišné části (formanty, F).

	F1	B	F2
1. vzor		lov	it
2. vzor	ú	lov	ek

Každou ze skupin slov, která do modelu vstupuje, musí uživatel definovat jako jeden vzor. Zadání vzoru může využít téměř všechny možnosti, které poskytují regulární výrazy a dotazovací jazyk systému Manatee/SketchEngine. Skupiny tvarů nebo lemmat vymezené jednotlivými vzory jsou pak aplikací porovnávány a podle uživatelovy specifikace jsou mezi nimi odhaleny formální podobnosti indikující slovotvorné vztahy.

Každý vzor je sekvencí částí společných všem vzorům a částí pro jednotlivé vzory specifických. Části všem vzorům společné (vyznačeny ve formuláři žlutou podkladovou barvou) jsou analogické ke slovotvorné bázi, části odlišné pak plní roli formantů. Zadávací

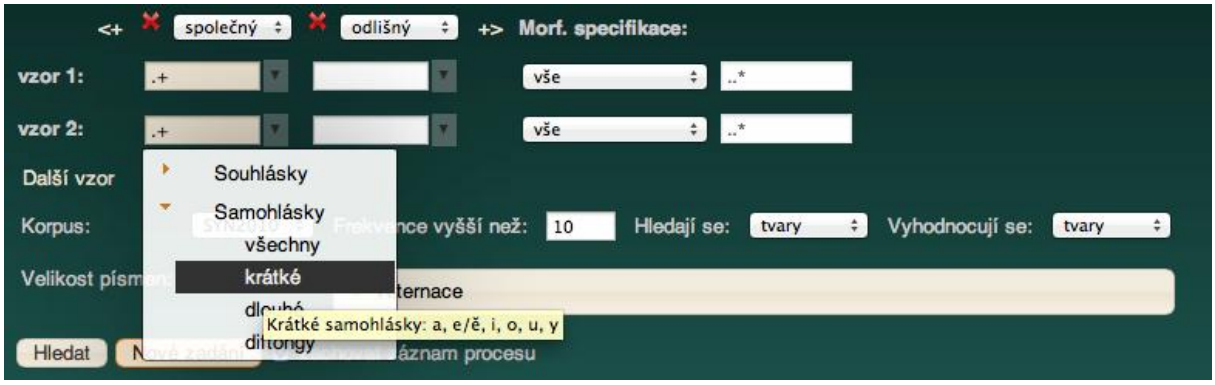
formulář automaticky zajišťuje, aby všechny společné části měly identické zadání, a naopak kontroluje, aby odlišné části nebyly stejné. Součástí specifikace vzorů může být i jejich morfologická charakteristika na úrovni slovního druhu (viz menu vpravo od vstupních polí) nebo i podrobnější (konkrétní morfologická informace zapsaná pomocí tagu).

Pomocí značek <+ a +> může uživatel přidávat další části (sloupce) ke vzorům, ikona křížku slouží k jejich smazání. Roletové menu nad jednotlivými částmi slouží k specifikaci toho, zda se jedná o část společnou/bázovou nebo odlišnou/formantovou. Pomocí těchto nástrojů je možné snadno vytvořit libovolnou sekvenci společných a odlišných částí pro všechny vzory. Odkaz „Další vzor“ umožňuje přidat další řádek do zadávání.

Do všech polí je možné zapisovat jak konkrétní znaky, tak regulární výrazy se speciálním významem umožňující vyhledávat podle obecného vzorce. Pro zjištění rozsahu a produktivity slovotvorného modelu je vhodné explicitně specifikovat formanty (tedy části odlišné) a pomocí nespecifických regulárních výrazů (např. .* pro sekvenci libovolných znaků nebo .+ pro sekvenci jednoho a více libovolných znaků) části společné. Zobecníme-li tak příklad uvedený výše, dostaneme modelový zápis obou vzorů.

	F1	B	F2
1. vzor		.+	it
2. vzor	ú	.+	ek

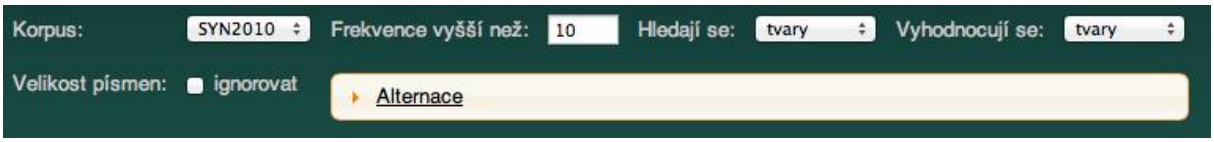
Vedle standardních regulárních výrazů nabízí aplikace pro účely slovotvorného výzkumu i předpřipravenou sadu fonémových skupin. Vedle každého zadávacího pole je pod ikonou trojúhelníku dostupné roletové menu, které umožňuje do dotazu vložit i některé relevantní skupiny hlásek.



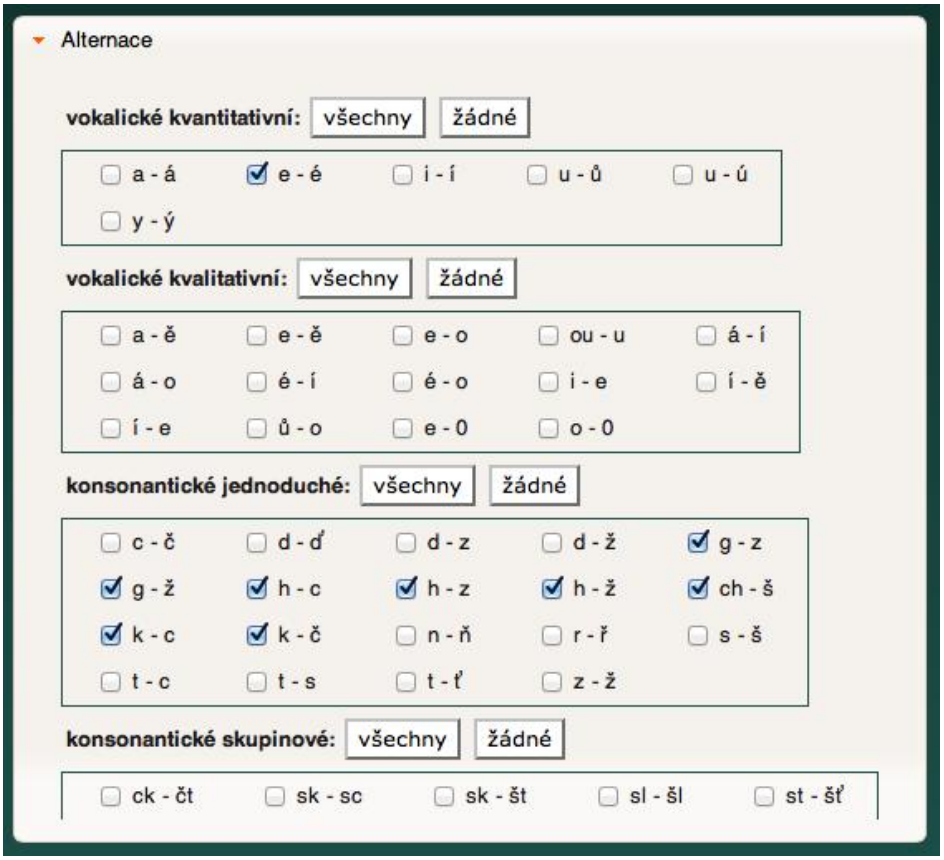
Při přejetí myši přes položku v menu se zobrazí stručná nápověda, po vybrání některé ze skupin se v příslušné části dotazu objeví zkratka, např. *[:Alveol:]*, která je při vyhodnocování dotazu nahrazena písmeny označujícími všechny české alveoláry.

Regulární výrazy reprezentující skupiny hlásek, které jsou v menu dostupné, byly vytvářeny s ohledem na možnou změnu fonologické hodnoty písmena ve specifickém kontextu. Zvolí-li uživatel např. palatální hlásky, bude skupina reprezentovat jak písmena *ď, ť, ň, j,* tak i *d, t, n* stojící před *i, í* nebo *ě*.

Pod specifikací vzorů je možné nastavit obecné parametry dotazu, které určují v jakých korpusech a na jakých typech jednotek (word/lemma) bude průzkum prováděn. Vedle toho může uživatel omezit prohledávání pouze na jednotky s určitou minimální frekvencí, příp. může ovlivnit i to, zda má být analýza citlivá na velikost písmen. Platí přitom, že čím větší objem dat je nutné prohledávat (větší korpus, menší minimální frekvence, slovní tvary namísto lemmat, s rozlišováním velikostí písmen), tím pomalejší analýza bude.



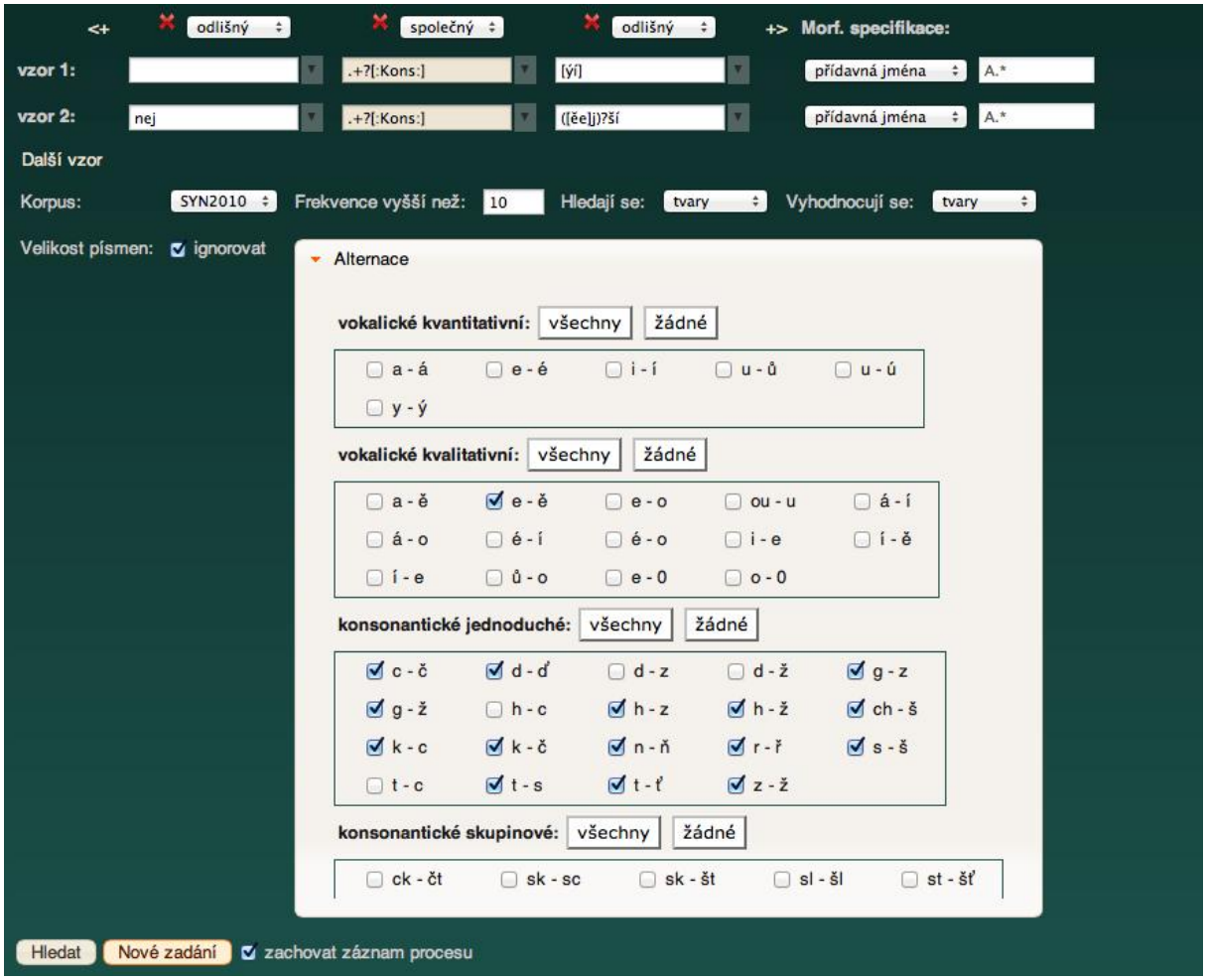
Další možnosti nastavení analýzy se skrývají v rozbalovací sekci Alternace. Po kliknutí na ni je možné z nabídky vybrat, jaké typy hláskových změn se můžou objevit ve sledovaném modelu.



Alternace umožňují vyhledávat i takové dvojice, kde shoda v bázi není stoprocentní. Např. u dvojice slov *moucha - muška*, dochází ke dvěma hláskovým obměnám: *ou* > *u* a *ch* > *š*. Pokud uživatel tyto alternace ve výběru nezvolí, nebude tato dvojice slov (a dvojice strukturně podobné) do výsledku zahrnuta, protože u ní nebude identifikována přesná formální shoda.

Alternace se aplikují pouze na společné části (báze), nikoli na formanty, a navíc pouze v případě, že vzor je definován pomocí nespecifických zástupných znaků (tečka s hvězdičkou či znaménkem plus). V případě, že společná část je definována např. výrazem *.+ch*, nebude u takto explicitně uvedených znaků (zde *ch*) provedena.

Příkladem využití jednotlivých možností aplikace může být hledání dvojic adjektiv, která se objevují jak v pozitivu, tak v superlativu. Kompletní zadání je patrné z následujícího obrázku.



- Vzor 1 specifikuje, že hledáme adjektiva, která nemají předponu a končí na *í/ý* (tvar pozitivu) a na konci báze, těsně před koncovým formantem, se nachází libovolný konsonant.

- Vzor 2 je rozšířený o předponu *nej-* a měl by tak zachycovat všechny superlativy končící na *-ší*, příp. *-ejší* (dotaz tedy ignoruje superlativy končící na *-čí*).

- Vzhledem k tomu, že standardně jsou kvantifikátory (*, +) v regulárních výrazech „hladové“, musíme do společné (tj. prostřední) části obou vzorů vložit symbol "?", který zajistí, že znaky *-ej-* v delší verzi formantu *-ejší* nebudou interpretovány jako součást báze.

- Analýzu je vzhledem k lemmatizaci korpusů řady SYN třeba provádět na slovních

tvarech (komparativ i superlativ jsou lemmatizovány dohromady pod pozitiv).

- Zároveň je třeba zapnout některé alternace, které zajistí, že pravidelné hláskové změny na švu mezi bází a formantem budou zohledněny.

Jak již bylo zmíněno, zadání a výsledky každého dotazu jsou uloženy na serveru pro případné další užití. Jejich opětovné vyvolání je možné za použití odkazu, který se po odeslání dotazu objeví v dolním pravém rohu formuláře. Tento odkaz je tak možné využít při citování výsledků v literatuře.

4 Výsledky

Po zadání specifikace jednotlivých vzorů se výsledky analýzy objeví uspořádány do jednotlivých záložek podle typu informace, které obsahují. Pro ilustraci možností nástroje uvedeme popis jednotlivých typů informací, které zprostředkovávají.

4.1 Výpis

V tabulce uvedené v záložce „Výpis“ jsou všechny doklady ze všech vzorů, které vstupují do zadaného modelu (každý sloupec pro jeden vzor). Červená část slov označuje společnou bázi (ta se může lišit pouze v případě aplikace alternací). V závorkách uvedený údaj představuje celkovou frekvenci jednotky ve zvoleném korpusu. Tabulku je možné přetřídit podle libovolného sloupce a to jak abecedně, tak frekvenčně pomocí šipek v záhlaví tabulky. Každé slovo zároveň funguje jako odkaz směřující k ukázce konkordancí ve zvoleném korpusu.

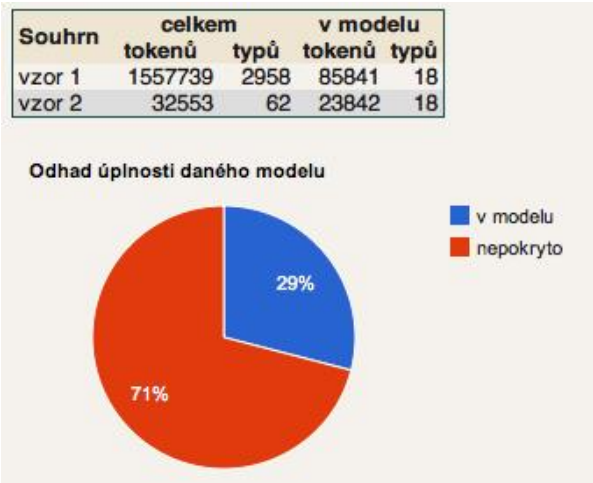
Páry vytvořené až díky aplikaci alternačních pravidel jsou zvýrazněné barevným pozadím. Jejich báze se proto budou lišit. V případě, že použitím alternačních pravidel dojde k situaci, že jednomu vzoru v dané dvojci odpovídá více slov, jsou všechna tato slova uvedena hromadně v jednom řádku tabulky.

4.2 Souhrn

Záložka souhrn prezentuje základní frekvenční charakteristiky jednotlivých vzorů i celého modelu (zejména pak jeho úplnost). Jsou zde uvedeny počty typů s nadlimitní frekvencí a součet jejich výskytů. Jedna sada údajů (sloupec „celkem“) se vždy týká vzoru samotného (chápaného izolovaně), druhá sada (sloupec „v modelu“) pak odkazuje k těm jednotkám příslušejícím ke vzoru, které zároveň patří do analyzovaného slovotvorného modelu, tj. slova, která mají k sobě odvozeninu identifikovanou v rámci druhého vzoru.

Odhad úplnosti daného modelu, resp. pokrytí slovotvorného problému zadaným modelem (jak ho znázorňuje graf), je vytvořen na základě následujících předpokladů:

1. Každému vzoru odpovídá určitý počet jednotek v korpusu (slovních tvarů nebo lemmat). Předpokládáme přitom, že většina slovotvorných vztahů je asymetrická – je-li určitá skupina slov odvozena od jiné skupiny (základových) slov, pak slov odvozených bude míň než slov základových (ne od každého základového slova vznikne odvozenina). Vzory tedy můžeme rozdělit na ty, které jsou základové (tj. ty, které zahrnují větší počet jednotek v korpusu) a vzory podmíněné (které obsahují menší počet jednotek). Jinými slovy, ten vzor, který vyděluje menší skupinu slov, je odvozen od vzoru druhého, je tedy ve vztahu ke slovům druhého vzoru vzorem podmíněným.



2. Úplnost modelu se pak počítá jako podíl, který tvoří celkový počet dvojic (příp. trojic) slov v modelu z celkového počtu typů vzoru, který je podmíněn (odvozen – tedy který vyděluje menší skupinu slov).

Příklad: Při zkoumání modelu, který je možné charakterizovat dvojicí slov *pracovat* – *pracovník*, resp. dvojicí formantů *-ovat* a *-ovník*, můžeme identifikovat 18 dvojic lemmat (s frekvencí vyšší než 10), která vstupují do tohoto slovotvorného vztahu (*bojovat - bojovník, hodovat - hodovník* atp.). Přitom samotné zadání prvního vzoru (*-ovat*) poskytuje 2958 různých slovesných lemmat, samotný druhý vzor (*-ovník*) pak 62 různých substantivních lemmat. Sloves zakončených formantem *-ovat* je tedy výrazně víc než substantiv s formantem *-ovník*. Každý ze vzorů přitom obsahuje jednotky, které do modelu nevstupují (např. sloveso *abdikovat* do modelu nevstupuje kvůli neexistenci substantiva **abdikovník*, k substantivu *činovník* nenacházíme sloveso **činovat* atp.). Vzor s menším počtem nalezených jednotek (v našem případě *-ovník*) představuje pro celý model větší omezení než vzor identifikující jednotek víc (v našem případě *-ovat*). Podmínkou pro existenci slovotvorného vztahu specifikovaného ukázkovým modelem je tedy existence slovesa *-ovat*, tvoření je ovšem limitováno především počtem substantiv typu *-ovník*, neboli vzor *-ovník* je podmíněn existencí vzoru *-ovat*. V takovém případě je smysluplné odhadovat úplnost navrženého modelu podle toho, jak moc se na něm podílí jednotky toho vzoru, který je podmíněn. V tomto případě se tedy odhad počítá jako poměr typů vzoru *-ovník*, které do modelu vstupují, ku všem typům tohoto vzoru, tedy 18/62, což odpovídá 29 %.

Pro slovotvornou analýzu je samozřejmě optimální, pokud pokrytí modelem je blízké 100 %. Znamená to, že pro všechny jednotky v podmíněném vzoru jsme v druhém vzoru identifikovali základové jednotky. Pokud takového pokrytí není možné dosáhnout, znamená to, že slovotvorný model vysvětluje pouze část jednotek formálně vymezených pomocí podmíněného vzoru; pro usouvztažnění nepokrytých jednotek se svými základy je třeba stávající model upravit nebo vytvořit další, komplementární model.

4.3 Vzor 1, vzor 2,...

V záložkách nadepsaných pouze označením vzoru jsou prezentovány výsledky analýzy jednotlivých vzorů jako samostatných dotazů, a to ve formě tabulky jednotek (slovních tvarů

nebo lemmat) spolu s jejich frekvencemi ve zvoleném korpusu. Tabulku je možné doplnit i o jednotky, které v modelu nebyly brány v potaz, protože jejich frekvence byla nižší než uživatelem stanovený limit. Údaje zvýrazněné barevným pozadím se účastní slovotvorného modelu (tj. existuje k nim v druhém vzoru protějšek se stejnou bází, lišící se pouze formanty).

Tabulky vyhodnocení jednotlivých vzorů slouží především ke korekci zadaného modelu. Je-li ve výsledku dotazu slovo, které by mělo být součástí modelu a přitom zahrnuto nebylo (není barevně zvýrazněno), je na místě specifikaci modelu pozměnit tak, aby byl odhad úplnosti a produktivity co nejpřesnější.

Tabulku je možné třídit vzestupně i sestupně pomocí značek v záhlaví tabulky. Vedle řazení podle frekvence, je možné využívat i třídění abecední, a to jak klasické, tak retrográdní (tedy řazení od konce slova). Pro lepší orientaci v abecedně setříděných datech je možné zapnout shlukování do skupin (viz odkaz nad záhlavím tabulky). Řádky se v takovém případě seskupují podle stejné počáteční nebo koncové sekvence znaků. Počet stupňů seskupení je možné regulovat pomocí prvku +/- (každé další písmeno od začátku či konce, kterým se slova liší, může tvořit další (pod)stupeň pro rozdělení skupin). U každé skupiny se přitom objevují údaje o počtu typů a tokenů dané skupiny, a to jak celkově, tak těch, které se účastní slovotvorného modelu (uvedeno v závorkách).

4.4 Produktivita

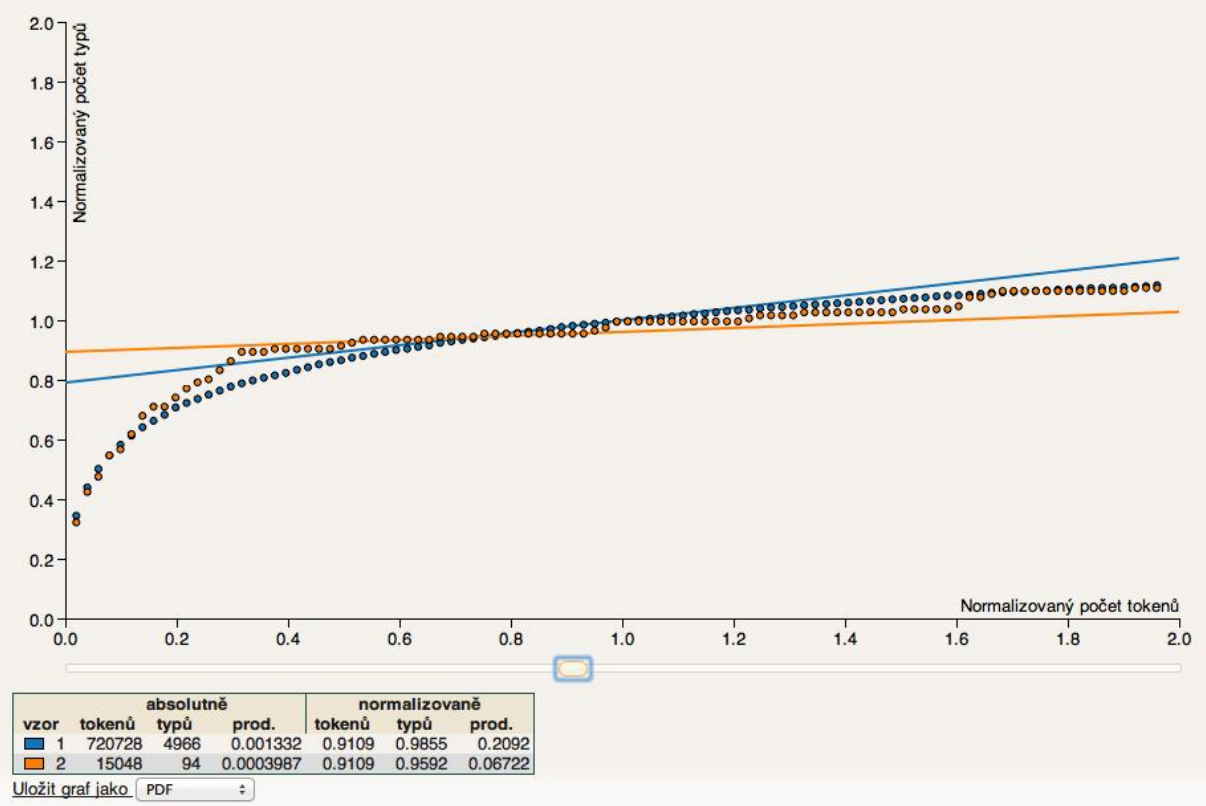
Odhad produktivity obou vzorů, který je k dispozici v poslední popisované záložce, a jejich vzájemné porovnání vychází z teoretických poznámek H. Baayena (1992). Morfologická produktivita se zde měří pomocí odhadu tendence přírůstku nových typů při přírůstku dokladů (tokenů) pro každý vzor samostatně. Ze srovnání pak vyplývá, který vzor je produktivní, protože počet jeho typů roste rychleji, s jeho formanty se pojí nové a nové báze, a který vzor je naopak neproduktivní a potenciálně uzavřený (i když třeba frekventovaný a rozsáhlý).

Zde je třeba podotknout, že chápání produktivity v současné teorii není ustálené. Na základě následujících úvah jsme vyloučili, že bychom morfologickou produktivitu zkoumaného formantu jednoduše ztotožnili s počtem tokenů, typů nebo směrem derivovanosti. Produktivitu nelze chápat jako počet výskytů slov (tokenů) s daným formantem zejména proto, že i velmi frekventované formanty můžou existovat pouze v rámci uzavřené, a tedy neproduktivní, třídy. Stejně tak není produktivita totožná s počtem různých slov (typů), na jejichž stavbě se formant podílí – v jazyce můžou existovat velmi rozsáhlé, avšak uzavřené slovotvorné třídy a vedle nich relativně malé skupiny (různých) slov s jedním formantem, jejichž přírůstek je potenciálně stále neomezen. Produktivita není vázána ani na směr derivace (odvozené slovo nemusí být méně produktivní než jednotka fundující). Skupina základových slov může být velmi rozsáhlá a přesto omezená, zatímco skupina slov od nich odvozených zdaleka ještě nerealizovala všechen potenciál k tvoření, nicméně vykazuje silnou tendenci (silnější než u slov základových) k dalšímu rozšiřování.

Odhad produktivity prezentovaný nástrojem Morfio je tedy založen na premise, že produktivitu bychom mohli definovat jako ochotu, s níž formant vstupuje do nových kombinací a účastní se tak na stavbě dalších lexémů. Takto vyjádřenou produktivitu bychom mohli ztotožnit s pravděpodobností, že po prozkoumání všech dokladů v rozsáhlém korpusu, budeme v okolním jazyce nacházet další nové typy.

Zkoumáme-li např. substantiva s příponou *-ost* (*společnost, možnost, činnost* apod.), bude pro nás zajímavé, zda budeme nacházet nové a nové doklady tohoto zakončení, když

prozkoumáme různě velkou část korpusu, ve srovnání např. se substantivy končícími na *-ín* (*odstín, kravín, modřín* apod.).



	absolutně			normalizovaně		
vzor	tokenů	typů	prod.	tokenů	typů	prod.
1	720728	4966	0.001332	0.9109	0.9855	0.2092
2	15048	94	0.0003987	0.9109	0.9592	0.06722

Uložit graf jako PDF

Produktivitu ve světle tohoto přístupu můžeme chápat jako sumární pravděpodobnost všech typů daného vzoru, které nejsou reprezentovány v korpusu. Je-li taková pravděpodobnost u nějakého vzoru po prozkoumání určitého počtu dokladů velká, znamená to, že vzor je produktivní, je-li naopak po prozkoumání stejně velkého počtu dokladů malá, znamená to, že daný vzor je relativně uzavřený. Sumární pravděpodobnost v korpusu nereprezentovaných typů daného vzoru lze obecně vypočítat pomocí Good-Turingova odhadu (Baayen, 2001: 57), jako počet hapaxů k celkovému počtu tokenů. V našem případě jsou hapaxy míněny ty typy, které se v daném vzoru vyskytnou právě jednou. Vyneseme-li údaje o tokenech a typech pro daný vzor do grafu, bude z podstaty konstrukce Good-Turingova odhadu tato sumární pravděpodobnost směrnicí jeho tečny v posledním bodě.

Pro porovnání vzorů, které mají nestejnou velikost, je ovšem třeba data o typech a tokenech normalizovat. Výsledky, které jsou uvedené v grafu, jsou tak normalizovány jak podle osy x, tak podle osy y, a to pro hodnotu mediánu tokenů, resp. typů. Znamená to, že hodnotu 1 pro normalizovaný počet tokenů na ose x představuje medián tokenů daného vzoru, a obdobně hodnota 1 pro normalizovaný počet typů na ose y koresponduje s prostřední hodnotou počtu různých slov (při různě velkém počtu tokenů).

Abychom předešli možnému zkreslení, které by bylo dáno tím, že texty jsou v korpusech uspořádány do žánrových a tématických celků, jsou konkordance sloužící pro odhad produktivity každého vzoru nejprve několikrát randomizovány. Počet náhodných permutací konkordančních řádků je proměnlivý a pohybuje se od jednoho promíchání po maximálně deset opakovaných znáhodňovacích cyklů.

5 Závěr

Závěrem je třeba připomenout, že ani rozsáhlá data, jakkoli vhodně zpracovaná, nezbavují badatele povinnosti výsledky ručně vyhodnotit a interpretovat. Korpusově založená slovotvorba, k jejímuž rozvoji by nástroj Morfio měl přispět, bude ovšem minimálně do doby, než bude k dispozici spolehlivé a systematické sémantické značkování, muset vycházet od formy a sémantickou interpretaci slovotvorných vztahů (ve smyslu onomaziologických kategorií) provádět až na základě zobecnění velkého množství pozorování.

Pomůcka Morfio byla koncipována mj. i za tím účelem, aby byla nápomocna při detailnější specifikaci a testování modelů. Zejména možnost retrográdního třídění se seskupováním podle sekvence znaků může přispět detailnější a přitom pohodlnější analýze formální stránky odvozování.

Určitým dluhem, který bude třeba řešit v budoucnu specifickým nástrojem, zůstává otázka kompozit. Jejich formální analýza a automatická identifikace v korpusech je stále velkým problémem. K jejímu vyřešení bude třeba data obohatit o morfematickou informaci, která by podávala informaci o existenci několika kořenů v rámci jednotky.

LITERATURA

BAAYEN, H. (1992): Quantitative aspects of morphological productivity. In G. E. Booij and J. van Marle (eds), Yearbook of Morphology 1991, Kluwer Academic Publishers, Dordrecht, 109-149.

BAAYEN, H. (2001): Word Frequency Distributions, Kluwer Academic Publishers, Dordrecht, The Netherlands.

CVRČEK et al. (2010): Mluvnice současné češtiny. Karolinum. Praha.

CVRČEK, V. - VONDŘIČKA, P. (2012) Morfio. přístupný na adrese *http://morfio.korpus.cz*

ČERMÁK, F. (2011): Morfématika a slovotvorba češtiny. NLN. Praha.

DOKULIL, M. (1962): Tvoření slov v češtině 1. ČSAV. Praha.

DANEŠE, F. - DOKULIL, M. - KUCHAŘ, J. (1967): Tvoření slov v češtině 2. ČSAV. Praha.

KARLÍK, P. et al. (1995): Příruční mluvnice češtiny. NLN. Praha.