K efektivitě manuální a poloautomatické excerpce neologismů

Jakub SLÁMA | Ústav pro jazyk český AV ČR

On the efficiency of manual and semi-automatic detection of neologisms

The paper presents a simple semi-automatic neologism detection procedure: a trivial Python script processes a text file, making use of a Czech morphological tagger, and extracts all words unrecognized by the tagger as potential neologisms. The list of these candidates has to be checked by a human (hence the label *semi*-automatic). This method was applied to a set of texts that were also analyzed in a more traditional way, by the "reading and marking" technique (i.e. the current practice). The comparison of the two methods has revealed that the semi-automatic procedure clearly outperforms the current practice both in speed and in efficiency.

Key words: data collection, manual detection of neologisms, neologisms, Python, semi-automatic detection of neologisms

Klíčová slova: manuální excerpce neologismů, neologismy, poloautomatická excerpce neologismů, Python, sběr dat

1 Úvod[1]

Pracovníci excerpčního úseku oddělení současné lexikologie a lexikografie Ústavu pro jazyk český AV ČR dlouhodobě budují elektronickou databázi Neomat (ze spojení „neologický materiál").[2] Excerptoři pročítají primárně tištěné publicistické texty, vyhledávají neologismy v širším smyslu[3] a zapisují je do databáze. Je oprávněné ptát se, zda není možné tuto činnost alespoň částečně automatizovat. O stávající praxi lze hovořit jako o manuální excerpci neologismů, o její částečně zautomatizované alternativě jako o excerpci poloautomatické. Tyto dva způsoby zachycování lexikálních neologismů jsou stručně charakterizovány ve druhé části tohoto textu, jehož primárním účelem je srovnat efektivitu manuální excerpce s efektivitou nejjednodušší možné metody excerpce poloautomatické (viz třetí oddíl). Východiskem pro srovnání jsou texty ze čtyř vydání týdeníku Reflex; posouzení obou metod excerpce spočívá ve srovnání seznamu lexikálních jednotek (LJ), které z daných

[1] Rád bych poděkoval oběma anonymním recenzentům za podnětné a cenné připomínky.

[2] Databáze je volně dostupná z adresy <http://www.neologismy.cz>. K historii excerpce a neologické databáze Neomat viz Goláňová (2011), pro aktuálnější údaje viz Filiačová (2016).

[3] Tj. včetně okazionalismů apod. (srov. Janovec, 2013, s. 106).

textů ručně zapsala do databáze Neomat excerptorka, a seznamu LJ, které z daných textů vyextrahoval jednoduchý skript v jazyce Python (viz třetí oddíl).

Za zmínku stojí fakt, že ačkoliv tento text vychází ze zdánlivě specifických potřeb neologické excerpce, může poskytnout metodologické východisko pro sběr jazykových dat pro slovotvorný či lexikologický výzkum obecně. Cílem tohoto textu není navrhnout optimální algoritmus pro účely poloautomatické excerpce, ale poukázat na to, že i „nekomputačnímu" lingvistovi dává minimální znalost programování do rukou cenný nástroj pro sběr jazykových dat.

2 Metody excerpce

2.1 Manuální excerpce

Jak již bylo zmíněno, stávající způsob práce na rozvoji databáze Neomat lze označit jako manuální excerpci, případně jako excerpci tradiční (srov. Radimský, 2003, s. 23). Excerptor jednoduše pročítá (obvykle) tištěné publicistické texty, a pokud se domnívá, že zaregistroval neologickou LJ, v textu ji označí. Vyznačené LJ jsou následně srovnávány s ověřovacími zdroji (tj. především s existujícími výkladovými slovníky češtiny a s heslářem NLA), na základě čehož je rozhodnuto, zda bude LJ zanesena do databáze Neomat. Tento postup práce tedy vzhledem k nesnázím obklopujícím snahy jednoznačně vymezit pojem neologismu vychází z kritéria nezachycenosti (srov. Filiačová, 2016, s. 101): není-li nalezená LJ zachycena v ověřovacích zdrojích, je považována za neologismus a zapsána do databáze. Zaznamenávána jsou jak nová slova, resp. jednoslovné formální neologismy (včetně odvozenin, kompozit, zkratek a přejímek), tak neosémantismy a víceslovné LJ, okrajově též doklady sloves v neobvyklých valenčních rámcích či slova, u nichž je zřejmý stylový posun.

Evidentní nevýhodou manuální excerpce je její časová náročnost a pracnost, dále malé množství a nízká různorodost zpracovaných textů (naprostou většinu excerpovaných textů tvoří psaná publicistika). Za výhodu manuální excerpce je obvykle považováno to, že excerptor automaticky odliší skutečné potenciální neologismy od slov s překlepy apod. Jako největší výhodu manuální excerpce lze pak chápat to, že excerptor snadno zachytí i víceslovné LJ, neosémantismy nebo například posuny ve valenci či stylové charakteristice.

2.2 Poloautomatická excerpce

O poloautomatické excerpci lze mluvit jako o poloautomatické proto, že ačkoliv zpracování textů je automatizováno (tj. je využit speciální nástroj, resp. skript či program), stále je třeba ručně vybrat analyzované texty a vyhodnotit výsledky této metody (tj. posoudit automatickým nástrojem vyextrahovaná slova a rozhodnout,

zda jde opravdu o neologismy). Základní postup poloautomatické excerpce lze zjednodušeně shrnout následovně:

a) nástroj texty, které má zpracovat, upraví (např. odstraní HTML značky apod.);
b) texty jsou tokenizovány a případně lemmatizovány (lemmatizace není nutná);
c) je vytvořen seznam všech textových slov (tento krok není nezbytně nutný);
d) jedno každé textové slovo je mechanicky porovnáno s předdefinovaným slovníkem (tzv. *exclusion list*);
e) textová slova, která nejsou zahrnuta v předdefinovaném slovníku, jsou považována za neologismy.

Tento postup může být rozšířen např. tak, aby nástroj rozeznal vlastní jména a z výsledného seznamu potenciálních neologismů je vyřadil.

Zjevnou výhodou poloautomatické excerpce je především možnost rychleji zpracovat textový materiál nejen rozsáhlejší, ale také různorodější. S ohledem na různorodost zpracovávaných textů se nabízí otázka, zda by nebylo při potenciálním přechodu k poloautomatické excerpci výhodné navázat spolupráci např. s Ústavem Českého národního korpusu FF UK, jehož korpusy obsahují texty různých žánrů. Obvykle se traduje, že texty z korpusů jsou pro potřeby excerpce příliš neaktuální; tato mezi excerptory rozšířená domněnka však, pokud vím, nikdy nebyla empiricky ověřena. Už i při velmi zběžném a namátkovém procházení seznamu hapax legomen v téměř 20 let starém korpusu SYN2000 jsem nicméně našel řadu slov, která dodnes nejsou ve slovnících ani v Neomatu zachycena, ale podle současných metodik by do Neomatu být zapsána mohla (a měla), ačkoliv nejsou zcela nová, např. *antiautoritářsky*, *blbounka*, *černopáska*, *docenitelnost*, *extracelulárně*, *kryptodiktatura*, *přesaditelně*, *sabaťárna*, *spoluhostitel*, *superlumpárna*, *šansonovitý*, *vymýšlitel*, *ženskolog*.

Dobře známou nevýhodu poloautomatické excerpce představuje především obvyklá nemožnost zachytit tímto postupem neosémantismy a víceslovné LJ. Podrobnější charakteristiku tohoto postupu, jeho možných rozšíření a jeho nevýhod podává spolu s přehledem vybraných nástrojů Sláma (2017).

3 Metodologie

Srovnání manuální excerpce a jednoduché metody poloautomatické excerpce je provedeno s využitím textů ze čtyř vydání týdeníku Reflex, která byla publikována v lednu 2018. Všechna čtyři čísla časopisu byla ručně zpracována excerptorkou v průběhu března; excerptorka o přípravě tohoto textu nevěděla, a lze tedy předpokládat, že při manuální excerpci pracovala zcela standardním způsobem. Excerptorka zanesla do databáze Neomat 39 záznamů, mezi nimiž je jedno sousloví (*smíšená realita*) a 37 jednoslovných LJ, jež zde uvádím pouze v abecedním výčtu:

antibabišismus, *antibabišovsky*, *arcimág*, *beďarovitý*, *bombík*, *cinefilek*, *cizo-kulturnost*, *kinematovýstava*, *konspirologický*, *kryptofašokomouš*, *kvazičistý*, *mikroláska*, *minimenšina*, *monstrprodukce*, *multietnicismus*, *namistrovanec*, *nedelegát*, *nemnichovan*, *netflixí*, *okamurák*, *pateorie*, *patobiont*, *protiinfarktní*, *rádobyspolečenskokritický*, *rádobysvětoobčan*, *rudokaštanový*, *sextremismus*, *supersmartfoun*, *technovládce*, *Trumpland*, *ublbnout*, *videodivize*, *vítač*, *vítačský*, *zajoulovat*, *znovunalézt*, *znovuvzedmutý*

Za účelem srovnání metod excerpce byly dané texty z týdeníku Reflex staženy v souboru formátu .html z mediálního archivu Mediasearch.[4] Následně byly zpracovány jednoduchým skriptem, který vychází z výše shrnutého postupu poloautomatické excerpce, modifikovaného způsobem naznačeným v tomto oddílu.

K vytvoření jednoduchého skriptu pro poloautomatickou excerpci byl použit skriptovací programovací jazyk Python (verze 3),[5] často využívaný pro lingvistické účely. Jeho základy si lze snadno osvojit na řadě webových portálů, ale dokonce také s pomocí příruček určených přímo pro lingvisty (srov. např. Bird – Klein – Loper, 2009).

Z textů v souboru formátu .html bylo nejprve automaticky odstraněno HTML tagování, k čemuž byla využita knihovna re[6] v Pythonu. Text byl tokenizován, lemmatizován a morfologicky označkován s využitím nástroje MorphoDiTa[7] (srov. Straková – Straka – Hajič, 2014). Důležité je, že při tom nebyl využit tzv. guesser; pokud se v textu objeví například spojení *malí cinefilkové*, tagger využívající guesser daným tokenům přiřadí lemmata *malý* a *cinefilkový* a obě slova považuje za adjektiva, a proto jim přiřadí morfologické tagy s *A* na první pozici. Když guesser aktivní není, slovu *cinefilkové* je přiřazeno lemma *cinefilkové* a tag začínající znakem *X*. Právě to umožňuje z textu efektivně – byť možná poněkud triviálním způsobem – automaticky vyextrahovat neznámá slova. Ústřední předpoklad, ze kterého skript vychází, je tedy ten, že tato neznámá slova budou slovy relativně novými, tj. neologismy v širším smyslu (včetně okazionalismů apod.).

Na základě výše popsaného základního postupu skript zcela jednoduše z otagovaného textu vyextrahuje slova s tagem začínajícím znakem *X* a seznam těchto slov s krátkým kontextem uloží do textového souboru. Následuje ukázka výstupu:

teátr
? Koho zčásti pobil a zčásti odvlekl do Říma Titus Vespasianus , koho hnal Vítězným obloukem , koho zotročil , aby mu postavil Flaviův amfi ***teátr*** *, známý jako Koloseum ? Nejde jen o svědectví Josefa Flavia , u něhož by zavádějící informace byla pochopitelná , protože sám byl Žid*

---

4 V mediálním archivu Newton Mediasearch (dostupný online z adresy <http://mediasearch.newtonmedia.cz>) je z lednových vydání Reflexu k dispozici 213 různých textů. V tomto archivu obvykle chybí texty reklam, vložených letáků apod.

5 Viz <https://www.python.org>. V češtině srov. např. Pilgrim (2010).

6 Viz <https://docs.python.org/3/library/re.html>.

7 Viz <http://ufal.mff.cuni.cz/morphodita>.

Juvenala
*by zavádějící informace byla pochopitelná , protože sám byl Žid . Ale co si máme počít se stejným tvrzením ve spisech Catulla , Tacita , **Juvenala** , ve verších Horatia ? Máme snad co dělat se sionistickým spiknutím , jehož tvůrci záměrně vytvořili mýtus o židovském Jeruzalémě , o Šalamounově*

konspirologická
*Nebo jejich díla byla zákeřně upravena až po roce 1948 , jakož i texty všech , kteří se na ně odvolávají ? Byla by to **konspirologická** operace nevídaného rozsahu . KOMISE PRO LIDSKÁ PRÁVA V podstatě jsme svědky procesu , při němž se utváří totalita . V Orwellově utopii 1984*

Zběžný pohled na výstupní seznam 1458 slov naznačuje, že jeho velkou část tvoří cizí vlastní jména (např. *Juvenala* v ukázce výstupu výše), a proto byla do skriptu pro zjednodušení přidána další podmínka: slovo s tagem začínajícím znakem *X* je do výstupního seznamu zařazeno pouze tehdy, pokud jeho první písmeno není jediné písmeno psané velkým písmenem (iniciálové zkratky tedy do seznamu zařazeny jsou). Tento krok sice výrazně urychlí manuální kontrolu výsledného seznamu (ten nyní obsahuje 872 slov), zároveň však může vést k tomu, že do seznamu nebude zařazeno např. slovo *Trumpland*, které bylo v textech nalezeno při manuální excerpci. (Toto slovo se však v analyzovaných textech objevilo také v grafické podobě *TRUMPLAND*, a skript jej proto zaznamenal.)

V dalším kroku byl seznam potenciálních neologismů ručně pročištěn; odstraněny byly části slov, které byly v textu získaném z mediálního archivu chybně oddělen mezerou, a proto je tagger považoval za samostatná slova (viz *teátr* v ukázce výstupu výše). Dále byla odstraněna nerelevantní slova psaná velkými písmeny (např. *DEPECHE* v názvu hudební kapely Depeche Mode, který byl v titulku zprávy psán verzálkami), cizí slova v názvech či cizojazyčných citátech (např. *du* v kontextu *restaurace L'Auberge du Pont de Collonges* či slovo *kradnú* v kontextu *multimilionář s komunistickou minulostí říká, že všeci kradnú*), sedm slov[8] obsahujících překlepy a další nežádoucí položky (např. názvy webových stránek typu *reflex.cz*).

V dalším kroku bylo ze seznamu vyřazeno 47 LJ, které mají v databázi Neomat dostatečný počet dokladů,[9] 16 LJ, které jsou doloženy v jednom z neologických slovníků (NSvČ; NSvČ2),[10] 13 LJ, které jsou zachyceny v některém ze dřívějších

[8] Konkrétně *členewm*, *koproduční*, *lékářkou*, *mainsteramoví*, *olomouc*, *prohlásili*, *přípěvku*.

[9] Jde o tyto LJ: *agrofertí*, *bifidobakterie*, *blockchain*, *blockchainový*, *cookie*, *debilizace*, *fake news*, *fejk*, *fotomodeling*, *ghostwriter*, *hejtr*, *hnojomet*, *hyperrealita*, *hyperúspěšný*, *infografika*, *kryptoměna*, *loser*, *mikrobiom*, *mileniál*, *náckovský*, *nobelistka*, *noisový*, *nominant*, *okamurovec*, *olajkovat*, *osmdesátkový*, *podcast*, *postpravda*, *pravdoláskařský*, *protibabišovský*, *protitrumpovský*, *protiuprchlický*, *předokolka*, *putinismus*, *retronálada*, *superzajímavý*, *tweet*, *tweetovat*, *účelovka*, *virál*, *youtuber*, *youtubový*.

[10] Jde o tyto LJ: *audiokniha*, *barbínovský*, *cool*, *cookies*, *demokratura*, *intoš*, *falokratický*, *megaúspěšný*, *minihra*, *ostalgie*, *protofašistický*, *technooptimista*, *transgender*, *ukritizovat se*, *unipolarita*, *ztransparentnit*.

výkladových slovníků češtiny,[11] a 13 LJ, které jsou zaznamenány pouze v NLA.[12] Žádná z těchto 89 vyřazených LJ by podle excerpčních zásad neměla být v souladu s výše zmíněným kritériem nezachycenosti do databáze Neomat zapsána.

Úspěšnost zde prezentované metody poloautomatické excerpce závisí na tom, jaké LJ jsou zahrnuty v hesláři morfologického slovníku, s nímž pracuje nástroj MorphoDiTa; ačkoliv jsem neprovedl jeho srovnání s hesláři slovníků, které patří mezi tzv. ověřovací zdroje (viz výše), zde uvedené počty LJ naznačují, že hesláře se značně překrývají; víceméně pouze u 13 slov, která jsou zachycena ve starších slovnících češtiny, překvapí, že byla označkována jako slova neznámá (obzvlášť s ohledem na to, že některá z nich introspektivně považuji za relativně běžná, např. *chrám*, *meeting*, *uskupení*, *viník*).

Ruční pročištění automaticky vytvořeného seznamu potenciálních neologismů je nezbytné, oproti manuální excerpci však nepředstavuje žádnou zátěž navíc (ruční ověřování toho, zda vyexcerpované LJ opravdu mají být zapsány do neologické databáze, se provádí i při manuální excerpci). Při použití sofistikovanější metody poloautomatické excerpce by navíc bylo možné pročištění seznamu do určité míry usnadnit například rozšířením skriptu o funkci, která by porovnávala automaticky vytvořený seznam s hesláři slovníků, jež existují v elektronické podobě, a s heslářem databáze Neomat. Tak by mohla být automaticky ze seznamu potenciálních neologismů vyřazena slova, která už jsou zachycena (v tomto případě výše uvedených 89 LJ), a proto by do databáze Neomat zaznamenána být neměla.

4 Výsledky srovnání

Manuální excerpcí ze čtyř sledovaných vydání týdeníku Reflex bylo zachyceno sousloví *smíšená realita* a 37 jednoslovných LJ (viz oddíl 3). Celkem 35 z těchto 37 LJ zachytil i jednoduchý skript pro poloautomatickou excerpci. Excerptorka do databáze tedy zapsala pouze jedno sousloví a dvě slova, která automatický skript nezachytil, *vítač* a *znovunalézt*. V prvním případě jde o neosémantismus; slovo *vítač* je vyloženo už v PSJČ, nově se však objevuje ve významu ‚kdo nekriticky, přehnaně podporuje přijímání uprchlíků'.[13] Víceslovné LJ a neosémantismy pro poloautomatickou excerpci představují dobře známý problém (srov. Sláma, 2017, s. 42). Slovo *znovunalézt* se v textu objevilo na začátku věty, tedy s velkým počátečním písmenem, a proto nesplnilo podmínku pro zařazení na seznam potenciálních neologismů, která byla do skriptu přidána ve snaze zásadně tento seznam zkrátit vyřazením

[11] Jde o tyto LJ: *alkylhalogenid*, *ethylový*, *chrám*, *kombuča*, *mecheche*, *meeting*, *meziplyn*, *oxymóron*, *ptydepe*, *ublbovat*, *ukázečka*, *uskupení*, *viník*.

[12] Jde o tyto LJ: *butyrát*, *hydropat*, *jednoplošný*, *klikařství*, *koherentnost*, *laktobacil*, *mention*, *náckovský*, *piedestálek*, *sociopat*, *štamtyš*, *štandopéde*, *vtipník*.

[13] Autorkou tohoto výkladu významu je M. Lišková (viz Kedroň, 2018).

vlastních jmen. V tomto případě tedy nejde o nedostatek poloautomatické excerpce, ale pouze o nedostatek zde použitého skriptu. Ten by mohl být snadno upraven tak, aby slova s velkým počátečním písmenem vyřadil pouze tehdy, nestojí-li na začátku věty (resp. po terminálním interpunkčním znaménku); v ideálním případě by pak pracoval s takzvaným rozpoznáváním pojmenovaných entit (srov. Straková, 2017; pro češtinu viz Ševčíková – Žabokrtský – Krůza, 2007).

Vedle 35 LJ, které byly zachyceny i excerptorkou, skript z textů použitých při manuální excerpci vyextrahoval 82 LJ, které ručně zachyceny nebyly, ale podle aktuálních excerpčních zásad by do databáze Neomat zapsány být měly. Jejich kompletní soupis je k dispozici v příloze tohoto textu. Distribuce těchto 82 LJ napříč sledovanými vydáními (viz Tabulka 1) ukazuje, že nedochází k výrazné akumulaci jednotek, které při manuální excerpci nebyly zachyceny, v jednom ze sledovaných vydání týdeníku. To naznačuje, že výsledek nemůže být například důsledkem opomenutí jednoho z vydání při ruční excerpci. Čtyři ručně nezachycené LJ (*binge-watching*, *lůzr*, *protidrahošovský*, *pixarovský*) se navíc objevily ve dvou různých vydáních, proto 82 LJ v tabulce odpovídá 86 údajům o datu.

Datum	Počet LJ
4. 1. 2018	26
11. 1. 2018	14
18. 1. 2018	19
25. 1. 2018	27
Celkem	86

Tabulka 1: Distribuce zachycených LJ ve sledovaných vydáních.

Zcela bezpečně lze tedy vyloučit možnost, že by mezi automaticky zpracované texty byly omylem zařazeny texty, které excerptorka manuálně nezpracovávala. Například z glosy s titulkem „Kazisvěti a víno“ ze 4. ledna byla navíc do databáze zapsána slova *protiinfarktní* (*víno obsahuje látky protirakovinné a protiinfarktní*), *ublbnout* (*jedna sklenička vám ublbne stovky mozkových buněk*) a *zajoulovat* (*jedna sklenička vás prý zajouluje na 481 000*). Skript z konce téhož krátkého textu vyextrahoval i slova *ublbovat* (zachyceno v PSJČ ve slang. významu ‚činiti blbým‘) a ve slovnících nezachycené *přichytřovat* (*nejen že vám buňky neublbuje, ono je naopak přichytřuje*). Přinejmenším druhé z těchto sloves by podle platných metodik excerpce mělo zcela jednoznačně být do databáze Neomat zapsáno.

Výše popsaná, nejjednodušší možná (ne-li naprosto triviální) metoda poloautomatické excerpce má tedy výrazně lepší výsledky než manuální excerpce, a navíc je spolehlivější a časově méně náročná.

5 Závěr

Excerpci neologismů (nebo vůbec excerpci lexikálního materiálu pro lexikologický výzkum) lze provádět několika způsoby. Obecně lze rozlišovat způsoby manuální a (částečně) automatizované. K automatizovaným postupům lze využít různých nástrojů. Obvykle se předpokládá, že pro specifické účely neologické excerpce je nutné vyvinout specializovaný nástroj. Kombinace existujícího taggeru pro češtinu s naprosto elementárním skriptem v jazyce Python však vede k překvapivě dobrým výsledkům. Manuální excerpcí bylo ze souboru textů vyexcerpováno jedno sousloví a 37 LJ. Skript z týchž textů vyextrahoval 35 LJ, které byly vyexcerpovány ručně, ale také dalších 82 LJ, které manuální excerpcí zachyceny nebyly, ačkoliv by podle platných excepčních zásad měly být zapsány do databáze Neomat. Skript texty zpracuje během několika vteřin, excerptor během několika hodin. Ruční ověření toho, zda nalezené potenciální neologismy opravdu představují neologismy (ve smyslu kritéria nezachycenosti), je u obou způsobů excerpce srovnatelně časově náročné; při použití sofistikovanějšího skriptu pro poloautomatickou excerpci by však mohlo být výrazně zkráceno.

Na jaře roku 2017 byla v rámci výzvy NAKI II na Ministerstvo kultury ČR podána přihláška návrhu projektu *Nová slova v češtině (nástroj pro automatické vyhledávání neologismů)*; tento společný projekt Ústavu pro jazyk český AV ČR a Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK však nebyl úspěšný. V současnosti je zvažována možnost znovu usilovat o získání grantového projektu; přesná podoba potenciálního nástroje pro vyhledávání neologismů (včetně toho, zda nebo jak by nástroj vyhledával např. i neosémantismy) zatím není známa. Zároveň však není vyloučena možnost, že by skript vytvořený v rámci přípravy tohoto článku mohl být prozatímně využit například pro automatické zpracovávání textů z internetu, které by mohlo dočasně alespoň doplnit excerpci manuální (samozřejmě s vědomím, že skript zachytí pouze jednoslovné formální neologismy v širším smyslu).

Primárním cílem tohoto textu bylo poukázat na to, že i naprosto triviální skript v jazyce Python svou rychlostí a efektivitou může překonat excerptora s dlouholetými zkušenostmi. Využívání poloautomatických nástrojů nejen při neologické excerpci by tak mohlo zásadním způsobem zjednodušit, ale zároveň radikálně zefektivnit sběr jazykových dat (nejen) pro lexikologický výzkum, zároveň by mohlo umožnit zachycení mnohem širšího okruhu slovní zásoby, včetně slovní zásoby z automaticky získaných dat z internetových diskusí, sociálních sítí apod. Smyslem tohoto textu však také bylo ilustrovat, že i lingvista s naprosto elementární znalostí programování má v ruce velice užitečný a spolehlivý nástroj.

LITERATURA

Bird, Steven – Klein, Ewan – Loper, Edward (2009): *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media.

Filiačová, Sylva [= Nzimba, Sylva] (2016): Blendy v elektronickém neologickém archivu Neomat. *Jazykovědné aktuality*, 53(3–4), s. 100–105.

Goláňová, Hana (2011): Novočeský lexikální archiv a excerpce v průběhu let 1911–2011. *Slovo a slovesnost*, 72(4), s. 287–300.

Janovec, Ladislav (2013): Neologie. In: Petra Martinková – Oldřich Uličný (eds.), *Studie k moderní mluvnici češtiny: 4, Dynamika českého lexika a lexikologie*. Olomouc: Univerzita Palackého v Olomouci, s. 105–130.

NSvČ: Martincová, Olga et al. (1998): *Nová slova v češtině: Slovník neologizmů*. Praha: Academia.

NSvČ2: Martincová, Olga et al. (2004): *Nová slova v češtině 2: Slovník neologizmů*. Praha: Academia.

Pilgrim, Mark (2011): *Ponořme se do Python(u) 3 / Dive into Python 3* [online]. Praha: CZ.NIC. Cit. 25. 8. 2018. <https://knihy.nic.cz/files/edice/python_3.pdf>.

Radimský, Jan (2003): *Italské a vybrané francouzské neologismy z oblasti informatiky a nových médií (1990–1996)*. České Budějovice: Jihočeská univerzita.

Ševčíková, Magda – Žabokrtský, Zdeněk – Krůza, Oldřich (2007): Named entities in Czech: annotating data and developing NE tagger. In: Václav Matoušek – Pavel Mautner (eds.), *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3–7, 2007: Proceedings*. Heidelberg: Springer, s. 188–195.

Sláma, Jakub (2017): K (polo)automatické excerpci neologismů. *Jazykovědné aktuality*, 54(3–4), s. 34–46.

Straková, Jana – Straka, Milan – Hajič, Jan (2014): Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In: Kalina Bontcheva – Zhu Jingbo (eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, MD: Association for Computational Linguistics, s. 13–18.

Straková, Jana (2017): Rozpoznávání pojmenovaných entit. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *Czech Ency – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 25. 8. 2018. <https://www.czechency.org/slovnik/ROZPOZNÁVÁNÍ POJMENOVANÝCH ENTIT>.

INTERNETOVÉ ZDROJE

Databáze excerpčního materiálu Neomat [online] (1991–2019). Praha: Ústav pro jazyk český AV ČR. Cit. 25. 8. 2018. <http://neologismy.cz>.

Kedroň, Radek (2018): Ohledáno a sečteno: Vítač se stal vítězem prezidentských voleb 2018. *iRozhlas.cz* [online], 26. ledna 2018. Cit. 25. 8. 2018. <https://www.irozhlas.cz/komentare/prezidentske-volby-2018-vitac-slovo-trend-google_1801260630_rak>.

NLA: *Novočeský lexikální archiv* [online] (2007–2009). Praha: Ústav pro jazyk český AV ČR. Cit. 25. 8. 2018. <http://bara.ujc.cas.cz/psjc>.

PSJČ: *Příruční slovník jazyka českého* [online] (2007–2008). Praha: Státní nakladatelství – Školní nakladatelství – Státní pedagogické nakladatelství. Cit. 25. 8. 2018. <http://bara.ujc.cas.cz/psjc>.

SYN2000: *Český národní korpus – SYN2000* [online] (2018). Praha: Ústav Českého národního korpusu FF UK. Cit. 25. 8. 2018. <www.korpus.cz>.

PŘÍLOHA

Následující tabulka obsahuje seznam lexikálních jednotek (LJ), které nebyly odhaleny manuální excerpcí, ale které ze stejných textů vyextrahoval jednoduchý skript pro poloautomatickou excerpci. Slova označená hvězdičkou jsou v databázi Neomat k 30. dubnu 2018 doložena méně než pětkrát, a proto by měla být podle aktuálních pravidel excerpce do databáze zapsána znovu. Slova neoznačená hvězdičkou v daném významu zachycena nejsou (např. výraz *kočena* vykládají některé starší slovníky jako nářeční označení kočky, zde je však v jiném významu; *trendovat* v databázi Neomat doloženo je, ale v jiném významu, a proto není označeno hvězdičkou). Ve druhém sloupci je poskytnut krátký kontext, z něhož lze obvykle alespoň přibližně vyvodit význam daného slova.

LJ	Příklad užití
akcionalismus	*nejvýznamnější představitel vídeňského akcionalismu*
*antiukrajinský**	*akce vznikla proti antiukrajinské církevní politice*
*aňák**	*trojlístek aňáků, komunistů a okamuráků*
*appka**	*tvůrci mobilní „appky“*
bicepsoid	*sovětský bicepsoid* [o muži]
*binge-watching**	*fenomén takzvaného binge-watchingu, kdy jednotlivci nebo skupiny přátel slupnou několik dílů či celou televizní sérii na jedno posezení*
*bullshit**	*oblast osobního rozvoje je principiálně plná „bullshitu“*
dohůdka	*předivo dohůdek mezi rodiči, mezi partnery, v práci, ve škole*
drakohafan	*legendární fantasy s malým Bastianem, létajícím drakohafanem Falkem a bojovníkem Atrejem se vrací*
dramedy	*rodinné dramedy o takové normální rodině*
erasing	*své technice říkám „erasing“. Ve filmu si najdu potřebný záběr a udělám z něj fotku*
evidence-based	*obecně se tomuto přístupu říká „evidence-based“*
foglarovec	*Foglarovci* [titulek] *Piráti svým důrazem na transparentnost působí foglarovsky správňáckým dojmem…*
fotorealista	*jeden z nejznámějších světových fotorealistů vedle toho portrétoval členy britské královské rodiny*
gastronovinářka	*vyrazila na vídeňskou premiéru „svého“ filmu o Marii Terezii se sestrou, známou gastronovinářkou Hanou Michopulu*
*havloidní**	*zkorumpovaný havloidní fašoun*
*hipísačka**	*nešťastná máma z populárního thrilleru se stává módní ikonou hipísaček*
hnusostroj	*Bojovníky čeká ještě osm dní a jezevec z Vysočiny může překvapit. Spustí svůj oblíbený hnusostroj. Obviní Drahoše, že jeho bratranec byl esesákem. Nebo že je Drahoš Arab.*
horáčkovec	*velká většina horáčkovců půjde volit Jiřího Drahoše*
hosip	*historický hosip neboli „hovno si pamatuju“ pokračuje strašně rychle*
*hulibrk**	*hovořit v přítomnosti homosexuála o hulibrcích nesvědčí o tom, že bychom překypovali taktem*

hypermechanismus	*uživatelé systému provázaného s platebním hypermechanismem*
islamák	*Sudeťáci a islamáci* [titulek]
*islámec**	*nejsou tu ani Sudeťáci, ani laviny islámců*
jako frasa	*za stěhováky říkám, že to bude drahý jako frasa*
*klinč**	*jak Zeman, tak Babiš se drží v klinči a vzájemně se potřebují*
kočena	*„Co jsem viděl fotky, je to kočena.“*
konopohrad	*V USA jde do prodeje první víno, které místo alkoholu obsahuje THC.* […] *Je načase založit nové konopohrady.*
kvantař	*kvantaři mají evidentně taková kvanta peněz, že si mohou platit kancelář na luxusní adrese*
kvazi-příraz	*při pátém nesmělém kvazi-přírazu se mi vrátí projekce sovětského bicepsoida a hlavou se mi rozjede nehumorný seznam všech potenciálních novosibirských breberek, které bych takto mohl posbírat pod předkožku*
literalistický	*prosazovat literalistický výklad koránu*
*lůzr**	*skvělá partička zamindrákovaných lůzrů*
*machista**	*ničím výjimeční venkovští machisté*
mariachiovský	*lyru nahradila mariachiovská kytara*
*marťanství**	*nachází na svém marťanství jen samá pozitiva*
*marvelovský**	*komiks s oficiálními marvelovskými hrdiny*
*maskulinistický**	*žádné maskulinistické protesty zatím nebyly zaznamenány*
*matfyzák**	*to jenom „matfyzáci“ si myslí, že dobrým naprogramováním systému mohou pudové inklinace přelstít*
matfyzácký	*„matfyzácká“ víra ve všemocnost digitálních technologií*
*mikrobiota**	*střevní mikroflóra (mikrobiota) čítá pět set až tisíc druhů mikroorganismů*
minibalkón	*1 + 1 s minibalkónem na květináč s bazalkou*
*moonwalk**	*Jackson moonwalk proslavil až v květnu 1983*
moudroprdnost	*na čtenáře brzy sedne depka z tlachavé vyprázdněnosti a moudroprdnosti textíků*
*multivan**	*zadní sedadlo multivanu*
*nadherecký**	*nadlidský či, chcete-li, nadherecký výkon Garyho Oldmana*
nouvelle cuisine	*hnutí „nouvelle cuisine“, které se v 60. a 70. letech zasloužilo o gastronomickou revoluci*
obchodovatel	*každý obchodovatel bude muset u každé transakce sesbírat, zveřejnit a pět let archivovat 65 různých údajů transakce se týkajících*
okamur	*republice hrozí něco mnohem horšího: komunisté a okamuři ve vládě*
okorigovaný	*čistá práce o práci, dobře přeložená, pečlivě okorigovaná a hezká*
orcký	*kamennost orckých masek*
*ork**	*fantasy thriller s bizarními prvky magie o soužití lidí, orků, elfů, kentaurů a další havěti*
*pelíškovský**	*používat adjektivum „pelíškovská“ pro jakousi až sentimentální útěšnost světa, kterou prý právě tento film do české kinematografie vnesl*
*pixarovský**	*nový pixarovský film COCO*
počítačověherní	*počítačověherní verze legendárního festivalu nezávislých filmů*
*po kunderovsku**	*jde na to pěkně po kunderovsku; jen ty Čecháčky hezky rozdráždit*

p(r)asáž	*četba vybraných p(r)asáží z letošní sklizně českých románů, novel a povídek*
prohulit se	*„Jednou bych chtěl dělat šanson, akorát se k tomu musím prochlastat, prohulit a prožít!“*
protidrahošovský	*protidrahošovský mediální obraz*
protireferendum	*mohl by nasadit protireferendum, zda minulé referendum platí*
protiweinsteinovský	*loni byl Oscar protitrumpovský, letos se čeká protiweinsteinovský*
*provzlykat se**	*Jenny se často provzlykala až k nervovému záchvatu*
psychedelic trance	*potřebuje hodně času, aby jako DJ mohl mixovat oblíbený psychedelic trance*
*putinista**	*škodolibost mnoha českých putinistů ohledně stavu Ukrajiny je založena na nevědomosti*
přichytřovat	*nejen že vám buňky neublbuje, ono je naopak přichytřuje*
rádobybásník	*my, kulturní redaktoři, kteří kdy byli obtěžováni agresivními grafomany, rádobybásníky, zpěváky s přerostlými egy*
*referendární**	*současná záliba v anonymní referendární demokracii*
retroklobouk	*Michaelův retroklobouk*
roulingstoun	*„V tom ten divoký roulingstoun položil dlaně na boky a začal se divoce kroutit.“*
*rukolapný**	*dobrodružný film s rukolapným názvem*
sebekochání	*Sebekochání* [titulek] *Prezident Trump v reakci na nepěknou kritiku o sobě na Twitteru napsal: „…jsem nejen chytrý, ale génius…“*
sebepráskání	*dobrovolné sebepráskání proteklo ze sociálních sítí až do oblastí, kde bylo ještě donedávna zvykem zachovávat jistou diskrétnost*
seniorchef	*seniorchef elektrotechnického závodu*
*smartfoun**	*utrácíme za své smartfouny horentní sumy*
spermofilní	*Je jisté, že během čtení Rimmerova právě vydaného spermofilního komiksu z vás nejspíš cosi vytryskne. Asi smích.*
*spoluobývat**	*Kaňkův palác, který dnes tak říkajíc spoluobývám*
*šéfvývojář**	*šéfvývojář a viceprezident automobilového gigantu*
*trapčit**	*Reflex je součást volebního týmu Drahoše? Už trapčíte nějak moc.*
trendovat	*teď trenduje blockchain a sexuální obtěžování*
upajcnutý	*vřele ho doporučujeme milovníkům thrillerů natočených podle skutečných událostí, jež mohly skončit (nebo taky skončily) pěkně blbě: 127 hodin o jedné upajcnuté ruce*
*virálně**	*pomluvy Drahoše se začínají mezi Zemanovými voliči šířit virálně*
xenomorf	*absurdní film Vetřelec: Covenant, v němž se ke xenomorfovi vrátil jeho původní tvůrce Ridley*
zemančík	*Koláž protidrahošovských plivanců ukazuje, jak neuvěřitelné šířky kampaň proti tomuto kandidátovi dosáhla. Kolegové z Deníku, kde si zemančíci zadali manipulativní inzerát, je pěkně vytrolili.*

Ústav pro jazyk český AV ČR
Letenská 123/4, 118 51 Praha 1
slama@ujc.cas.cz