Český sufix *-ák* – případová studie

Dana Hlaváčková – Karel Pala
Centrum zpracování přirozeného jazyka, Fakulta informatiky, Masarykova univerzita
hlavack@fi.muni.cz – pala@fi.muni.cz

ABSTRACT: New techniques in Czech derivational morphology are discussed in the paper. They are based on the use of the tool Deriv with integrated access to the main Czech dictionaries and corpora (SYN2000c and the new large Czech corpus CzTenTen12). The case study deals especially with the Czech suffix *-ák* – we describe its behaviour as completely as possible. The paper brings some new results in comparison with standard Czech grammars which, as a rule, do not rely on large data and software tools. The semantics of the suffix *-ák* has been explored thoroughly and 10 categories have been recognized. The most frequent nouns with the suffix *-ák* are agentive ones and the nouns denoting bearers of properties. The obtained results can be found in Table 1.

1 Úvod

V článku věnujeme pozornost derivačnímu chování českého sufixu *-ák.* Pracujeme s nástrojem Deriv (dostupný po registraci na http://deb.fi.muni.cz/deriv), který umožňuje pracovat pokud možno se všemi českými substantivy se sufixem *-ák*, jež lze získat standardními derivačními procesy.

Je třeba říci, že zpracování tohoto i dalších sufixů ve standardních českých gramatikách je založeno na omezených datech, jejich autoři nepracovali s korpusy a neměli k dispozici velké strojové slovníky kmenů. Z toho nutně plyne, že jejich výsledky musí být jen parciální. Totéž platí i pro Dokulilovy relevantní výsledky (Dokulil, 1962), jimiž položil teoretické základy české slovotvorby, i když musel vycházet z omezených dat.

2 Motivace

Naším hlavním cílem je podívat se detailněji a hlouběji na derivační vztahy s použitím větších a úplnějších českých dat. Usilujeme o lepší pokrytí zkoumaných jevů spolu s vyšší přesností, což jsou parametry, s nimiž se ve standardních gramatikách vůbec nepracuje, a otázky orientované tímto směrem se v nich vůbec nekladou.

Vycházíme z výsledků získaných v oblasti české formální morfologie, zejména z českého morfologického slovníku, jenž je součástí českého morfologického analyzátoru (Šmerk, 2010). Na ně navazuje zkoumání derivačních vztahů a konkrétně chování zmíněného sufixu *-ák.* Pracujeme se softwarovými nástroji, proto náš popis musí být co nejformálnější. Jsme přesvědčeni, že hlubší poznání derivačního chování jednotlivých sufixů (a přirozeně i prefixů) je nutným předpokladem pro vývoj chytrých vyhledávacích nástrojů použitelných v různých webových aplikacích.

3 Počítačové zpracování české slovotvorby

Hlubší poznání českých slovotvorných vztahů vyžaduje rozsáhlejší data, než se zatím užívala v českých standardních gramatikách (MČ I, Petr et al, 1986, PMČ, Karlík et al., 1995). Taková data se ovšem nedají dost dobře zpracovávat manuálně a pokud ano, bylo by to časově velmi náročné a zpracování by se nevyhnulo chybám. V současnosti máme k dispozici velký strojový slovník českých kmenů čítající cca 400 000 položek, jehož pokrytí představuje

cca 95 % současné češtiny. Nástrojem, jímž zkoumáme chování sufixu *-ák* a dalších, je vyhledávač Deriv (Šmerk, 2010a), který umožňuje na základě formálního popisu morfematické segmentace sledovat slovotvorné vztahy. Toho se dosahuje jednoduchým vyhledáváním možných kombinací kmenů a sufixů (a podobně i prefixů) a také využitím regulárních výrazů dovolujících zachycovat také alternace v kmenech nebo na morfematických švech.

Tímto způsobem získáváme nejprve potřebný seznam substantiv obsahujících konkrétní české sufixy. Nástroj Deriv v sobě integruje přístup ke dvěma českým korpusům, a to SYN2000c (SYN2000) a CzTenTen12 (dostupný po registraci na http://ske.fi.muni.cz) a také slovníkový prohlížeč DEBDict, takže uživatel (lingvista) může pozorovat chování zkoumaného sufixu co nejkompletněji a sledovat zejména frekvenční údaje z uvedených korpusů.

4 Výchozí data – morfologický slovník, korpusy

Jak jsme se již zmínili, pracujeme s velkým strojovým slovníkem českých kmenů čítajícím cca 400 000 položek, pokrývajícím 95 % české slovní zásoby v psaných synchronních korpusech (pro srovnání uveďme, že rozsah SSJČ je asi 192 000 hesel). Je integrální součástí českého morfologického analyzátoru Ajka (Šmerk, 2010) a nástroj Deriv jej využívá k vytváření seznamů českých substantiv (sloves, adjektiv, …) se zadanými vlastnostmi. Zbývajících 5 % slovní zásoby zahrnuje výrazy jako e-mailové a webové adresy, telefonní čísla, časové údaje, různé typy zkratek a také výrazy pocházející z cizích jazyků, nejčastěji angličtiny a slovenštiny. Toto pokrytí znamená, že analyzátor Ajka dovede pracovat s prakticky libovolným českým textem a rozpoznávat v něm veškeré slovní tvary.

Protože Deriv je propojen se dvěma českými korpusy a šesti hlavními českými slovníky prostřednictvím prohlížeče DEBDict, získané výsledky lze vzájemně dobře porovnávat, zejména s ohledem na frekvence výskytů. To platí zejména pro porovnání čísel získaných z korpusu SYN2000c čítajícím 114 363 813 tokenů a korpusu CzTenTen12 s 5 414 437 666 tokeny, jež jasně prokazují potřebu mít data co největší. Díky vazbě Derivu na Word Sketch Engine lze získávat též kolokační data. Naše data také přirozeně obsahují příslušné seznamy českých sufixů (a prefixů).

5 Případová studie – český sufix *-ák*

Nástroj Deriv umožňuje vygenerovat seznamy lemmat vymezených morfologickými charakteristikami a zakončených zvoleným řetězcem znaků (sufixem). V případě této studie byly získány dva seznamy lemmat zakončených řetězcem *-ák*. Jeden pro substantiva rodu mužského životného (morfologická značka k1gMnSc1), druhý pro substantiva rodu mužského neživotného (morfologická značka k1gInSc1). Počet českých substantiv končících řetězcem *-ák* činí 1351, z nich 724 jsou mužská životná a 627 mužská neživotná (včetně substandardních jmen nevyskytujících se v SSJČ a vlastních jmen). U sledovaných substantiv jde převážně o názvy expresivní, slangové a dnes již zastaralé, na druhou stranu je sufix *-ák* velmi produktivní v procesu univerbizace. Pomocí nástroje Deriv je možné vidět a srovnávat frekvence jednotlivých slov v korpusech SYN2000c a CzTenTen12.

6 Slovotvorné kategorie a jejich klasifikace

Vzhledem k tomu, že dále chceme pracovat pouze se substantivy odvozenými od českých základových slov, bylo nutné ze získaných seznamů odstranit lemmata, která tento požadavek nesplňují.

V případě substantiv rodu mužského životného šlo o (celkem 200 lemmat):

- substantiva neodvozená, např. *čmelák, luňák, pašák, pták, slimák, žák*;
- vlastní jména, např. *Barták, Čermák, Dvořák, Horák, Plzák, Rybák*;
- názvy nářeční, terminologické a neterminologické názvy zvířat a rostlin, např. *doupňák, jespák, krchňák, strapák*;
- substantiva odvozená od cizojazyčných základů (němčina, ruština, francouzština), např. *ajznboňák, burlák, funebrák, landšturmák, saperák, valcverák*;
- substantiva odvozená od problematického základového slova, např. *bubák, mulisák, tumák, ťuťmák*.

V případě substantiv rodu mužského neživotného (celkem 138 lemmat):

- substantiva neodvozená, např. *hák, lák, maják, mák, tabák*;
- vlastní jména, pouze *Šišák, Žebrák*;
- názvy nářeční, terminologické a neterminologické názvy zvířat a rostlin, např. *brymburák, čerpák, brutnák, kukmák, portulák, sevlák*;
- substantiva odvozená od cizojazyčných základů (němčina), např. *execírák, hapták, maršbaťák, méblák*;
- substantiva odvozená od problematického základového slova, např. *čvaňhák, raťafák, vančák, žehrovák*.

Pro zbývající (odvozená) substantiva v seznamech navrhujeme klasifikaci obsahující následující kategorie, přičemž vycházíme z klasifikací, jež můžeme najít v MČ 1 a PMČ. Do jednotlivých kategorií byla lemmata zařazována manuálně.

Substantiva odvozená ze substantiv:

- jména konatelská, např. *dudák*, *koňák, sedlák, tramvaják*;
- jména obyvatelská, např. *Brňák, Hanák, Malostraňák, Pražák*;
- jména označující členství ve skupině lidí, např. *devětsilák, esenbák, tatrovák*;
- jména označující zvířata (přechýlená od feminin), např. *lišák, myšák, opičák*;
- augmentativa, např. (rod m. živ.) *chlapák, sršňák*, (rod m. neživ.) *šutrák, kyják*.

Substantiva odvozená z adjektiv:

- jména označující nositele vlastností, např. (rod m. živ.) *blonďák, dobrák, chudák, silák*, (rod m. neživ.) *dřevák, gumák, stříbrňák, širák, tvrďák*.

Substantiva odvozená z číslovek:

- jména označující pořadí, např. *prvák, druhák, páťák*;
- jména označující srnčí zvěř podle paroží, např. *desaterák, dvanácterák*.

Substantiva odvozená ze sloves:

- jména činitelská, např. *divák, honák, pašerák, zpěvák*;
- jména označující nástroje, např. *bodák, drapák, naviják*.

V tabulce 1 jsou uvedeny počty lemmat v jednotlivých slovotvorných kategoriích.

Kategorie	**Rod mužský životný**	**Rod mužský neživotný**
	724	**627**
konatelská	100	-
obyvatelská	78	-
skupiny	63	-
zvířata	16	-
augmentativa	9	6
nositelé vlastností	141	286
pořadí	5	-
paroží	7	-
činitelská	105	-
prostředky činnosti	-	197
Celkem	**524**	**489**

Tabulka 1 Pokrytí substantiv patřících k jednotlivým kategoriím

Zařazení slov do slovotvorných kategorií je v řadě případů samozřejmě pouze předběžné a může si vyžádat ještě další přezkoumání. U některých slov se nejednoznačně vyjadřují i MČ I a PMČ, např. slovo *světák* je zde zařazeno současně pod jména konatelská (*ten, kdo zná svět*), jména obyvatelská (*obyvatel světa*) a jména nositelů vlastností (*světový člověk*). Některá slova svým charakterem nespadají do žádné ze jmenovaných kategorií, např. slovo *vědmák* rodu mužského životného je sice přechýlené od ženského rodu, ale neoznačuje zvíře. Byla by také možná ještě další podrobnější subkategorizace, např. u jmen označujících prostředky činnosti je možné vydělit kategorii skutečných nástrojů (*naběrák*), druhy tanců (*obkročák*) a části zvířecího těla (*běhák*). V počítačovém zpracování přirozeného jazyka a případném využití seznamů v počítačových aplikacích je však příliš detailní vymezování významů spíše nežádoucí.

7 Výsledky a závěry

Věnovali jsme pozornost českému sufixu *-ák* a popsali jsme jeho derivační chování. Je potřeba zdůraznit, že jeho popis slouží jako vzorec, jenž lze aplikovat i na další české substantivní sufixy. K tomu dodejme, že pro naznačený způsob zpracování je připraven soubor cca 40 českých sufixů, přičemž chování některých z nich je již zčásti popsáno.

Hlavní výsledky jsou obsaženy v tabulce 1, jež ukazuje, jaké významy nese sufix *-ák* a s jakými četnostmi. Výchozí seznamy substantiv byly získány pomocí nástroje Deriv, analýza významů a jejich klasifikace byla provedena převážně manuálně. Výsledkem klasifikace předložené v tabulce 1 je také vyjasnění víceznačnosti sufixu *–ák*, což pokládáme za relevantní výsledek popisované práce. Tato homonymie nepředstavuje problém pro lidské uživatele, ovšem pro počítačové aplikace je její řešení nutnou podmínkou.

Rádi bychom zdůraznili, že tento článek je jednou z prvních studií věnovaných popisu chování českého sufixu *-ák*, jež představuje východisko pro zpracování dalších substantivních

sufixů. Dodejme ještě, že bez nástroje Deriv a zmíněných morfologických a korpusových dat by tyto v dané situaci úplné výsledky nemohly být získány.

REFERENCE

PETR, J. (1986): Mluvnice češtiny I. Praha: Academia

KARLÍK, P. – NEKULA, M. – RUSÍNOVÁ, Z. (1995): Příruční mluvnice češtiny. Praha: Lidové noviny

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

ŠMERK, P. (2010): K počítačové morfologické analýze češtiny. Brno: FI MU. Disertační práce

ŠMERK, P. (2010a): Deriv (software). Brno: FI MU. Dostupný z WWW: <http://deb.fi.muni.cz/deriv>

Korpus SYN2000 (2000): *Český národní korpus – SYN2000.* Praha: Ústav Českého národního korpusu FF UK. Dostupný z WWW: <http://www.korpus.cz>.

GRANTOVÁ PODPORA

Tato práce byla částečně podpořena projektem MŠMT ČR LM2010013 a projektem GAČR P401/10/0792.