



Nová příručka o tvoření slov:
Slovník afixů užívaných v češtině
(automatická morfologická analýza: prostředek
a past při práci s korpusovými daty)

1

Klára Osolsobě

FF MU

osolsobe@phil.muni.cz

Šimandl, J. (ed.). Slovník afixů užívaných v češtině, Praha : Karolinum. 2016.

Slovník afixů užívaných v češtině

Šimandl Josef (ed.)

Karolinum 2017

brožovaná, 654 str.

ISBN 9788024635446

**SLOVNÍK
AFIXŮ
UŽÍVANÝCH
V ČEŠTINĚ**

JOSEF ŠIMANDL (ED.)

KAROLINUM

-15% 490,-
417,-

SKLADEM



<http://www.ujc.cas.cz/electronicke-slovniky-a-zdroje/Slovník-afixu-html>

The screenshot shows the website interface. At the top left is the logo of the Ústav pro jazyk český Akademie věd České republiky. To the right are navigation links: Úvod, Intranet, English, and a search bar labeled 'Vyhledávání'. Below this is a dark blue navigation bar with white text links: O ústavu, Internetová jazyková příručka, Jazyková poradna, Elektronické slovníky a zdroje, Věda a výzkum, Expertní činnost, and Kontakty. The main content area has a title 'Slovník afixů užívaných v češtině' and a sub-header 'Vyhledávat ve Slovníku afixů můžete zde:'. Below this is a small thumbnail of the website's 'Úvod' page. The URL www.slovnikafixu.cz is displayed. The main text describes the dictionary's focus on prefixes and suffixes in contemporary Czech, its non-traditional perspective on word formation, and its utility as a corpus-based source for research. A sidebar on the left contains a list of links to various linguistic resources, with the current page highlighted in a dark blue box.

Ústav pro jazyk český
Akademie věd České republiky

Úvod Intranet English

Vyhledávání

O ústavu Internetová jazyková příručka Jazyková poradna **Elektronické slovníky a zdroje** Věda a výzkum Expertní činnost Kontakty

- Akademický slovník současné češtiny
- Bibliografické ročenky
- Český jazykový atlas
- Internetová jazyková příručka
- Kartotéka lexikálního archivu (1911–1991)
- Korpus DIALOG
- Korpus MONOLOG
- Lexikální databáze humanistické a barokní češtiny
- Lexiko
- Místní jména v Čechách
- Neomat
- Příruční slovník jazyka českého (1935–1957)
- Slovník afixů užívaných v češtině**

Slovník afixů užívaných v češtině

Vyhledávat ve Slovníku afixů můžete zde:

www.slovnikafixu.cz

Slovník afixů zpracovává předpony (prefixy) a přípony (sufixy), představuje však i četné části složených slov, jimž se někdy přiznává blízkost afixům. Je to tedy kompendium popisující všechny důležité odvozovací prvky slov užívaných v současných českých textech – včetně slov přejatých. Netradiční pohled na tvoření slov od afixů dává trochu odlišný obraz než běžný směr od derivátů: do ohniska pozornosti se leckdy dostanou doklady z různých příčin opomíjené. Výhodou je i korpus jako zdroj dat: materiál je dostatečně reprezentativní a poskytuje vesměs spolehlivý základ pro odborné závěry. Pomocí slovníku je možné studovat systém afixů v závislosti na jejich frekvenci či produktivitě. Lze z něho vyjít při řešení otázek synonymie afixů, návaznosti na povahu základu a mnoha dalších jevů.

Co v něm lze najít?

- ▶ **Česká utvořená slovní zásoba**
- ▶ **Slovní zásoba se zřetelnou morfologickou stavbou**

Sufix -ouš

-ouš

Stavba: *-ouš(0)*, kde (0) reprezentuje koncovky vzoru muž (*bělouš, komouš*), zcela výjimečně stroj (slovo *rohous*, které vzniklo expresivní modifikací slova *rohypno*).

Při odvozování I. od substantiv vyjadřuje (1) **předmětný vztah** ke skutečnosti označené základovým slovem: *chocholouš, břehouš, vodouš* (typicky jde o termíny z oblasti zoologie). V rámci tohoto typu lze rozlišit pojmenování (1a) **podle výrazného znaku**: *chocholouš, rypouš*; případy, kdy základové slovo označuje (1b) **místo typického výskytu**: *břehouš, vodouš*; pojmenování (1c) **podle podobnosti**: *vlkouš*. V řadě případů sufix základovému substantivu pouze (2) **dodává expresivitu**; vlastní derivaci pak obvykle předchází mechanické krácení: *komouš, homouš, úchylouš*. K typu (2) náleží i deriváty od křestních jmen: *Fanouš, Jarouš, Zdendouš*.

Při odvozování II. od adjektiv vyjadřuje význam (3) **nositele vlastnosti**: *starouš, bělouš, teplouš*.

Sufixu *-ouš* konkuruje zejm. sufix *-och*; deriváty se obvykle liší stylovou hodnotou (srov. *černo*ch × *černo*uš, *staro*ch × *staro*uš), méně často lexikálním významem (srov. *bělo*ch × *bělo*uš). Stylově neutrálním synonymem k substantivu *svalouš* je jméno *svalovec*, utvořené formantem *-ovec*.

S výjimkou zoologických termínů jsou slova utvořená sufixem *-ouš* obvykle expresivní.

SYN2010

Dotaz [lemma="*.ouš"] dává 87 lemmat, z toho 41 relevantních.

20 lemmat s nejvyšším počtem dokladů: *starouš* (K3) 203, *bělouš* (K3) 155, *teplouš* (K3) 148, *Bohouš* (K2) 117, *komouš* (K2) 63, *rypouš* (K1a) 53, *Fanouš* (K2) 47, *Jarouš* (K2) 25, *chocholouš* (K1a) 19, *drahouš* (K3) 18, *vodouš* (K1b) 15, *konzouš* (K2) 12, *Zdendouš* (K2) 10, *dědouš* (K2) 9, *svalouš* (K1a) 8, *cukrouš* (K1c) 8, *černo*uš (K3) 7, *homouš* (K2) 7, *křivouš* (K3) 4, *prdlouš* (K3) 3.

Řetězec *-ouš* je součástí řady příjmení: *Jirouš, Vondrouš, Kotrouš*. Slovo tvorná stavba těchto slov dnes již není průhledná.

Stavební prvek *-hypno*

hypno-

První část přejatých složených slov. Z řec. *hypnos* ‚spánek‘; odkazuje ke spánku, i navozenému uměle. Příklady lemmat: *hypnoterapie*, *hypnoterapeut*, *hypnogram*, *hypnoanalýza*, *hypnopedie*, *hypnagogický*.

JŠ

Frekvenční zpráva

- Korpus SYN2010
- Korpus SYN (v3)
- Automatická morfologická analýza jako: prostředek a past při práci s korpusovými daty

- ▶ Tokenizace
- ▶ Nejednoznačná automatická analýza (lemma+tag): slovník
- ▶ Desambiguace
- ▶ Závěr

Hesla

- ▶ **na- -o** (<http://www.slovníkafixu.cz/heslar/na-%20o>) VV
- ▶ **za- se** (<http://www.slovníkafixu.cz/heslar/za-%20se>) KO
- ▶ **-oš** (<http://www.slovníkafixu.cz/heslar/-o%C5%A1>) KO
- ▶ **-cí** (<http://www.slovníkafixu.cz/heslar/-c%C3%AD>) KO, JŠ
- ▶ **-í** (<http://www.slovníkafixu.cz/heslar/-%C3%AD>) KO
- ▶ **sou- -í** (<http://www.slovníkafixu.cz/heslar/sou-%20-%C3%AD>)
KO

Tokenizace

- ▶ První krok automatické morfologické analýzy = rozdělení textu na jednotky, s nimiž pracují další kroky. Token odpovídá přibližně tomu, co klasická lingvistika nazývá **grafické / textové slovo**.
- ▶ V rámci NLP se problémy takto pojaté tokenizace řeší v rámci zpracování tzv. **MWE (Multiword Expression)**.

Tokenizace: *na- -o* *na-tvrd-o × ~~na tvrd-o~~*

- ▶ Příslušná adverbia obvykle charakterizuje dvojí způsob psaní, srov.: *načerno / na černo, nahrubo / na hrubo, naměkko / na měkko*. První způsob zápisu je nutno považovat za základní. Do statistické části hesla jsou zahrnuty pouze varianty psané dohromady.
- ▶ Výsledky frekvenční zprávy ukazují pouze **frekvenci jedné z variant grafického úzu** (navíc jde pouze o varianty zachycené ve slovníku automatického analyzátoru, viz níže).

Tokenizace: za- se

za-bouchnout se × za-bouchnul prý se × ~~za-bouchly (se)~~
dveře

- „Statistická zpráva vznikla ruční editací, s omezenou zárukou přesnosti dat.“
- Zohledněné byly pouze případy, kdy se stojí **bezprostředně před/za slovesem <-1,1>** (~~on se asi do ní na plese zabouchl~~).
- Ruční eliminace případů jiných typů reflexivity: ~~dveře se zabouchly~~.

Nejednoznačná automatická morfologická analýza (lemma+tag na základě porovnání se slovníkem)

- Výsledky lemmatizace jsou závislé na rozsahu a obsahu použitého slovníku.
(HAJIČ, J., HLAVÁČOVÁ, J. (2013). *MorfFlex CZ*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9>.)
- Přestože je slovník rozsáhlý a stále se doplňuje, je počet **hapaxových výrazů** v textech stabilní veličina. **Údaje o produktivitě mohou být selektivní (závislé na slovníku).**

Neúplný slovník: *na- -o*

na-prav-o, na-lev-o × na-těsn-o, na-kratičk-o

- Zopakujeme-li u hesla *na- -o* dotaz **s vynecháním morfologické značky (D.*)**, nalezneme další relevantní doklady.
- Slova s malou frekvencí **odpovídají modelu tvoření.**
- **Obraz produktivity lze tudíž napadnout.**

Neúplný slovník: -oš

Mil-oš, Jug-oš × **Káj-oš, Tal-oš**

- Slovník automatického analyzátoru budovaný primárně pro analýzu **psaného spisovného jazyka zachycuje expresivní slovní zásobu včetně proříí velmi selektivně.**
- Dotaz `[lemma="*.oš" & tag="NN[MI].*"]` dává 125 lemmat, z toho **69** relevantních. Dotaz `[lemma="(*.oš) | (*.oš[eiů]) | (*.oších) | (*.ošům) & tag="X.*"]` dává 282 tvarů, z toho **36** relevantních lemmat.

Desambiguace

- Výsledky desambiguace (výběru interpretace ze všech nabízených automatickou morfologickou analýzou) jsou závislé na použité metodě desambiguace.
- **Problém homonymie (transpozice, polyfunkční afixy, náhodné shody** při formálním zadání dotazu) je řešen selektivně.

Desambiguace: -cí vedou-cí

- ▶ **vedoucí** (↔1/2/3) 8.348
- ▶ adjektivum vyjadřuje **(1) aktuální vlastnost plynoucí z děje** (např. **cesta vedoucí lesem** = 'cesta, která vede lesem'), může se **(2) dezaktualizovat** (**vedoucí složky armády** = 'vedení armády') a **(3) substantivizovat** (**vedoucí skupinky** = 'vůdce skupinky').
- ▶ **(1) a (2)** automatická morfologická analýza **nerozlišuje**.
- ▶ **(3)** je sice **zachycen**, nicméně výsledky desambiguace jsou velmi **nespolehlivé**.

Chyby v desambiguaci: cestující-cí

Výskytů: 5 457 | i.p.m. 0: 44,85 (vztaženo k celému "omezeni/syn2010") | ARF 0: 1 864,26 | Výsledek je promíchán

1 / 137 ▶▶

Výběr řádků: základní ▾

- | | | | |
|--------------------------|---|--|--|
| <input type="checkbox"/> | , na kterou jsou křehce napojeny jednotlivé prsty . Pro | cestující/cestující/NNMP4-----A----- | je toto letiště skutečným zážitkem . Baskická metropole Bilbao tak |
| <input type="checkbox"/> | jezdít po stejné trase pod číslem 4 . " Pro | cestující/cestující/NNMP4-----A----- | to přinese změnu v tom , že už nebudou moci |
| <input type="checkbox"/> | " ukončení letu " pro sestřelení jihokorejského dopravního letadla plného | cestujících/cestující/AGMP2-----A----- | sovětskými barbary) . K takovým eufemismům patří i " |
| <input type="checkbox"/> | korun . Týdenní jízdenka platí od pondělí do neděle . | Cestující/cestující/AGMP1-----A----- | si ji mohou na následující týden koupit od úterka až |
| <input type="checkbox"/> | - někdejší oslí stezka se změnila ve vysokohorskou silnici . | Cestujících/cestující/AGFP2-----A----- | na trati nikdy moc nebylo a náklad , který železnici |
| <input type="checkbox"/> | četníci , kteří na nádraží zatýkali chmatáka , jenž okrádal | cestující/cestující/AGMP4-----A----- | o jejich hodinky a další cennosti . Svě umění s |
| <input type="checkbox"/> | . * Železniční společnost Eurostar přepravila loni rekordních 8,26 mil. | cestujících/cestující/AGMP2-----A----- | (meziročně o 5,1 % více) . Tržby z |
| <input type="checkbox"/> | . Ve vlaku bylo 34 osob , z toho 31 | cestujících/cestující/AGMP2-----A----- | a 2 členové posádky v osobním vagónu (požární odolnost |
| <input type="checkbox"/> | " uvedl Petr Žáček z Mostu . Na zvýšený počet | cestujících/cestující/AGMP2-----A----- | do Německa se připravují i České dráhy . " Víkendové |
| <input type="checkbox"/> | 154 poté , co v sobotu zemřela jedna ze zraněných | cestujících/cestující/AGMP2-----A----- | . Podle madridské zdravotní služby boj o život prohrála María |
| <input type="checkbox"/> | mu podařilo posadit stroj na hladinu , zachránil alespoň části | cestujících/cestující/AGIP2-----A----- | život . V noci na včerejšek byly trosky letadla vytaženy |

Desambiguace: **sou-** **-í** **soutěžení, souručenství, soukromí**

► **Náhodné shody řetězců**

- „Řetězec **sou-** **-í** mají i substantiva, která **nejsou tvořena příslušným** cirkumfixem. Jsou to a) **dějová jména** od sloves s prefixem/počátečním řetězcem **sou**, např. *soustředění, soužití, soutěžení*; b) deriváty jmen na **-ství/-ctví** s počátečním řetězcem/prefixem **sou-**, např. *sousedství, souručenství*; c) substantivum **soukromí**, utvořené od adjektiva/adverbia. V korpusu SYN2010 je jich celkem 28.“

Závěr

- ▶ § **limity automatické analýzy** počítáme a snažili jsme se na ně upozorňovat uživatele v takovém rozsahu, aby nebyl uveden v omyl.
- ▶ Veškeré údaje (pokud není uveden opak) odkazující na korpus jsou vzaty z referenčního korpusu (2010) a způsob jejich zjištění je dostatečně popsán. Každý uživatel slovníku má tudíž možnost uvedený postup **zopakovat i s použitím jiných dat** (jiných korpusů) (požadavek empirické testovatelnosti výsledků).
- ▶ **Bez využití výsledků automatické morfologické analýzy** by vznik slovníku (psaní hesel) byl a) nesrovnatelně **časově náročnější**, b) podstatně **nákladnější** a ve svém výsledku c) **méně objektivní**.

DĚKUJI VÁM ZA POZORNOST!

► Více soch je **sou-soš-í**,
více žen je **s-ouž-en-í**
nikoli **sou-žen-í**.