Typologie víceslovných jednotek v češtině a frekvenční zastoupení jejich hlavních vlastností v žánrově vyváženém korpusu[1]

Vladimír Petkevič — Marie Kopřivová — Milena Hnátková — Tomáš Jelínek — Pavel Kopřiva — Alexandr Rosen — Hana Skoumalová — Pavel Vondřička

ABSTRACT:

Typology of Multiword Expressions in Czech and Frequency of Their Main Features in a Genre-Balanced Corpus

The paper consists of two main parts:

(a) In the first part, a typology of multiword expressions (MWE) in Czech is described in a detailed way. This typology is part of the description of MWE database entries in the lexical database LEMUR containing more than 10,500 MWE entries as of June 2020. MWE properties reflected in this typology are accounted for by categories and their values. Each MWE is identified by a unique lemma; a group of related MWEs is assigned a "superlemma". A MWE is described by the following properties: a MWE definition, characteristic examples, lemmas and morphological features of MWE components (words), as well as the following key categories: MWE style/register, type of usage, syntactic structure (including its representation by a dependency and a phrase-structure tree), aspects of flexibility (variants and fragments, internal modifiability of individual MWE components, possibilities of syntactic transformations of the main MWE components and morphological constraints) and types of idiomaticity on the lexical, morphological, syntactic, semantic and pragmatic level.

(b) In the second part of the paper, the authors focus on the frequency of the main features of the adopted typology in the real language material represented by the genre-balanced SYN2015 corpus, containing 100 mil. word forms (excluding punctuation): a type of usage correlated with a syntactic type and frequency of various kinds of idiomaticity. Our paper seems to be the first attempt at approaching the MWE properties from the point of view of MWE frequencies as types rather than tokens (i.e. frequencies of occurrences of a given MWE).

ABSTRAKT:

Příspěvek má dvě hlavní části:

(a) V první části je podrobně popsána typologie (vlastnosti) víceslovných lexikálních jednotek (dále VLJ) v češtině, přičemž tato typologie je součástí popisu databázových hesel těchto jednotek v lexikální databázi LEMUR, obsahující k červnu 2020 více než 10 500 hesel.[2] Jednotlivé vlastnosti těchto

1 Příspěvek vznikl jako součást projektu *Mezi slovníkem a gramatikou* (Between Lexicon and Grammar), podpořeného Grantovou agenturou České republiky, reg. č. 16-07473S.

2 Databáze LEMUR je podrobně charakterizována v článku Vondřička (2019). Vznikla v Ústavu Českého národního korpusu FF UK a výhledově bude zpřístupněna uživatelům. Bude rovněž postupně propojována s korpusem, kde budou víceslovné lexikální jednotky anotovány, takže bude možné podle anotovaných vlastností vyhledávat. Na vyžádání v Ústavu Českého národního korpusu FF UK je ovšem možné zpřístupnit databázi k nahlédnutí již nyní.

jednotek jsou zachyceny prostřednictvím kategorií a jejich hodnot. U každé jednotky uvádíme její identifikační lemma a tzv. superlemma, definici, typické příklady; dále popisujeme lemmata a morfologické vlastnosti jednotlivých komponent (slov) a poté takové charakteristiky jako styl/varieta VLJ, její typ užití, syntaktická struktura (včetně reprezentace v podobě závislostního a frázového stromu), aspekty ustálenosti/flexibility (včetně variant a fragmentů VLJ, vnitřní modifikovatelnosti jednotlivých komponent VLJ, možností syntaktických transformací hlavních komponent VLJ a též morfologických omezení) a konečně typy idiomatičnosti na rovině lexikální, morfologické, syntaktické, sémantické a pragmatické.
(b) V druhé, hlavní části příspěvku sledujeme frekvenční zastoupení hlavních aspektů této typologie u dosud zpracovaných VLJ: typ užití v korelaci se syntaktickým typem a dále zastoupení různých druhů idiomatičnosti, a to v reálném jazykovém materiálu reprezentovaném žánrově vyváženým korpusem SYN2015 (obsahuje sto milionů slovních tvarů mimo interpunkci). Jde patrně vůbec o první pokus zaměřit se na vlastnosti víceslovných lexikálních jednotek z hlediska četnosti jejich výskytů jakožto typů, nikoli tokenů (tj. četností výskytů dané jednotky).

KLÍČOVÁ SLOVA / KEY WORDS:
víceslovná lexikální jednotka v češtině, typologie víceslovných lexikálních jednotek, frekvence typů víceslovných lexikálních jednotek, idiomaticita, lexikální databáze, žánrově vyvážený korpus
multiword (lexical) expressions in Czech, typology of multiword expressions, frequency of types of multiword expressions, idiomaticity, lexical database, genre-balanced corpus

1. ÚVOD

Víceslovnými jednotkami (dále VLJ) nazýváme ustálená spojení, která mají charakter vícečlenných, avšak významově celistvých lexikálních jednotek, reprodukovatelných pouze jako celek; o jejich celistvosti svědčí mj. to, že jejich členy nelze nahrazovat synonymy, antonymy, případně je rozvíjet[3] (srov. Karlík et al., 2012, s. 70). Představují významnou součást jazyka, přičemž mnoho z nich má v jazyce frekvenčně vysoké zastoupení. Z hlediska jazykového systému jsou zajímavé tím, že se nacházejí na pomezí slovníku a gramatiky (srov. například Mathesius, 1942, s. 88). Srovnáme-li své pojetí s funkčními typy víceslovných lexémů, které uvádí Čermák (2010, s. 222nn), pak my zahrnujeme mezi víceslovné lexikální jednotky víceslovné termíny, idiomatické kolokace a frekventované běžné uzuální kolokace. Stranou naopak ponecháváme víceslovná propria. V odborné literatuře se VLJ označují podle sledovaných důrazů a hledisek rozličně (například slovní spojení, sousloví, víceslovná pojmenování, víceslovné lexémy, srov. příslušnou terminologii užívanou v literatuře například in Hnátková et al., 2018).

Autorský tým vytvořil reprezentativní počítačovou databázi (slovník) víceslovných lexikálních jednotek v češtině, zvanou LEMUR; tato databáze obsahuje k červnu 2020 více než na 10 500 hesel a neustále roste. Jednotlivá hesla jsou popisována a klasifikována na základě podrobné typologie zachycující vlastnosti víceslovných lexi-

3 Někteří odborníci (např. Čermák, 2007, s. 263) ještě rozlišují *lexikální frazémy*. My je neuvádíme, neboť se zaměřujeme výhradně na jednotky víceslovné, tj. jednotky formálně tvořené více slovy.

kálních jednotek na různých jazykových rovinách (morfologie, syntax, sémantika, pragmatika, slovní zásoba).

V tomto textu nejprve popisujeme tuto typologii (kapitola 2), přičemž se zaměřujeme zejména na kategorie a jejich hodnoty, jež považujeme za hlavní: typ užití, syntaktický typ a různé typy idiomatičnosti VLJ (lexikální, morfologická, syntaktická, sémantická a pragmatická). V kapitole 3 pak sledujeme frekvenční zastoupení těchto kategorií a hodnot v reálných jazykových datech, zastoupených žánrově vyváženým korpusem češtiny SYN2015 o rozsahu 100 milionů slov (Křen et al., 2015).

2. TYPOLOGIE VÍCESLOVNÝCH LEXIKÁLNÍCH JEDNOTEK A OBSAH HESLA V DATABÁZI

Při stanovování typologie českých víceslovných lexikálních jednotek jsme vycházeli z návrhu klasifikace VLJ obsaženého v článku Baldwin et al. (2010) a z návrhu projektu PARSEME,[4] navazujícího na uvedený článek. Autoři kategorizují VLJ podle tří hlavních kritérií, kterými jsou:

(a) syntaktická struktura
(b) ustrnulost/flexibilita
(c) idiomatičnost (včetně nepravidelností ve VLJ).[5]

Toto třídění jsme přijali a rozšířili:

(i) se zřetelem k osobitým vlastnostem češtiny, kterými jsou zvláště morfologická idiomatičnost, daná velmi bohatou a spletitou morfologií češtiny, a syntaktická specifika, zejména volný slovosled
(ii) se zřetelem k tomu, že databázová hesla jsou užitečná pro lidského uživatele a dají se rovněž vhodně využít v počítačových aplikacích.

Rozšíření zahrnuje tyto údaje a vlastnosti: definici, charakteristické příklady, typ užití, valenční vzorec, fragmenty a varianty VLJ, styl/varietu VLJ a rovněž podrobnější charakteristiku některých typů (zvláště těch, jež jsou typické pro češtinu).

V popisu předpokládáme platnost standardních gramatických pravidel češtiny a zachycujeme jen odchylky od těchto pravidel s výjimkou zachycení valence, u níž nezachycujeme pouze valenční anomálie, nýbrž i standardní valenční struktury.

4 Srovnej http://typo.uni-konstanz.de/parseme.

5 Ve svém pojetí chápeme pojem *idiomatičnost*/*idiomaticita* jako „obecnou vlastnost" frazému, která se vztahuje nejen k sémantice, ale i k dalším rovinám (lexikální, morfologické, syntaktické, pragmatické). V české frazeologické literatuře se v tomto obecném smyslu hovoří o *anomálii* (srov. Čermák, 2007) a idiomatičnost je pak omezena pouze na sémantiku.

2.1 OBSAH DATABÁZOVÉHO HESLA

U hesla víceslovné lexikální jednotky v databázi zachycujeme tyto údaje:

- *lemma*, reprezentující a jednoznačně identifikující VLJ jak v databázi, tak v anotovaném korpusovém textu
- *superlemma*, jednoznačně identifikující skupinu nějakým způsobem příbuzných VLJ, identifikovaných svými lemmaty
- *lemmata* a *morfologické vlastnosti* jednotlivých komponent (slov) VLJ
- *definice,* slovně, neformalizovaně objasňující význam VLJ
- prototypické *příklady*
- *styl/varieta*, popisující VLJ podle stylu/registru
- *typ užití*, popisující VLJ z tradičního frazeologického hlediska
- *syntaktická struktura*, charakterizovaná těmito údaji:
  - syntaktický typ, který kategorizuje VLJ podle syntaktické kategorie celé jednotky
  - základní strukturní vzorec jako posloupnost kódovaných slovních druhů vyvozená ze syntaktického stromu VLJ
  - syntaktický strom závislostní i frázový (bezprostředněsložkový), obsahující rovněž příslušné syntaktické funkce
  - valenční vlastnosti celé VLJ i jejích jednotlivých slov (hlavně sloves a adjektiv)
- *ustálenost/flexibilita,* vystižená těmito údaji:
  - lexikální a další varianty VLJ, neboť VLJ se zdaleka nemusí vyskytovat pouze ve své standardní podobě
  - fragmenty standardní VLJ
  - slovosledné zvláštnosti VLJ
  - vnitřní modifikovatelnost, tj. možnost syntakticky rozvíjet nějakou z komponent VLJ
  - syntaktické transformace: (ne)možnost nominalizace, (ne)možnost adjektivizace, (de)pasivizace
  - morfologická omezení jako morfologické zvláštnosti jednotlivých komponent (slov) VLJ
- *idiomatičnost*, dále členěná na idiomatičnost
  - lexikální: monokolokabilita, negativa tantum, cizojazyčné výpůjčky…
  - morfologickou: morfologicky nestandardní tvary
  - syntaktickou: anakolut, atrakce, aposiopeze…
  - sémantickou: ta odráží míru významové kompozicionálnosti VLJ
  - pragmatickou: užití VLJ ve specifické situaci
  - statistickou: týká se uzuálních kolokací majících nadprůměrnou frekvenci a nápadně omezenou kolokabilitu.

Nyní popíšeme jednotlivé sledované údaje a pojmy podrobněji a rovněž je osvětlíme příklady.[6]

6 Všechny příklady uvedené v tomto článku jsou převzaty z lexikální databáze LEMUR nebo z korpusu SYN2015 a SYN v8 či ze slovníkových zdrojů, zejména ze Slovníku české

2.1.1 LEMMA

Lemma, tvořené posloupností slovních tvarů dané víceslovné lexikální jednotky, reprezentuje a jednoznačně identifikuje tuto jednotku jak v databázi, tak v anotovaném korpusovém textu, přičemž řídící člen je v základním tvaru, ostatní členy jsou vyjádřeny svým tvarem. Například lemmatem VLJ *dostali přes kušnu* je *dostat_přes_kušnu*, kde syntakticky řídícím členem je infinitiv *dostat*. U některých VLJ se řídící člen vyskytuje vždy ve tvaru, který se liší od tvaru základního, a v tomto případě je i řídící člen v tomto tvaru: *padni_komu_padni*.

2.1.2 SUPERLEMMA

Superlemma jednoznačně identifikuje skupinu nějakým způsobem příbuzných víceslovných lexikálních jednotek identifikovaných svými lemmaty. Například VLJ *bojovat pro čest a slávu* má lemma *bojovat_pro_čest_a_slávu*, ale superlemma *čest_a_sláva*; VLJ *dostat ultimátum* i *dát ultimátum* jsou zastřešeny jediným superlemmatem *dát_ultimátum*.

2.1.3 DEFINICE

Definice slovním, neformalizovaným způsobem objasňuje lidskému uživateli význam víceslovné lexikální jednotky. Například víceslovná lexikální jednotka *dát někomu do nosu/po nose* je podle Slovníku české frazeologie a idiomatiky (SČFI, Čermák et al., 1983–2009) definována takto:

> Zvl. silný čl. vůči slabšímu, aby zastavil jeho nepříjemné chování, neodbytnost n. čl. omylem při náhodném pohybu vůči druhému ap.: dát někomu silnou ránu do nosu, udeřit někoho do nosu, popř. do obličeje.

Víceslovná lexikální jednotka *na blind* je podle SČFI definována takto:

> 1. Bez zevrubnějšího prohlížení n. bez dívání se vůbec 2. bez urč. záměru, úmyslu, účelu, cíle 3. bez předběžné (důkladnější) přípravy, bez zajištění obvyklých průvodních okolností.

Je-li k dispozici, přebírá se definice z dostupných slovníkových zdrojů (v drtivé většině z SČFI). Pokud tato definice z nějakého důvodu nevyhovuje nebo v dostupných zdrojích schází, je v hesle uvedena další/nová definice, reflektující rovněž významové posuny plynoucí z úzu, jak se obráží v korpusových datech.

frazeologie a idiomatiky (SČFI, Čermák et al., 1983–2009) a též ze Základního slovníku českých přísloví (Čermák, 2013).

2.1.4 PŘÍKLADY

Prototypické větné příklady, vzaté z korpusů SYN2015 nebo SYN v8 (srov. Křen et al., 2015; 2019), osvětlují význam víceslovné lexikální jednotky a doplňují tak její definici.

2.1.5 STYL/VARIETA

Styl/varieta popisuje jak celou víceslovnou lexikální jednotku, tak její jednotlivé komponenty (slova) z hlediska stylu/registru; hodnoty komponent se přitom mohou lišit od hodnoty přiřazené celé VLJ. Například ve VLJ *klepat kosu* jsou všechna slova spisovná, zato VLJ jako taková je charakterizována jako kolokviální. Naopak VLJ *bejt vyhublej na kost* obsahuje kolokviální slovní tvary *bejt* a *vyhublej*, celá VLJ je však označena jako expresivní.

Víceslovné lexikální jednotce přiřazujeme tyto stylové hodnoty (zde i u dalších kategorií jsou příklady převzaty z lexikální databáze LEMUR, ze SČFI (Čermák et al., 1983–2009) a z korpusů SYN2015 a SYN v8):

— *spisovný*: např. *vládnout mistrně perem*; *být upoután na lůžko*; *být navýsost spokojený*
— *kolokviální*: např. *mouchy snězte si mě*; *něco mi do toho vlezlo*; *co ty na to jako svazák*
— *nářeční*: např. *bejt jako Boží vědro*; *náčiňová hadra*
— *expresivní*: např. *špinit si s* [něčím/někým] *ruce*; *obrátit oči v sloup*; *vytřískat ze všeho / z* [něčeho] *peníze*; *mít po srandě*
— *slangový*: např. *kroutit míč / pálit do šibenice* (např. v kopané); *spadl mu praporek* (překročil v šachu časový limit a prohrál); *udělat špagát* (ve slangu gymnastickém); *dát zkoušku* (ve slangu studentském)
— *jiný*: například knižní výrazy, archaismy, biblismy, antická a jiná rčení: *odejít / vrátit se do lůna Abrahamova*; *prodat* [někoho] *za třicet stříbrných*; *zlomit nad* [někým] *hůl*; *jidášský groš*; *nic lidského mi není cizí*; *peleš lotrovská*; *překročit Rubikon*; *království za koně*.

2.1.6 TYP UŽITÍ

Typ užití víceslovné lexikální jednotky vychází z lingvistického pojmenování obvyklého v české odborné literatuře[7] (srov. Čermák, 2016b) a navazuje na dělení v SČFI (Čermák et al., 1983–2009); uživateli pomáhá v celkové orientaci. Některé typy jsou zavedeny nově (*cizojazyčné spojení* a *otevřený frazém*), neboť máme za to, že mohou uživateli přinést novou, zajímavou informaci. Rozlišujeme tyto hodnoty:

— *přísloví*: podle paremiologa Wolfganga Miedera je to „krátká, všeobecně známá věta lidové slovesnosti vyjadřující lidovou moudrost, pravdu, morální maximu či tradiční názor. Je vyjádřena metaforickou, ustálenou a zapamatovatelnou formou a je předávána z generace na generaci“ (Mieder, 2012, s. 394). Příklady: *Koho chleba jíš, toho píseň zpívej.*; *Kdo jinému jámu kopá, sám do ní padá.*; *Čím hloupější sedlák, tím větší brambory.*; *Mráz kopřivu nespálí.*; *Proti gustu žádnej dišputát.*

7 Některá tradiční označení jako rčení a pořekadlo neužíváme pro jejich nejasné vymezení.

— *pranostika*: meteorologické přísloví, tj. v podobě větné vyjádřená souvislost mezi roční dobou, zpravidla vázanou na svátky, typem počasí a (doporučovanou) venkovskou polní a jinou prací, činností ap. (srov. Čermák, 2016c). Příklady: *Svatá Alžběta se sněhem přilétá.*; *Na svatého Jiří vylézají hadi a štíři.*
— *přirovnání*: ustálená VLJ, typicky tvořená slovesnou nebo adjektivní frází obsahující komparátor. Na subjekt této fráze v konkrétním užití se nekladou žádné omezující podmínky(srov. Čermák & Hladká, 2016). Příklady: *vyvádět jako smyslů zbavený*; *být papežštější než papež*; *hulit/kouřit jako lokomotiva*; *líný jako veš*
— *citace*: část jiného textu uvedená obvykle v doslovném znění. Bývá převzata z literatury či z jiného média (film, divadlo): *Svět chce být klamán, ať je tedy klamán.* (Lat. *Mundus vult decipi, ergo decipiatur.*); *Knihy mají své osudy.* (Lat. *Habent sua fata libelli.*); *Všechno své nosím s sebou.* (Lat. *Omnia mea mecum porto.*); *Tak pomíjí světská sláva.* (Lat. *Sic transit gloria mundi.*); *Kostky jsou vrženy.* (Lat. *Alea iacta est.*); *Hliník se vodstěhoval do Humpolce*.[8]
— *cizojazyčné spojení*: kolokace převzatá z cizího jazyka (typicky z latiny, angličtiny, němčiny, francouzštiny) a objevující se v českém textu: *mutatis mutandis*; *hora ruit*; *libri prohibiti*; *alma mater*; *perpetuum mobile*; *raison d'être*; *by the way*
— *termín*: „odborný název, technický termín, terminus technicus jsou výrazy označující jazykový výraz (slovo nebo sousloví), který má v určitém oboru, řemesle či povolání specifický, ostře vymezený význam, často odlišný od základního a obecného významu, nebo obecně odborné označení v daném oboru užívané" (srov. https://cs.wikipedia.org/wiki/Term%C3%ADn).[9] My pracujeme samozřejmě pouze s víceslovnými termíny, tedy „souslovími", nikoli s termíny jednoslovnými (srov. též Kováříková, 2017, s. 8nn). Příklady: *státní/nestátní aktér*; *diferenciální rovnice*; *reliktní záření*; *synoptická evangelia*; *bělásek zelný*; *Turingův stroj*
— *víceslovné synsémantikum*: patří sem zejména víceslovné předložky (*bez ohledu na*; *za účelem*; *na rozdíl od*), víceslovné spojky (*až na to, že*; *i když*), víceslovné částice (*tak či onak*; *tím spíše*; *kdesi cosi*), víceslovná citoslovce (*basta fidli*; *ajta krajta*), ale i kombinace jiných synsémantik, např. *beze všeho*; *svůj k svému*; *přece jen* (tzv. gramatické frazémy, srov. Čermák, 2007)
— *(nespecifický) slovesný frazém*: sémanticky nekompozicionální VLJ obsahující slovesný tvar jako řídící: *na tom nesejde*
— *neslovesný frazém*: sémanticky nekompozicionální VLJ neobsahující sloveso: *básnická žíla*; *něžné pohlaví*
— *kvazifrazém*: verbonominální spojení abstraktního substantiva s několika slovesy, která lze zařadit do široce chápaných obecných fází: inchoativní, durativní a terminativní (srov. Čermák et al., 1994, s. 26nn; Štícha et al., 2013, s. 428): *věnovat pozornost*; *vzbudit/strhnout/upoutat pozornost*; *vznášet obvinění*
— *větný frazém*: frazém jiný než přísloví, pranostiky a citace: *povídali, že mu hráli*; *co tě nemá; já vím*; *co ty na to?*; *děkuji pěkně*; *promiňte, že ruším*

8 Příklad *Hliník se vodstěhoval do Humpolce* pochází z filmu *Marečku, podejte mi pero* (1976, režisér Oldřich Lipský).

9 Další víceméně synonymní definice: „Verbální označení obecného pojmu v určitém oboru" (ČSN ISO 1087-1, 2002); „Název (lexém) užívaný jednoznačně pro denotáty specifické v určité vědě, oboru, ale i řemeslu či speciálním povolání" (Čermák, 2010, s. 132).

— *otevřený frazém*: VLJ, která vyžaduje nějaké pokračování, zpravidla se jedná o rutinní formulace uvádějící text nebo konverzaci (srov. Coulmas, 1981; Aijmer, 1996), které bývají dále rozvíjeny; v jiném pojetí bývají zahrnuty pod pojmy, jako jsou *formulae*[10] nebo *set phrases* (srov. Moon, 2012), například: *v okamžiku, kdy…*; *nezbývá než…*; *pokud vím,…*; *je někomu jasné, že…*; *představte si, že…*; *abych pravdu řekl…*; *ale k věci…*; *ani se nenaděješ…*
— *kolokace* (běžné kolokace uzuální): *akademická diskuse*; *adresná kritika*. Viz podrobněji odstavec 2.1.10.6 (Idiomatičnost statistická).

2.1.7 SYNTAKTICKÁ STRUKTURA

Popis syntaktické struktury zahrnuje údaj o syntaktickém typu, základní strukturní vzorec, závislostní i frázový (bezprostředněsložkový) syntaktický strom a valenční vlastnosti VLJ jako takové i jejích relevantních komponent (slov).

2.1.7.1 SYNTAKTICKÝ TYP

Syntaktický typ představuje syntaktickou kategorii celé víceslovné lexikální jednotky ve frázovém (bezprostředněsložkovém) pojetí. Zejména podle syntakticky řídícího členu syntaktické struktury (tedy podle formy, nikoli podle funkce) rozlišujeme tyto kategorie:

— *jmenná fráze* (*nominální skupina*): např. *svatý klid*;[11] *čirá náhoda*; *časová tíseň*; *nezvratné přesvědčení*; *uznávaná veličina*; *úřední šiml*; *svatá prostota*; *utahování opasků*; *štika soutěže*; *hlava děravá*
— *adjektivní fráze*: např. *draze zaplacený*; *přísně tajné*; *všeho schopný*; *níže uvedený*
— *slovesná fráze plnovýznamová*: např. *neprodat svou kůži lacino*; *nabrat* [někoho] *na rohy*; *sehrát*/*hrát roli mouřenína*; *vypadnout z role*; *splnit* [něco] *do puntíku*; *dát si nohu za krk*; *mít hodně co dohánět*
— *slovesná fráze s lehkým (kategoriálním) slovesem*: např. *přinést oběť*; *věnovat úsilí*; *dělat drahoty*; *dát na srozuměnou*; *dostat vynadáno*; *projevit účast*; *mít zaděláno* [na něco]; *vzít zavděk*; *vzít v potaz*; *nést odpovědnost*; *mít tušení*; *mít* [na někoho] *pifku*; *činit* [někoho] *odpovědným*; *vznášet obvinění*; *mít obavu*
— *adverbiální fráze*: např. *mírnyx tírnyx*;[12] *znovu a znovu*; *volky nevolky*; *časně zrána*
— *předložková fráze*: např. *po přeslici*; *o překot*; *na věky věků*; *s lehkým srdcem*; *ve vší počestnosti*; *ve vší tichosti*; *za tímto účelem*; *pro pána jána*; *na jeden zátah*[13]
— *složená předložka (předložkový výraz)*: např. *ve vztahu k*; *v souladu s*; *ve srovnání s*; *v rozporu s*; *se zaměřením na*

10 Srov. „Formule: obvykle ustálená lexikální kombinace vstupující do textu jako hotová, relativně frekventovaná a známá jednotka s výraznou komunikativní funkcí; zvláště specifického kontaktu n. aktu a často i vyvolání žádoucího cíle, účinku“ (Čermák, 2016a).

11 Podtrženě je vyznačen syntakticky řídící člen víceslovné lexikální jednotky.

12 U této VLJ a dalších dvou nevyznačujeme řídící člen.

13 Například zde se projevuje zaměření našeho přístupu na formu, nikoli na funkci (srovnej však odst. 2.1.8): např. VLJ *za tímto účelem* klasifikujeme jako předložkovou frázi, nikoli jako frázi adverbiální (tj. nikoli jako adverbiale okolnostního určení).

— *složená spojka (spojková skupina)*: např. *buď — anebo*; *i když*; *čím — tím*; *místo aby*; *ergo kladívko*[14]
— *složené citoslovce*: např. *ajta krajta*; *basama fousama*; *basta fidli*[15]
— *klauze (větná jednotka)*: např. *stokrát nic umořilo osla*; *pravda vítězí*; *každý provaz má dva konce*; *devátá rozhodne*; *škoda je jen dobrýho člověka*; *s tím je teď utrum*
— *souvětí*: např. *zákony jsou od toho, aby se porušovaly*; *mluviti stříbro, mlčeti zlato*
— jiný: např. *no jasně*.

2.1.7.2 ZÁKLADNÍ STRUKTURNÍ VZOREC

Základní strukturní vzorec víceslovné lexikální jednotky je odvozen ze syntaktického stromu této VLJ, který je vytvořen automatickou syntaktickou analýzou (parsing) a následně manuálně zkontrolován. Vzorec má podobu posloupnosti rozšířených slovnědruhových kódů, např.:

mít z ostudy kabát: (Verb — Prep.gen — Subst.gen — Subst.acc)
udělat [něco] *bez říkání*: (Verb — Prep.gen — Subst.gen)
ztratit [za někoho] *dobré slovo*: (Verb — Adj.Acc — Subst.acc)

2.1.7.3 ZÁVISLOSTNÍ A FRÁZOVÝ STROM

Víceslovná lexikální jednotka je vyjádřena dvěma druhy syntaktických stromů:

(i) stromem závislostním (podle formátu užívaného na tzv. analytické rovině Pražského závislostního korpusu, srov. Hajič, 2006)
(ii) stromem bezprostředněsložkovým (frázovým).

V obou druzích stromů jsou rovněž vyznačeny syntaktické funkce. Stromy vznikají automatickou syntaktickou analýzou a jsou poté manuálně zkontrolovány.

Příklad. Víceslovná lexikální jednotka

(i) *vrátit se do lůna Abrahamova*

je níže zachycena:

— jako linearizovaný *závislostní strom* (ii) a také jako jeho grafické znázornění (Obrázek 1)
— jako *složkový strom* (iii) a také jako jeho grafické znázornění (Obrázek 2).

(ii) (*vrátit*.Pred (*se*.AuxT *do*.AuxP (*lůna*.Adv (*Abrahamova*.Atr))))

14 U uvedených VLJ nevyznačujeme řídící člen.
15 U uvedených VLJ nevyznačujeme řídící člen.

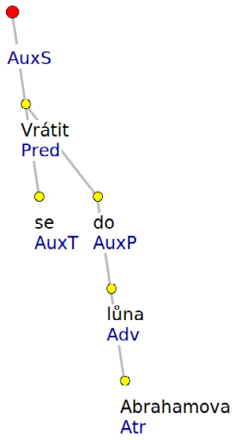
(iii) [vrátit.HD se.ReflTant [do.sHead [lůna.HD Abrahamova.attrPlain].dHead] .AdvbPlain]

V závislostním stromě symbolizuje AuxS celou větu, funkce Pred vyjadřuje hlavní predikát, AuxT označuje formální užití *se/si* u reflexiv tantum (AuxT), AuxP označuje předložku, Adv adverbiale (*do lůna* je zde totiž jako celek funkčně chápáno jako adverbiale) a Atr označuje atribut substantiva *lůna*.

Ve složkovém stromě označuje HD (head) hlavní syntaktický prvek složky, SH (sHead) označuje povrchovou syntaktickou hlavu (Surface Head), DH (dHead) označuje hloubkovou syntaktickou hlavu (Deep Head), advbPlain označuje adverbiale a attrPlain atribut.

Závislostní strom:

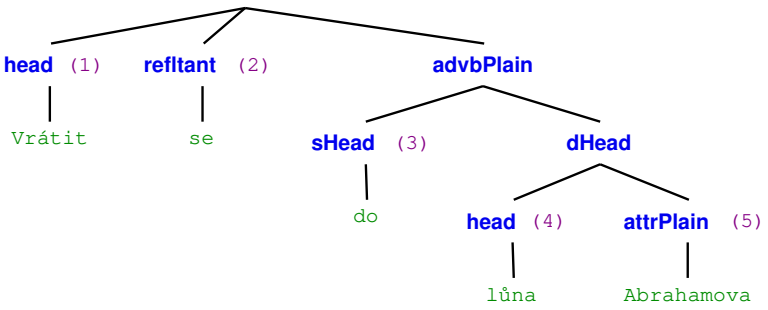
Vrátit se do lůna Abrahamova



OBRÁZEK 1: Grafické znázornění závislostního stromu.

Složkový strom:

```
Vrátit se do lůna Abrahamova
```



OBRÁZEK 2: Grafické znázornění složkového stromu.

2.1.7.4 VALENCE

Heslo víceslovné lexikální jednotky v lexikální databázi LEMUR obsahuje valenční údaje přiřazené jednak VLJ jako takové, jednak jednotlivým relevantním lexémům ve VLJ (především lexémům slovesným, adjektivním a substantivním), přičemž se nezachycují pouze valenční anomálie, nýbrž i standardní valenční struktury (valenčně očekávané argumenty a adjunkty s očekávanými morfosyntaktickými vlastnostmi). Valenční údaje se dají vyvodit ze syntaktických stromů. Upozorňujeme na to, že valence celé VLJ a její syntaktické hlavy může být odlišná. Například ve VLJ *věnovat úsilí*, kde sloveso *věnovat* je tzv. lehké (kategoriální) sloveso, je hloubkověsyntaktický Patiens *úsilí* slovesa *věnovat* součástí VLJ a — na rozdíl od nepřímého dativního objektu — není přítomen ve valenčním rámci VLJ jako takové.

2.1.8 SÉMANTICKÝ TYP ADVERBIÁLNÍ

U adverbiálních VLJ rozlišujeme základní sémantické kategorie:

— místní určení: *na pokraji*; *do popředí*; *směrem do*; *po vlastech českých*
— časové určení: *na poslední chvíli*; *dnem i nocí*
— určení způsobu: *po vzoru* [někoho]; *v neposlední řadě*; *u vědomí* [čeho]
— okolnostní určení: *u/při příležitosti*.

2.1.9 USTÁLENOST/FLEXIBILITA

Víceslovné lexikální jednotky se nemusí vyskytovat jen ve standardní podobě, mohou být totiž různě modifikovány a mohou také vykazovat jistá specifická omezení. V této části se budeme stručně zabývat lexikálními/morfologickými variantami VLJ (odst. 2.1.9.1) a částmi standardních VLJ (odst. 2.1.9.2); dále se budeme věnovat slovosledným specifikům (odst. 2.1.9.3), omezením možností rozvíjet (modifikovat) jednotlivé komponenty (slova) VLJ (odst. 2.1.9.4), omezením možností některých syntaktických transformací (odst. 2.1.9.5) a omezením morfologickým (odst. 2.1.9.6). O dalších specifických vlastnostech VLJ pojednáváme v části 2.1.10 (Idiomatičnost).

2.1.9.1 VARIANTY

Některé komponenty víceslovné lexikální jednotky se mohou vyskytovat ve variantních podobách ať už lexikálních, nebo morfologických, například:

— ve standardní VLJ *bojovat o čest a slávu* může být sloveso *bojovat* nahrazeno i dalšími slovesy: *zápolit*, *zápasit*, *válčit*, *hrát*…; ke standardní VLJ *utéct hrobníkovi z lopaty* existuje i lexikální varianta *uprchnout hrobníkovi z lopaty*
— ve standardní VLJ *překročit Rubikon* může být sloveso i ve své nedokonavé podobě: *překračovat Rubikon*; ke standardní VLJ *hodit flintu do žita* existují i vidové varianty: *házet/zahodit/zahazovat flintu do žita*.

Synonymní varianty téže VLJ (projevující se typicky v synonymii jednotlivých lexémů) pojímáme vskutku jako varianty VLJ, nevytváříme pro ně samostatná hesla. VLJ lišící se konverzivností predikátů vedeme v databázi jako samostatná hesla spjatá až na úrovni tzv. superlemmat: *dát ultimátum* vs. *dostat ultimátum* (srovnej výše odst. 2.1.2).

2.1.9.2 FRAGMENTY VÍCESLOVNÉ JEDNOTKY

Víceslovné lexikální jednotky se nemusí vyskytovat jen ve své úplné kanonické podobě, ale i v podobě pouhých částí/fragmentů. Například standardní VLJ *až naprší a uschne* se může v textech vyskytnout v podobě pouhého fragmentu: *až naprší*.

Tato dvě slova jsou charakteristická (jádrová) slova uvedené víceslovné lexikální jednotky *až naprší a uschne* a lexikální databáze by měla umožňovat, aby se v textech daly nacházet i takovéto torzovité frazémy (podrobněji v článku Jelínek et al., 2018). Fragmenty se formálně vyjadřují prostřednictvím identifikátorů jednotlivých slov a (pod)struktur.

2.1.9.3 SLOVOSLED

Standardní slovosledné vlastnosti a omezení v databázovém hesle příslušné víceslovné lexikální jednotky nevyznačujeme, explicitně uvádíme pouze slovosledné anomálie. Například ve VLJ *v tomto/pravém/dobrém/nejlepším slova smyslu* je tvar *slova* preponovaným genitivním přívlastkem tvaru *smyslu*, což je syntaktická anomálie, neboť standardně se řídící substantiva rozvíjejí postponovaným, nikoli preponovaným genitivním přívlastkem.

2.1.9.4 OMEZENÍ VNITŘNÍ MODIFIKOVATELNOSTI

Vycházíme z toho, že obecně lze komponenty (slova) víceslovné lexikální jednotky rozvíjet. Například ve VLJ *nést zodpovědnost* je jistě možné rozvíjet substantivum *zodpovědnost* (*nést* velkou/značnou/malou… *zodpovědnost*). Je-li známo lexikální/morfologické omezení při rozvíjení určitého slova, uvádíme je v databázovém hesle příslušné VLJ. Pokud naopak nějaké slovo ve VLJ nelze rozvíjet, pak na to u daného slova v databázovém hesle VLJ upozorňujeme: například ve VLJ *být na štíru* [s něčím] nelze rozvíjet slovo *štíru*.

2.1.9.5 TRANSFORMACE

V lexikálních heslech popisujeme explicitně jen ty struktury a vzorce, jež představují anomálie z hlediska standardní gramatiky češtiny. Evidujeme zejména tyto transformace: (ne)nominalizaci, (ne)adjektivizaci a (de)pasivizaci.

(Ne)nominalizace:

— Předpokládáme, že standardně je nominalizace možná: *věnovat pozornost → věnování pozornosti*; *ztratit pozornost → ztráta pozornosti*; *nést odpovědnost → nesení zodpovědnosti*; *dávat pokyny → dávání pokynů*; *hrát si s ohněm → hra s ohněm*.

— Někdy je však nominalizace vyloučena: *dostat za vyučenou* → **dostání za vyučenou*; *zasloužit na zadek* → **zásluha na zadek*.

(Ne)adjektivizace:

— Předpokládáme, že standardně je adjektivizace možná: *dávat pokyny* → *dávající pokyny*; *stát za starou belu* → *stojící za starou belu*; *dostat za vyučenou* → *dostávající za vyučenou*.
— Někdy je však adjektivizace vyloučena: *kape ti na karbid* → **kapající ti na karbid*.

(De)pasivizace:

— Některé VLJ nemohou mít pasivní podobu: taková je např. VLJ *Nevědomost hříchu nečiní.*, byť sloveso *činit* je přechodné.
— Některé VLJ naopak nemohou mít podobu aktivní: rčení (citaci) *Svět chce být klamán*. nelze převést do aktiva, přestože samo tranzitivní sloveso *klamat* pochopitelně může být, ba dokonce i zhusta bývá v aktivu.

2.1.9.6 MORFOLOGICKÁ OMEZENÍ

Specifická morfologická omezení (např. v čase, vidu, způsobu sloves; pádě, jmenném rodě, čísle jmen) se zachycují u všech morfologicky anomálních komponent příslušné víceslovné lexikální jednotky. Omezení mohou vyplývat ze syntaktické funkce slova (subjekt jména je jen v nominativu či genitivu), nebo mohou být dána sémantikou či územ. Například ve VLJ *ber, kde ber* mohou být výskyty slovesných tvarů pouze v imperativu 2. os. sg., ve VLJ *širší než delší* mohou být obě adjektiva pouze v komparativu, přičemž tvar komparativu *delší* je užit nesystémově.

Evidujeme i morfologické zvláštnosti v užití stylových variant, například obecněčeské tvary obligatorně přítomné ve víceslovné lexikální jednotce. Například VLJ *basama fousama* obsahuje obecněčeský tvar *fousama*, který v této VLJ nelze nahradit spisovným tvarem *fousy*.

2.1.10 IDIOMATIČNOST

Pojmem idiomatičnost vyjadřujeme stupeň anomálnosti víceslovné lexikální jednotky na různých jazykových rovinách: lexikální, morfologické, syntaktické, sémantické, pragmatické.

Idiomatičnost VLJ se podle článku Baldwin et al. (2010, s. 4) „vztahuje k příznakovosti / odchylce od základních vlastností jednotlivých lexémů tvořících VLJ […] Daná jednotka je často idiomatická na více rovinách […] V úzkém vztahu k pojmu idiomatičnosti je kompozicionalita, kterou chápeme jako stupeň, v jakém se vlastnosti částí VLJ spojují, aby se daly předpovědět vlastnosti celku. Zatímco kompozicionalita se často chápe tak, že se týká výhradně sémanticky idiomatických jednotek (výrazem ‚nekompozicionální VLJ' mají proto odborníci obvykle na mysli sémanticky idiomatické jednotky), v praxi ji lze uplatňovat na týchž rovinách jako

idiomatičnost.“ V souladu s uvedeným přístupem rozlišujeme následující typy idiomatičnosti:

2.1.10.1 IDIOMATIČNOST LEXIKÁLNÍ

Lexikálně idiomatické jsou ve víceslovné lexikální jednotce tyto druhy slovních tvarů a/nebo lexémů:

— *monokolokabilní slovní tvary* (srov. i Čermák, 2014): například *vzít zavděk*; *do třetice*; *mírnyx týrnyx* (s variantami *mírnix dýrnix*, *mírnix týrnix* ad.); *křížem krážem*, *vyjít najevo*; *tma tmoucí*; *na přeskáčku*
— *téměř monokolokabilní slovní tvary*: např. tvary, které se spojují jen s velmi omezenou množinou slov: *žabomyší spory*; *učinit* [něčemu] *přítrž*; *chléb vezdejší*; *jáma lvová*; *stojí to za starou belu*; *zavdat důvod/příčinu*; *utkvělá myšlenka/představa*; *stanné právo*
— *slovní tvary a lexémy mající jen zápornou podobu (negativa tantum)*: např. *jen ať se nepotento*; *nepřeberné množství*; *nezadatelné právo*; *neodolatelná touha*
— *výpůjčky z cizích jazyků*: např. *ad infinitum* (z latiny); *raison d'être* (z francouzštiny)
— *makarónská struktura*: VLJ složená ze slov různých jazyků, například *baj voko* — anglická předložka *by* či německá předložka *bei* a české substantivum *voko*; další příklad: *je to v oukeji*, kde *oukeji* je přizpůsobený tvar anglického *OK*; *per hubam* — latinská předložka *per* a české substantivum *huba* s latinským pádovým morfem *-m*, který v latině vyjadřuje akuzativ feminina singuláru první deklinační třídy
— *jiná lexikální idiomatičnost*.

2.1.10.2 IDIOMATIČNOST MORFOLOGICKÁ

Morfologicky idiomatické jsou nestandardní/nesystémové či neobvyklé tvary s nesystémovými nebo řídce užívanými gramatickými morfy vyskytující se pouze v příslušné víceslovné lexikální jednotce. Například v přísloví *Podle nosa poznáš kosa*. je použit nestandardní tvar Gsg *nosa* s morfem *-a* (oproti systémovému *nosu*). Ve VLJ *v kuse* je nestandardní morf *-e* (oproti systémovému *kusu*).

2.1.10.3 IDIOMATIČNOST SYNTAKTICKÁ

Tento typ idiomatičnosti zachycuje syntaktické nepravidelnosti celé víceslovné lexikální jednotky. Rozlišujeme tyto hodnoty:

— *anakolut*: například v pozměněném novozákonním přísloví: *Kdo po tobě kamenem, ty po něm chlebem.*
— *atrakce*: např. *v řadě případech* (tvar *případech* nesystémově atrahuje lokál od tvaru *řadě*); *padni, komu padni*; *stůj co stůj* (zvláštní druhý výskyt tvaru *padni*, resp. *stůj*, je dán totožností s výskytem prvním); *lidé jsou různé* (tvar *různé* atrahuje morf *-é* od *lidé*)
— *zvláštní valence*: například v archaických biblismech: *očekávat na Hospodina*; nebo užití dnes již ze systémového hlediska zastaralého genitivu záporového u nego-

vaných tranzitivních sloves právě ve VLJ: *nemám námitek*; *není divu*; nebo výjimečná akuzativní valence slovesa *stát*: *stát čestnou stráž*. V náboženském diskurzu se často užívá 2. os. sg. imperativu pro vyjádření 3. osoby imperativu: *chraň tě ruka Páně*, *pozdrav Pán Bůh*, navíc lexémy *ruka* a *Pán Bůh* jsou v nominativu.

— *aposiopeze*: odmlka typu *ty jsi teda…*; *já se na to…*
— *elipsa*: výpustka nějakého slova: *nevím, co* [mám dělat] *dřív*; *mladost* [je] *radost*; termín *zobrazení na*, kde schází jméno (= *množinu/množině*) po předložce[16]
— *slovosled*: například adjektivum z hlediska gramatického systému výjimečně (zato však v daných VLJ obligatorně) následuje své řídící substantivum: *mše svatá*; *hlava děravá*; *Duch svatý*; *chléb náš vezdejší*; *liška podšitá*; *mládež školou povinná*; [film] *mládeži nepřístupný*
— *jiná syntaktická idiomatičnost*: například *kluk šikovná*; *chlap mizerná*. Ve VLJ *kdo chce kam, pomozme mu tam* je nadřazená klauze *pomozme mu tam* rozvita vztažnou klauzí *kdo chce kam*, kde se nacházejí dva spojovací výrazy vyjadřující vztažnost: *kdo* a *kam*, přičemž antecedentem vztažného zájmena *kdo* je v nadřazené klauzi *mu*, antecedentem příslovce *tam* je zájmenné příslovce *kam* — je to v každém případě konstrukce syntakticky dost neobvyklá.
— *syntakticky neidiomatické*: VLJ mající standardní syntaktickou strukturu.

2.1.10.4 IDIOMATIČNOST SÉMANTICKÁ

U víceslovných lexikálních jednotek rozlišujeme míru sémantické kompozicionality (= sémantické neidiomatičnosti/nemetaforičnosti), tj. míru toho, nakolik lze odvodit význam VLJ z jejích komponent (slov), a to na škále:

— *nekompozicionální* (= vždy idiomatické, označovaná situace nemůže v reálném světě nikdy nastat): např. *to pravé ořechové*; *chladný počtář*; *hotový poklad*; *udělat* [něco] *levou zadní*; *mít ocelové nervy*; *vymřít po meči*; *mít z ostudy kabát*; *viset* [někomu] *na rtech*; *boží mlýny*; *zaječí úmysl/úmysly*
— *zřídka kompozicionální* (= zhusta idiomatické): např. *jestřábí oko*; *kočičí hlavy*; *strouhat* [někomu] *mrkvičku*
— *často kompozicionální* (= málo idiomatické): např. *černočerná tma*; *ve vší tichosti*; *sloní tlapa*; *holý zadek*
— *vždy kompozicionální* (= neidiomatické, doslovné): např. *neodolatelná touha*; *doba kamenná*; *dobrá investice*.

2.1.10.5 IDIOMATIČNOST PRAGMATICKÁ

Pragmaticky idiomatické jsou takové víceslovné lexikální jednotky, které se užívají pouze ve zvláštních situacích, například: *smím prosit?*; *no a co?*.

16 *Zobrazení na* je matematický termín, jeho synonymem je *surjekce*. Znamená takové zobrazení *f* z množiny A do množiny B, kde každý prvek množiny B má v definici zobrazení *f* svůj vzor v definičním oboru tohoto zobrazení.

2.1.10.6 IDIOMATIČNOST STATISTICKÁ

V databázi LEMUR evidujeme i *uzuální kolokace*, které jsou typicky sémanticky neidiomatické, vyznačují se nadprůměrnou frekvencí a nápadně omezenou kolokabilitou — a v tomto smyslu jsou anomální, „statisticky idiomatické". Komponenty takové kolokace nelze bezvýhradně nahradit synonymy: *živelní pohroma* vs. **veliká pohroma* / **pohroma živlů*; *očistná lázeň* vs. **očišťující lázeň*; *oslovská lavice* vs. **trestná lavice*,[17] *v dezolátním stavu* (omezená kolokabilita). Kolokace může být sémanticky kompozicionální a zpravidla v době svého vzniku i bývá, ale jedná se o hraniční kategorii na pomezí frazeologie a právě statistická idiomatičnost může naznačovat posun v kolokabilitě. Komponenty se sice objevují v různých kolokacích, ale kolokace zařazené do databáze jsou častější. Z hlediska typu užití jsou kromě „kolokace" statisticky idiomatické:

— termíny: např. *vektorový prostor*; *mnichovská dohoda*
— víceslovná synsémantika:
  - složené předložky: *s ohledem na*
  - složené spojky: *i když*.

3. FREKVENČNÍ ZASTOUPENÍ HODNOT HLAVNÍCH KATEGORIÍ VLJ V REÁLNÝCH JAZYKOVÝCH DATECH

Po podrobně popsané typologii se nyní zaměřme na to, jaké hodnoty hlavních kategorií se objevují v korpusu, tedy jak jsou klasifikované frazémy zastoupeny v reálných textech.

Přestože víceslovné lexikální jednotky dosud zachycené v databázi LEMUR představují jen část VLJ přítomných v jazyce jako takovém (databáze LEMUR se neustále doplňuje), přistoupili jsme už nyní k frekvenčnímu ověření zachycovaných vlastností u dosud popsaných jednotek. Tato sonda nám umožňuje nahlédnout alespoň v hrubých obrysech systém jazyka v oblasti VLJ prizmatem jejich frekvenční distribuce podle sledovaných vlastností a naznačit směr, kudy by se měl výzkum dále ubírat. V něm pak bude vhodné se nejprve zaměřit na frekventované typy a u nich případně uvažovat o jejich jemnější subklasifikaci. Sonda nám také poskytne zpětnou vazbu, zda jsou sledované vlastnosti vhodné pro zařazení do databáze, případně zda tam některé frekventované typy nechybí a zda je zapotřebí je do databáze doplnit. V budoucnu by mohlo být zajímavé mezijazykové nebo mezižánrové srovnání.

Jako datovou základnu jsme zvolili žánrově vyvážený korpus SYN2015 (Křen et al., 2015) obsahující 122 mil. tokenů (100 milionů výskytů slovních tvarů a 22 milionů výskytů interpunkčních znamének), který se z hlediska jazykového úzu pokládá za žánrově vyvážený: obsahuje ve stejném početním zastoupení

17 *Trestná lavice* je ovšem také neidiomatická uzuální kolokace, užívaná však výhradně v hokejovém diskurzu a nesynonymní s uzuální kolokací *oslovská lavice*.

(i) texty beletristické
(ii) texty oborové literatury
(iii) texty publicistické.

V této sondážní fázi výzkumu frekvenční distribuce typů VLJ jsme zatím nezkoumali, od jakých autorů či z jakých (typů) publikací VLJ v korpusu pocházejí, ani jsme se nezabývali korelací typů VLJ a žánrové skupiny (i)–(iii). Jakmile bude databáze co do počtu zachycených VLJ výrazně reprezentativnější, bude vhodné přikročit k výzkumu tohoto typu.

Anotace VLJ bude ostatně součástí budoucích korpusů, a bude tedy možné, aby si uživatel sám zjistil informace o jejich distribuci v příslušných textech.

Na materiálu 9745 zpracovaných víceslovných lexikálních jednotek (jsou rozděleny na základní úrovni, tj. nejsou v nich např. sdruženy synonymní podoby) sledujeme tyto kategorie:

— typ užití (srov. výše část 2.1.6)
— syntaktický typ (srov. výše část 2.1.7.1)
— idiomatičnost
  - lexikální (srov. výše část 2.1.10.1)
  - morfologická (srov. výše část 2.1.10.2)
  - syntaktická (srov. výše část 2.1.10.3)
  - sémantická (srov. výše část 2.1.10.4).

Četnosti vlastností VLJ sledujeme z hlediska typů, nikoli tokenů (!): každou VLJ bereme jako jeden element, v tomto příspěvku nás nezajímá četnost dané VLJ v korpusu.

Vzhledem k tomu, že zjištěné údaje považujeme za pouhou první frekvenční sondu do celé problematiky, a vzhledem k tomu, že lexikální databáze LEMUR se postupně buduje (jakkoli je dnes již plně funkční a naplněná více než deseti tisíci hesly) a reviduje (i co do klasifikace ve směru jejího zjemňování), nepokládali jsme za nezbytné již v této fázi ručně značkovat data nezávisle více anotátory. Šlo spíše o ujasnění celkové klasifikace a její rozumné jemnosti (nemá nyní tedy ani smysl vyhodnocovat přiřazování vlastností mírami *pokrytí* (*recall*) a *přesnost* (*precision*)).[18] Jakmile bude databáze obsahovat výrazně větší počet hesel VLJ, a bude tedy reprezentativním modelem jazyka, bude náležité přikročit ke klasickému ručnímu značkování VLJ více anotátory a na základě výsledků takového značkování popis VLJ v databázi mj. dále zpřesnit.

18 K dispozici je pouze dílčí ověřování zaměřené na somatické frazémy v mluveném korpusu. Bylo provedeno na původní frazémové anotaci, která je východiskem i pro naši databázi (Kopřivová, 2015). Ukázalo se, že použitá metoda má vyšší *precision*, neboť se snaží vyhýbat frazémům, které mohou mít doslovný význam.

3.1 TYP UŽITÍ A SYNTAKTICKÝ TYP

V Tabulce 1 níže zachycujeme vztah mezi typem užití (řádky) a syntaktickým typem (sloupce).

Z hlediska typu užití upozorňujeme na tyto údaje v Tabulce 1:

1. Převážná část přísloví je tvořena klauzemi (*Cesta do pekla je dlážděná dobrými úmysly.*), občas souvětími (*Dvakrát měř, jednou řež.*). Mezi „jiné" jsme zařadili přísloví neobsahující finitní sloveso, sloveso je v nich elidováno (*Hlas lidu, hlas boží.*).
2. Pranostik je v našem vzorku velmi málo, v korpusových datech je jich patrně více. Není vyloučeno, že pranostiky se často uvádějí v podobě fragmentů, a proto nebyly rozpoznány.
3. Přirovnání jsou typicky slovesná (*čučet jako tele na nová vrata*; *zařvat jako tur*), bývají však i adjektivní (*zelený jak sedma*), adverbiální (*jako žába na prameni*)[19] a jmenná (*zima jako v morně*) nebo jsou tvořena klauzí (*jde to jak na drátku*).
4. Citace jsou tvořeny klauzí (*Karty jsou rozdány.*) či souvětím (biblismus *Hospodin dal, Hospodin vzal.*), někdy jsou i jmenné (*Království za koně!*).
5. Cizojazyčná spojení jsme přiřadili k syntaktickému typu „jiné" (*ad multos annos*).
6. Nepřekvapí, že termíny mají vesměs podobu jmenné fráze (*babočka admirál*; *aktivní saldo*).
7. Víceslovná synsémantika jsou syntakticky spojkové skupiny (*a když už*; *buď — anebo*) nebo skupiny adverbiální/částicové (*a tak podobně*) či citoslovečné (*ach tak*; *a hergot*).
8. Ze syntaktického hlediska řadíme slovesné frazémy pochopitelně k slovesným frázím (*dělat ve dne v noci*; *naservírovat až pod nos*), přičemž vydělujeme konstrukce s lehkým (kategoriálním) slovesem (*brát zřetel*; *mít tušení*).
9. Neslovesné frazémy jsou syntakticky pochopitelně především jmenné fráze (*Adamovo žebro*; *básnické střevo*), mnohem méně pak fráze předložkové (*bez bázně a hany*) či fráze adverbiální (*ani vidu ani slechu*).
10. Kvazifrazémy mají samozřejmě slovesnou syntaktickou strukturu, jde o kombinace slovesa s abstraktem (*věnovat pozornost*), přičemž zvlášť jsou vyděleny konstrukce s lehkým slovesem (*budit respekt*; *dostat zálusk*; *držet se zásady*; *ztratit vědomí*).
11. Větné frazémy (jiné než přísloví, pranostiky, přirovnání a citace) jsou tvořeny téměř vždy klauzemi (*Absolutní moc korumpuje absolutně.*).
12. Mezi tzv. otevřené frazémy nalezené v korpusu SYN2015 patří například tyto klauze: *aby bylo jasno*; *aby toho nebylo málo*; *teď si představte, že*; *dlužno přiznat, že*.
13. Uzuální kolokace jsou typicky jmenné fráze a předložkové fráze, v menší míře fráze adverbiální: *aktivní odpočinek*; *bez nároku na*.

19 Jako adverbiální chápeme toto přirovnání v kontextech, v nichž je užito jiného slovesa než *sedět*, např.: *Škodí zdravotnictví jako žába na prameni*.

↓ Typ užití	Syntaktický typ												celkem
	nom	**adj**	**vrb**	**cat**	**adv**	**prep**	**prep_sl**	**cnj_sl**	**int_sl**	**cls**	**cmp**	**jiné**	
1. přísloví										**321**	15	14	350
2. pranostika										5	2		7
3. přirovnání	85	**236**	**1910**		**343**					49	3		2626
4. citace	10		2			1				**51**	**30**		94
5. cizojazyčné spojení												21	21
6. termín	**163**												163
7. víceslovné synsémantikum					20*			**40**	11				71
8. slovesný frazém			**3280**	120									3400
9. neslovesný frazém	**1512**	20			**91**	**280**		4	4				1911
10. kvazifrazém			101	20									121
11. větný frazém										**315**	2		317
12. otevřený frazém										40			40
13. kolokace	**379**	5			**64**	**173**		3					624
celkem	2149	261	5293	140	518	454		47	15	781	52	35	9745

TABULKA 1: Vztah mezi typem užití víceslovné jednotky a jejím syntaktickým typem.
Záhlaví sloupců: nom = jmenná fráze, adj = adjektivní fráze, vrb = slovesná fráze plnovýznamová, cat = slovesná fráze s lehkým (kategoriálním) slovesem, adv = adverbiální fráze, prep = předložková fráze, prep_sl = složená předložka (předložkový výraz), cnj_sl = složená spojka (spojková skupina), int_sl = složené citoslovce, cls = klauze, cmp = souvětí.
* Víceslovnou lexikální jednotku, jež je tvořena skupinou částic, zahrnujeme mezi víceslovná adverbia.

Z pohledu syntaktického typu vidíme, že:

— jmenné fráze jsou hlavně neslovesné frazémy, kolokace a termíny
— adjektiva a adverbia se využívají především v přirovnáních
— slovesnými frázemi se vyjadřují hlavně přirovnání a slovesné frazémy
— předložky jsou především součástí neslovesných frazémů a kolokací
— klauzemi se vyjadřují přísloví, citace a přirovnání
— souvětím se vyjadřují citace a také přísloví.

3.2 LEXIKÁLNÍ IDIOMATIČNOST

Vybrali jsme pět typů užití relevantních z hlediska lexikální idiomatičnosti: přísloví, přirovnání, slovesné frazémy, neslovesné frazémy a kolokace. V drtivé většině VLJ se lexikální idiomatičnost projevuje přítomností (téměř) monokolokabilních slov, což je právě jeden z typických rysů VLJ.

↓ Druhy lexikální idiom.	**Relevantní typy užití**					**celkem**
	přísloví	**přirovnání**	**slovesný frazém**	**neslovesný frazém**	**kolokace**	
obsahuje monokol. slovo	5	10	51	60	3	129
obsahuje téměř monokol. slovo	15	47	209	85	11	367
obsahuje cizí výpůjčku			2	8	1	11
makarón. struktura						
jiná lex. idiomat.						
celkem	20 z 350*	57 z 2626	262 z 3400	153 z 1911	15 z 624	507

TABULKA 2: Korelace lexikální idiomatičnosti a hlavních typů užití.

* Míní se zde 20 z celkového počtu 350 přísloví, srov. tabulku 1 výše. Podobně je nutno interpretovat údaje v dalších sloupcích tohoto řádku: 57 z celkového počtu 2626 přirovnání, 262 z celkového počtu 3400 slovesných frazémů, 153 z celkového počtu 1911 neslovesných frazémů, 15 z celkového počtu 624 kolokací.

Tabulka 2 obsahuje frekvenční údaje týkající se korelace lexikální idiomatičnosti a hlavních typů užití. Vysvítá z ní, že monokolokabilní slovní tvary/lexémy[20] (níže vyznačené podtržením) jsou přítomné zvláště v neslovesných frazémech (*baba princ-metálová*; *hokus pokus*; *do třetice všeho dobrého i zlého*; *Pandořina skříňka*; *ani za zlámanou grešli*; *léta letoucí*) a frazémech slovesných (*dát do pucu*; *dát na srozuměnou*; *dělat si šoufky*), někdy i v přirovnáních (*koukat/tvářit se jako kakabus*; *jako na zavolanou*) a příslovích (*Co se stalo, nedá se odestát.*; *Jez do polosyta, pij do polopita.*; *Pozdě bycha honit.*).

20 Čermák chápe monokolokabilnost jako kombinaci tvaru/lexému s až devíti komponenty (Čermák, 2014).

Jako téměř monokolokabilní slovní tvary/lexémy označujeme elementy s výrazně omezenou kolokabilitou, která je však vyšší než u monokolokabilních tvarů/lexémů. Je jich pochopitelně více než ryze monokolokabilních a vyskytují se hlavně ve slovesných frazémech (*ani ve snu se nenadát*; *ani nepáchnout*; *vzít si čas na rozmyšlenou*; *hodit se / dát se do gala*; *být fuk*; *být k pohledání*; *být na maděru*; *být na mizině*), ale i frazémech neslovesných (*Daniel v jámě lvové*; *bez cavyků*; *bez skrupulí*; *dvojsečná zbraň*). Lze na ně však narazit také v přirovnáních (*křičet jako pominutý*; *oblečený jako hastroš*), příslovích (*Práce kvapná málo platná.*; *Chudoba cti netratí.*) a kolokacích (*do popředí*).

V tabulce zvlášť neuvádíme řádek s počty VLJ obsahujících negativa tantum, neboť všechny takové výskyty jsou přiřazeny k VLJ obsahujícím ryze či téměř monokolabilní tvary: 17 slovesných frazémů (např. *být k nepotřebě*; *být k nerozeznání*; *div si voči nevykoukat*; *a to ani nemluvě*), 4 neslovesné frazémy (např. [něco] *do nepohody*; *kouzlo nechtěného*) a jedno přirovnání (*chovat se jako neotesanec*).

3.3 MORFOLOGICKÁ IDIOMATIČNOST

Případů morfologické idiomatičnosti bylo nalezeno velmi málo, uveďme tu pouze jednotlivé případy:

— příslovíˇ: *podle nosa poznáš kosa*; *dobrého pomálu*
— přirovnání: *slzy jako hrachy*
— neslovesná: *jakýs takýs*

3.4 SYNTAKTICKÁ IDIOMATIČNOST

Ve zkoumaném vzorku byly nalezeny tyto druhy syntaktické idiomatičnosti:

— elipsa (typicky elipsa sponového *být*)[21] — je hojná v příslovích (116 výskytů), jež co nejúsporněji a nejpregnantněji vystihují lidovou moudrost: *Bližší košile než kabát.*; *Bez božího požehnání marné lidské namáhání.*; *Co na srdci, to na jazyku.*; *Co Čech to muzikant.*; *Světská sláva polní tráva.*; *Člověk člověku vlkem*; *Devět řemesel — desátá bída*. Dále se vyskytuje v pranostikách: *Březen, za kamna vlezem*; ve slovesných frazémech: *v nejlepším přestat*; *jak řečeno*; *jak vidno*; v neslovesných frazémech a kolokacích: *všeho s mírou*; *těžko na cvičišti, lehko na bojišti*; *škoda každé rány* (s pokračováním *která padne vedle*); *blahoslavení chudí duchem*; *co platno*
— zvláštní valence v příslovíˇ: *Poturčenec horší Turka.*, kde upozorňujeme na zvláštní genitivní valenci vyjadřující srovnání (v češtině výjimečný genitivus comparationis)
— jiná syntaktická idiomatičnost: vyskytuje se například v neslovesném frazému *kluk ušatá*. Tvar feminina Nsg *ušatá* sám o sobě není nijak zvláštní, je jenom příznakově užit v expresivním vyjádření na úkor shody v rodě ve jmenné skupině, a proto chápeme celou syntaktickou jmennou skupinu jako syntakticky idiomatickou.

21 Struktury s elidovanou sponou je podle našeho názoru vhodnější popisovat právě jako eliptické a nepopisovat je z hlediska lexikálně přítomného jmenného přísudku.

Jiným podtypem je také užití imperativu 2. os. sg. ve funkci imperativu 3. os. sg., nechápeme-li ovšem níže uvedené (a podobné) případy jako pozůstatky imperativu 3. osoby. V datech jsme nalezli tyto příklady frazémů z náboženského diskurzu: *Co Bůh spojil, člověk nerozlučuj.*; *Děj se vůle Boží/Páně.*; *dejžto pánbůh*; *Chraň Tě ruka Páně*.
Mezi slovesnými frazémy je zajímavá VLJ *stát čestnou stráž/stát frontu*, kde sloveso *stát* je valenčně uspokojeno předmětem v akuzativu (sic!), což je u slovesa *stát* v daném významu (nikoli např. ve významu *stát* o ceně) zcela výjimečné.

3.5 SÉMANTICKÁ IDIOMATIČNOST

Opět jsme vybrali pět typů užití relevantních z hlediska lexikální idiomatičnosti víceslovné lexikální jednotky: přísloví, přirovnání, slovesné, neslovesné frazémy a kolokace a zkoumáme u nich míru sémantické idiomatičnosti podle škály navržené v odstavci 2.1.10.4. Připomeňme, že na této škále rozlišujeme míru sémantické idiomatičnosti (sémantické nekompozicionality) VLJ:

— *vždy idiomatická*: VLJ nelze nikdy chápat doslova, tj. sémanticky kompozicionálně
— *zřídka neidiomatická*: VLJ lze jen málokdy chápat doslova, tj. VLJ je málokdy sémanticky kompozicionální
— *často neidiomatická*: VLJ lze často chápat doslova, je tedy často sémanticky kompozicionální
— *vždy neidiomatická*: VLJ je vždy sémanticky kompozicionální.

Tabulka 3 níže ukazuje výraznou převahu vždy idiomatických (sémanticky nekompozicionálních, metaforických) VLJ oproti ostatním druhům na škále ve všech sledovaných typech užití (sloupce). Zvlášť výrazné je to — jak lze ostatně očekávat — u přirovnání. Nutno ovšem mít na paměti, že přiřazování dané VLJ k tomu či onomu typu je někdy (možná často) arbitrární, zvláště k typům uprostřed škály: zřídka neidiomatická a často neidiomatická.

↓ Druhy sémantické idiom.	**Vybrané typy užití**					**celkem**
	přísloví	**přirovnání**	**slovesný frazém**	**neslovesný frazém**	**kolokace**	
vždy idiom.	292	2431	2879	1354	0	6956
zřídka neidiom.	40	129	77	109	0	355
často neidiom.	16	31	173	119	0	339
neidiomat.	2	35	313	329	624	1303
celkem	350	2626	3442	1911	624	8953

TABULKA 3: Korelace míry sémantické idiomatičnosti a vybraných typů užití.

Uveďme nyní charakteristické příklady na jednotlivé typy:

vždy sémanticky idiomatické:

— přísloví: *Jak se do lesa volá, tak se z lesa ozývá.*; *Kdo šetří, má za tři.*; *Každý pes jiná ves.*
— přirovnání: *přetrhnout jako hada*; *připadat si jako Alenka v říši divů*; *sedět jako mumie*; *růst jako dříví v lese*; *mít se jako ve vatičce*; *mít peněz jako železa*; *mít nervy jako špagáty*; *jako na potvoru*; *rozbrečet se jako želva*; *vypadat jako strašák do zelí*
— slovesný frazém: *doslova šílet*; *být z obliga*; *loupnout očima/pohledem*; *lámat si s/nad* [něčím] *hlavu*; *měřit dvojím metrem*; *obrátit v prach a popel*; *obrnit se trpělivostí*; *nést kůži na trh*
— neslovesný frazém: *do morku kostí*; *do nebe volající*; *do očí bijící*; *hlava skopová*; *ostřílený kozák*

zřídka neidiomatické:

— přísloví: *Každá hůl má dva konce.*; *Kdo chce psa bít, hůl si najde.*
— přirovnání: *působit jako magnet*; *rovný jako mlat*; *vypadat jako šílenec*
— slovesný frazém: *nebýt z cukru*; *přinést až pod nos*; *obrátit list*; *obletět svět*
— neslovesný frazém: *otrávený šíp*; *nepřítel lidstva*; *stará bačkora*; *spojenými/společnými silami*; *tichá domácnost*

často neidiomatické:

— přísloví: *Co se škádlívá, rádo se mívá.*; *Kdo neokrádá stát, okrádá rodinu.*
— přirovnání: (*mít*) *nohy jako hůlky*; *jako na pouti*
— slovesný frazém: *mýt si ruce*; *nést ovoce*; *octnout se na ulici*
— neslovesný frazém: *sluha/služebník Boží*; *spavá nemoc*; *stočený/á do klubíčka*; *uzavřená společnost*

neidiomatické:

— přísloví: *Co Bůh činí, dobře činí.*; *Každý svého štěstí strůjcem.*; *Každý začátek je těžký.*; *Kdo mlčí, souhlasí.*
— přirovnání: *přibývat jako hub po dešti*; *píchat jako ježek*; *mít oči jako kočka*; *rozplynout se jako mlha*
— slovesný frazém: *dospět ke konci*; *obrátit se* [na někoho / k někomu] [*s prosbou*] *o pomoc*; *obstát se ctí*; *obětovat život*
— neslovesný frazém: *do dnešního dne*; *otevírací doba*; *osudová/osudná chyba*; *spotřební daň*; *spravedlivá odplata*
— kolokace: *absolutní ticho*

3.6 SHRNUTÍ — VĚROHODNOST VÝSLEDKŮ

Na základě poměrně velkého vzorku 9745 víceslovných lexikálních jednotek obsažených v žánrově vyváženém korpusu SYN2015 jsme se pokusili ukázat frekvenční zastoupení hlavních typů užití VLJ ve vztahu k jejich syntaktickému typu a dále k různým druhům idiomatičnosti: lexikální, morfologické, syntaktické a sémantické. Vzhledem k tomu, že korpus je žánrově vyvážený, považujeme zjištěné frekvenční poměry VLJ za poměrně relevantní, majíce za to, že — samozřejmě cum magno grano salis — jakožto docela slušné odhady zhruba vypovídají o frekvenčních poměrech v jazyce jako takovém; pro češtinu jde snad o první takový pokus. Přitom jsme si ovšem vědomi velkých úskalí, jež mohou výsledky našeho šetření zpochybňovat:

— Výběr víceslovných lexikálních jednotek ve vzorku je víceméně arbitrární, data jsme čerpali hlavně ze SČFI (Čermák et al., 1983–2009) a ověřovali jejich výskyt v korpusu.
— V korpusu SYN2015 jsou jistě i jiné VLJ než ty, jež jsme zahrnuli do svého šetření; identifikovat VLJ v textech je z povahy věci velmi obtížné a spojené s velkým objemem ruční práce, jakkoli se dají využít různé statistické kolokační míry; nicméně při vyhledávání například sémanticky idiomatických VLJ tyto míry příliš nepomohou. Dalším problémem je to, že delší VLJ bývají často uváděny jako fragmenty nebo jsou mezi ně vkládána další slova, a takováto VLJ lze pak obtížněji identifikovat — to se týká především pranostik a přísloví.
— Není vyloučeno, že některé VLJ častěji se vyskytující v úzu se neobjevují v korpusu SYN2015 (např. přísloví). Je však velmi pravděpodobné, že se budou vyskytovat v padesátkrát větším korpusu SYN v8. Další výzkum by tedy měl vycházet z dat tohoto velkého, ač nereprezentativního korpusu či korpusu podobné velikosti.
— Výše uvedená typologie je pouze jednou z možných typologií (naše typologie je v některých aspektech hlavně zaměřena na formu, nikoli funkci, například v syntaxi): některá rozlišení hodnot/druhů jsou případně příliš jemná, jiná hrubá, některé relevantní hodnoty/druhy možná scházejí; sám výběr sledovaných kategorií je do jisté míry arbitrární.

Velká většina víceslovných lexikálních jednotek zjištěných v korpusu SYN2015 je již obsažena v lexikální databázi LEMUR, anotace zbývajících VLJ je rozpracována. Jakmile budou tyto VLJ kompletně zpracovány, budou zařazeny do této databáze.

Přes uvedená úskalí máme za to, že výsledky jakožto kvalifikovaný odhad poskytují určitou představu o frekvenčním rozložení sledovaných vlastností VLJ z hlediska typů v datech korpusu, a tedy — vzhledem k žánrové vyváženosti korpusu — i v psaném jazyce. Bylo by navíc zajímavé porovnat výskyty VLJ a jejich vlastností v textech odlišných žánrů: v beletrii, oborové literatuře a publicistice.

4. ZÁVĚR

V tomto příspěvku autorský tým:

(a) představil podrobnou typologii víceslovných jednotek v současné češtině v lexikální databázi LEMUR a
(b) zkoumal zastoupení jednotlivých klasifikovaných kategorií a jejich hodnot v reálných jazykových datech žánrově vyváženého korpusu SYN2015.

V budoucnu se výzkumný tým hodlá zaměřit na:

— vyhledávání dalších víceslovných lexikálních jednotek v korpusových datech, anotaci těchto VLJ a jejich zařazování do databáze LEMUR
— opravy anotace hesel, jež v databázi již jsou
— zpřesňování představené typologie
— další práce na propojování VLJ v textových datech a v databázi LEMUR
— publikace dalších poznatků o vlastnostech VLJ, a to zvláště na datovém základě rozsáhlých korpusů typu SYN v8.

LITERATURA

Aijmer, K. (1996). *Conversational Routines in English: Convention and Creativity*. London: Routledge.

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, 2nd edn. (s. 267–292). Boca Raton: CRC Press.

Coulmas, F. (Ed.). (1981). *Counversational Routine. Explorations in Standardized Communication Situations and Prepatterned Speech.* The Hague: Mouton.

Čermák, F. et al. (1983–2009). *Slovník české frazeologie a idiomatiky* (*SČFI*), vol. 1–4. Praha: Academia/Leda.

Čermák, F. (2007). *Czech and General Phraseology*. Prague: Karolinum.

Čermák, F. (2010). *Lexikon a sémantika*. Praha: Nakladatelství Lidové noviny.

Čermák, F. (2013). *Základní slovník českých přísloví. Výklad a užití*. Praha: Nakladatelství Lidové noviny.

Čermák, F. (2014). *Periferie jazyka. Slovník monokolokabilních slov*. Praha: Nakladatelství Lidové noviny.

Čermák, F. (2016a). Formule. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *Nový encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny. Dostupné z https://www.czechency.org/slovnik/FORMULE

Čermák, F. (2016b). Frazeologie a idiomatika. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *Nový encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny. Dostupné z: https://www.czechency.org/slovnik/FRAZEOLOGIE%20A%20IDIOMATIKA.

Čermák, F. (2016c). Pranostika. In P. Karlík, M. Nekula, & J. Pleskalová, (Eds.), *Nový encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny. Dostupné z: https://www.czechency.org/slovnik/PRANOSTIKA

Čermák, F., & Hladká, Z. (2016). Přirovnání. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *Nový encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny. Dostupné z:https://www.czechency.org/slovnik/P%C5%98IROVN%C3%81N%C3%8D

ČSN ISO 1087-1 (2002). *Terminologická práce — Slovník — Část 1: Teorie a aplikace.* Dostupné z https://www.nlfnorm.cz/terminologicky-slovnik/39582

Hajič, J. (2006). Complex corpus annotation: The Prague Dependency Treebank. In M. Šimková (Ed.), *Insight into the Slovak and Czech Corpus Linguistics* (s. 54–73). Bratislava: Veda.

Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., & Vondřička, P. (2018). Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. *Korpus — gramatika — axiologie*, *17*, 3–22.

Jelínek, T., Kopřivová, M., Petkevič, V., & Skoumalová, H. (2018). Variabilita českých frazémů v úzu. *Časopis pro moderní filologii*, *100* (2), 151–175.

Karlík, P., Nekula, M., & Rusínová, Z. (Eds.) (2012). *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny.

Kopřivová, M. (2015). Evaluating automatic idiom annotation in spoken corpora: The case of somatic idioms. In K. Gajdošová & A. Žáková (Eds.), *Natural Language Processing, Corpus Linguistics, Lexicography* (s. 72–76). Bratislava: Slovenská akadémia vied..

Kováříková, D. (2017). *Kvantitativní charakteristiky termínů*. Praha: Nakladatelství Lidové noviny / Ústav Českého národního korpusu.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Dostupný z http://www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A. (2019). *Korpus SYN, verze 8 z 12. 12. 2019*. Praha: Ústav Českého národního korpusu FF UK. Dostupný z http://www.korpus.cz

Mathesius, V. (1942). O soustavném rozboru gramatickém. *Slovo a slovesnost*, *8*, 88–92.

Mieder, W. (2012). *Proverbs Are Never Out of Season: Popular Wisdom in the Modern Age*. New York: Peter Lang.

Moon, R. (2007). Corpus linguistic approaches with English corpora. In H. Burger, D. Dobrovol'skij, P. Kühn & N. R. Norrick (Eds.), *Phraseology. An International Handbook of Contemporary Research*. Berlin — New York: Walter de Gruyter.

Štícha et al. (2013). *Akademická gramatika spisovné češtiny*. Praha: Academia.

Vondřička, P. (2019): Design of a multiword expressions database. *Prague Bulletin of Mathematical Linguistics*, *112*, 83–101.

Vladimír Petkevič — Marie Kopřivová — Milena Hnátková — Tomáš Jelínek — Pavel Kopřiva — Alexandr Rosen — Hana Skoumalová — Pavel Vondřička
Ústav teoretické a komputační lingvistiky FF UK
Ústav Českého národního korpusu FF UK
<milena.hnatkova@ff.cuni.cz>
<tomas.jelinek@ff.cuni.cz>
<P.Kopriva@seznam.cz>
<marie.koprivova@ff.cuni.cz>
<vladimir.petkevic@ff.cuni.cz>
<alexandr.rosen@ff.cuni.cz>
<hana.skoumalova@ff.cuni.cz>
<pavel.vondricka@ff.cuni.cz>