O příbuznosti lingvistiky a biologie

Hana Owsianková — Ondřej Kučera — Dan Faltýnek[1]

ABSTRACT:
In this paper, we discuss the history and current state of research where linguistics and biology meet on the base of interdisciplinary approach. We would like to introduce the use of quantitative text features (entropy, vocabulary richness, etc.) and quantitative text laws (Zipf's law, Menzerath-Altman's law) in the analysis of genetic text. We point out research in which similar text features of natural languages are observed on the genetic text. In particular, we focus on the field of molecular phylogenetics, the subject of which is to investigate the relationship of species and the reconstruction of the evolutionary tree of organisms. In the conclusion, we present step-by-step genetic analysis of interspecific relationships of selected crops of the Brassicaceae family using the Bag-of-words model. Our aim is to stress that linguistic methods can be a good tool for gaining new knowledge about genetic texts, and we want to introduce a method for mapping evolutional relationships.

KEYWORDS:
Bag-of-words, biology, genetics, linguistics, molecular phylogenetics

ABSTRAKT:
V tomto článku se zabýváme historií a současným stavem výzkumu, v němž se potkávají biologie a lingvistika. Poukazujeme na výzkumy, při nichž jsou na genetickém textu pozorovány podobné vlastnosti jako na textech přirozených jazyků. Konkrétně se zaměřujeme na oblast molekulární fylogenetiky, jejímž předmětem zkoumání je rekonstrukce stromu příbuznosti organismů. V závěru v jednotlivých krocích prezentujeme genetickou analýzu mezidruhových vztahů vybraných kulturních plodin čeledi Brassicaceae (brukvovité) pomocí metody Bag-of-words. Našim cílem je ukázat, že lingvistické metody mohou být vhodným nástrojem k získání nových poznatků také o genetických textech, nově pak chceme představit metodu Bag-of-words využitelnou k mapování příbuzenských vztahů.

KLÍČOVÁ SLOVA:
Bag-of-words, biologie, genetika, lingvistika, molekulární fylogenetika

ÚVOD

Od přelomu osmnáctého a devatenáctého století, kdy byla vymezena jako vědní disciplína, se lingvistika ocitla hned několikrát v těsném vztahu k biologii, a to v souvislosti s analogizováním jazyka a živého a také z důvodu podobnosti metodologie obou disciplín, např. při popisu původu jazyků a druhů. Vzájemná inspirace obou disciplín byla v průběhu času obousměrná, historickosrovnávací lingvistika se např. inspirovala teoretickým rámcem evoluční teorie, genetika využila analogie s textem při popisu DNA. V tomto článku chceme shrnout vývoj kontaktu obou disciplín a upozornit na výhledy jejich dalšího setkávání.

1 Článek vznikl v rámci projektu Sinophone Borderlands Reg. no. CZ.02.1.01/0.0/0.0/16_019/0000791, Excellent research.

Na vztah biologie a lingvistiky bylo v lingvistice mnohokrát poukázáno. Na začátku připomeňme práci Augusta Schleichera (1869), který ovlivněn darwinismem považoval jazyk za vyvíjející se živý organismus. Schleicher ukázal, že darwinistický výklad evoluce založený na náhodné mutaci a přirozeném výběru se vztahuje nejen na organismy, ale také na jazyky. Tímto způsobem postavil na roveň přístup evoluční biologie a historickosrovnávací lingvistiky. O vlivu darwinismu na lingvistiku, přijetí prací Schleichera dobovými biology a také o „ztotožňování řeči s přírodninou" později referuje Emanuel Rádl (2006, s. 170–172; německý originál 1909). Rádl vztahy lingvistiky a biologie ovšem reflektuje z pozice dobové kritiky darwinismu a analogie předmětu studia biologie a jazykovědy pojímá v kontextu vitalistické teorie, která je vzhledem k dnešnímu všeobecnému akceptování Darwinovy teorie neakceptovatelná (vyjmeme-li z diskuse kreacionismus a dnes známou paletu evolučních mechanismů doplňující koncept přirozeného výběru). O vlivu darwinismu na českou lingvistiku píše Jan Petr (1982, 1988), všímá si především Jana Gebauera, který se s darwinismem setkal prostřednictvím svého učitele Martina Hattaly — stejně jako on pojetí jazyka jako organismu odmítá, a naopak považuje jazyk za historický společenský jev, u nějž je třeba zkoumat především jeho vnitřní psychickou stránku (k tomu viz Syllaba, 1983, s. 25–27).

Další výraznou inspiraci biologií můžeme nalézt u Pražského lingvistického kroužku (PLK). V jeho Tezích (Vachek, 1972, s. 37) nacházíme vliv nomogenetického pojetí vývoje jazyka. Kroužek se v tomto směru nechává inspirovat antidarwinistickou koncepcí zákonem řízené evoluce ruského teoretického biologa a ichtyologa L. S. Berga (1926) — teleologické pojetí jazyka se výrazně promítá do profilu pražského strukturalismu, který usiluje o strukturní (tzn. systémový) popis nejen synchronních, ale také diachronních jevů, čímž se odlišuje od ostatních strukturalismů, především ženevského (viz Kořenský, 2008; Leška, 1986; Němec, 1989; Savický, 1991; Šoltys, 1991; Vachek, 1968). Teleologický výklad se týká především vysvětlení hláskových změn (Vachek, 1968; pro ilustraci viz vysvětlení vzniku hlásek ť a ď vlivem fonologické pozice hlásky ň v Historické mluvnici češtiny, Lamprecht — Šlosar — Bauer, 1986, s. 88).

Jedna z nejvýraznějších osobností PLK, Roman Jakobson, znovu promlouvá o vztahu obou věd později, když analogizuje strukturu přirozeného jazyka a strukturu DNA (1971; k tomu viz Katz, 2008). Jakobson upozorňuje např. na dvojí artikulaci obou kódů — stejně jako jsou v přirozeném jazyce věty členěny do slov, a ty dále do morfémů/fonémů, které jakožto jednotky druhé artikulace nenesou význam, ale rozlišují jej, tak lze sekvence DNA členit na triplety a ty pak na jednotlivé nukleotidy, přičemž tripletům je přisouzen význam v podobě aminokyseliny a nukleotidy význam nemají, jejich funkcí je rozlišovat triplety. Jakobson dále připodobňuje řetězce DNA k textu z hlediska jejich linearity/sekvencionality. Svou teorii binárních opozic, aplikovanou např. na morfologický systém ruštiny (viz Jakobson, 1932), převedl na vztah protilehlých bází v DNA: cytosin — guanin, adenin — thymin (Jakobson, 1971, s. 678–681). Dokonce mluví o pružné stabilitě jazyka DNA (ibid., s. 681). Analogii DNA ve vztahu k jazyku a jeho projevech v textu dále rozvíjí např. Searles (2002) nebo Raible (2001).

Soustavně se historickým souvislostem vztahů biologie a lingvistiky věnoval Simeon Romportl (1989, 1994) — jazyk a genetický kód srovnává z hlediska komplex-

nosti jejich struktury, upozorňuje na jejich hierarchičnost (jazykové plány, struktura genetického kódu) a diskrétnost veličin reprezentovaných v textu (ve smyslu jazykových jednotek a de Saussurovy diskrétnosti jakožto vlastnosti jazykového znaku) (1989, s. 261). Romportl ukazuje na podobnosti a odlišnosti v evoluci jazyka a živého — podobnosti např. z hlediska mutace jakožto prostředku vývoje nových forem (Romportl mluví o transformacích jazykových a genetických prvků), odlišností např. z hlediska niterného charakteru genetického textu oproti komunikační funkci jazyka. V souvislosti s taxonomizací jazyků a druhů informuje o kladistice (viz níže) a ukazuje způsoby rekonstrukce genealogického stromu (1997, 1999).

Na druhou stranu čeští biologové vyhledávají kontakt s lingvistikou, např. v oblasti sémiotického popisu procesů v živých organismech kolem sebe ustavil celou školu Anton Markoš (2003; Markoš & Švorcová, 2009). Podle Markoše je textům přirozených jazyků a genetickým textům společné to, že se jedná o lineární záznam znaků (prvků) nesoucích informaci. Nezdůrazňuje ovšem samotnou povahu informace, ale především způsob, jakým je přečtena — hovoří o procesech interpretace v buňce, které formují výsledný fenotyp. V oblasti zoosémiotiky, při popisu systémů zvířecí komunikace, kriticky navazuje na Hockettův výčet charakteristických vlastností jazyka (1982, „*language design features*“) Lucie Čadková (2015) — na základě moderních poznatků o komunikaci zvířat upozorňuje především na to, že fyziologické vlastnosti organismu nelze pokládat za konstitutivní rysy jazyka a že k univerzáliím lidského jazyka nemůžeme přistupovat jako k univerzáliím jazyka obecně. Signalizační systémy mnohobuněčných bakteriálních konsorcií sloužící k výstavbě kolonií popisují Jaroslav Čepl et al. (2010).

Genetické predispozice jazyka (např. tzv. „jazykový gen“ FOXP2), jeho vznik a evoluci zkoumá biolingvistika, která má s českým prostředím také kontakt (Augustyn, 2013; Berwick & Chomsky, 2011; Kosta & Krivochen, 2012). Programem biolingvistiky je především nalezení genetických dispozic člověka, vysvětlení jejich vztahu ke kognitivnímu zpracování řeči, rekonstrukce vývoje schopnosti řeči u člověka a popis odlišnosti zvířecí komunikace oproti lidské (Berwick & Chomsky 2016).

TAXONOMIE

Z hlediska dnešních vztahů biologie a lingvistiky se nejdříve zastavíme u metod rekonstrukce vývoje organismů. Zkoumání příbuzenských vztahů je v současné biologii úkolem nejen morfologie, ale také molekulární fylogenetiky, která rekonstruuje genetický strom na základě analýzy dlouhých řetězců genetického zápisu organismů. Původní taxonomizace (Linnæi, 1758) je založena na srovnávání morfologických znaků, tzn. celkové vnější stavby organismu — u rostlin se jedná např. o tvar, velikost a barvu plodu, květu nebo listu, velikost a tvar semen, trichomů („chloupků“, výčnělků na pokožce rostlin) apod. K taxonomizaci jsou využívány také anatomické znaky, jako je stavba pletiv, např. typ cévních svazků, jejich počet apod. Zvažovány jsou také znaky chemické — obsah určité sloučeniny v organismu. Proti taxonomizaci organismů na základě podobnosti výše zmíněných znaků (dále tělních ve smyslu fenotypických) se od padesátých let 20. stol. staví kladistika, která taxonomizaci

provádí prostřednictvím rekonstrukce evolučních linií podobnosti (příbuznosti) znaků, na základě takového srovnání se pak evoluční biologové pokouší rekonstruovat průběh evoluce. To znamená, že kladistika rekonstruuje příbuznost linií na základě přítomnosti určitých znaků, které vyznačují evoluční oddělování skupin organismů (k tomu viz Hennig, 1966, 1975; Dupuis, 1984; Brinkman, 2001; ke kritice fylogenetické taxonomie viz Mayr, 1974; Wheeler, 2004). Taxonomizace fenotypická je postupně doplňována i metodami molekulární fylogenetiky (mezi něž patří i kladistika), která příbuznost druhů a jejich taxonomické zařazení zkoumá pomocí molekulárních znaků, tj. úseků DNA, RNA a proteinů.

Určování příbuznosti druhů na základě molekulárních znaků — podobnosti řetězců DNA, RNA a proteinů — přináší oproti určování příbuznosti na základě morfologických a dalších tělních znaků určité výhody. Tělní znaky se totiž vzájemně podmiňují (např. tvar semene, jeho chemické složení a tvar kořene (např. Stace, 1989)). Tělních znaků je oproti tomu, co lze získat z lineární posloupnosti genetických textů, omezené množství. Oproti morfologickým znakům zde mluvíme o tzv. genetických markerech — ty lze vymezit jako oblast genetické informace, kterou srovnávané organismy sdílí, a to i s diferencemi získanými v evoluci. Markery jsou oproti fenotypickým znakům proměnlivé spíše z hlediska toho, zda jsou brány z oblastí kódujících částí DNA, kde je jejich variabilita vázána na funkci řetězce, nebo z částí nekódujících — i v tomto případě je ale míra jejich proměnlivosti různá, vzhledem k tomu, že i nekódující části DNA mohou mít určitou funkci, např. v podobě regulace genové exprese. Molekulární fylogenetika (nikoliv pouze kladistika) umožňuje porovnávat evolučně velmi vzdálené organismy, jejichž tělní znaky nevykazují dostatečnou podobnost k vzájemnému srovnávání s ostatními, ale struktura jejich molekulárních znaků ano, nebo takové organismy, které mají velmi rozdílnou morfologii, ale jsou příbuzensky blízké (Flegr, 2005, s. 439–442; podrobněji Page & Holmes, 1998).

LINGVISTICKÉ METODY V ANALÝZE GENETICKÉHO TEXTU

Analýzou posloupnosti bází DNA či aminokyselin v proteinech neboli „genetických textů" se zabývá bioinformatika, jejíž metody srozumitelně i pro laického čtenáře přibližuje F. Cvrčková (2006). Vzhledem k tomu, že centrální úlohu v analýze genetických markerů hrají textualizované, tzn. jako texty zapsané, řetězce biopolymerů (proteinů, nukleových kyselin), se zároveň stává molekulární fylogenetika zájmem lingvistiky, která má s analýzou textu zkušenosti. Výzkumné metody biologie a lingvistiky jsou vzájemně přejímány. Např. kladistika jako molekulárně fylogenetická metoda je včetně dílčích postupů (např. Maximum parsimony, Maximum likelihood) postupně aplikována na analýzy slovesné kultury (Ross et al., 2013) a jazyka (Gray & Atkinson, 2003; Rexová et al., 2003, 2006; Forster & Renfrew, 2006; Grollemund & Hombert, 2012; Fangerau et al., 2013; Rabinovich et al., 2017). Naopak metody využívané současnou bioinformatikou často pocházejí z lingvistiky, jedná se např. o Damerau-Levenshteinovu vzdálenost (Damerau, 1964; Levenshtein, 1966).

Obecně je Damerau-Levenshteinova vzdálenost definována jako minimální počet transformací jednoho řetězce (textu) na druhý. Transformací se rozumí vložení,

odstranění a nahrazení části textu nebo změna její pozice v textu. Čím méně je třeba transformací od výchozího textu k srovnávanému, tím jsou si texty podobnější (Damerau, 1964; Levenshtein, 1966). Tato metoda byla navržena jako editační vzdálenost mezi texty — vyjadřovala množství ortografických změn, jimiž se dva texty odlišují. Použita byla také v oblasti lexikostatistiky (glottochronologie) v souvislosti s rekonstrukcí genetického stromu jazyků na základě analýzy výpůjček, k čemuž je využíván tzv. Swadeshův seznam (Swadesh, 1952, 1955; Embleton, 2000) — obsahuje slova, která jsou odolná vůči jazykovým výpůjčkám: osobní zájmena, části těla, zvířata, slovesa základních činností, barvy, číslovky „jedna" a „dvě", nebeská tělesa atd. Množství rozdílů mezi slovy určuje vzdálenost mezi jazyky ve stromu (metodu představuje např. Skalička, 1967; Králík, 1976; Serva & Petroni, 2007). V oblasti biologie jsou tímto způsobem srovnávány konkrétní sekvence DNA a na základě jejich míry shody je vyjadřována příbuznost jednotlivých druhů a jimi zastupovaných skupin organismů. Částmi řetězce jsou v případě analýzy genetických textů nukleové báze DNA nebo aminokyseliny proteinového řetězce. Dodejme, že kromě lingvistiky a biologie byla Damerau-Levenshteinova vzdálenost využita např. v medicíně pro výzkum evoluční trajektorie chronické lymfocytární leukémie (Sutton et al., 2014).

Další metodou, která je využívána pro analýzu podobnosti textů, je tzv. Bag-of--words model. Tato metoda je využívána v oblasti zpracování přirozeného jazyka (*natural language processing*), vyhledávání informací (*information retrieval*), počítačovém vidění (*computer vision*) a klasifikaci dokumentů (Toldo et al., 2009; Zhang et al., 2010). Pro analýzu genetických textů je převzata nověji (Bolshoy et al., 2010; Lovato, 2015). Metoda Bag-of-words spočívá v reprezentaci textu jeho slovy, bez ohledu na jejich pořadí v textu, ale se zohledněním jejich frekvence. Umožňuje vyhodnocovat podobnost lexika textů a z ní vyvozovat podobnost tematickou, obsahovou, stylovou či autorskou. V případě našeho zájmu se jedná o podobnost textů genetických, z nichž můžeme usuzovat na blízkost fylogenetickou (vývojovou, příbuzenskou). V případě textů genetických však nemáme přirozeně dány hranice slov, proto je k reprezentaci slova v genetickém textu vhodné využít n-gramovou analýzu. Při ní jsou genetické texty segmentovány na stejně dlouhé části (substringy): dvoj-kombinace bází nebo aminokyselin (2-gram), troj-kombinace bází nebo aminokyselin (3-gram) atd. Tento nepřirozený zásah do struktury genetického textu můžeme podpořit zkušeností s tím, že n-gramová analýza je citlivá na přirozené hranice jednotek (např. slov) a při vhodném užití velikosti n-gramu kopíruje výsledky analýzy přirozeně segmentovaného textu — dokládá to např. využití n-gramové analýzy při určování autorství, kdy tato technika dosahuje podobných výsledků jako určování autorství založené na přirozeně segmentovaném textu (Peng a kol., 2003; viz také Faltýnek, 2017, s. 72, kde jsou ukázány výsledky n-gramové analýzy textů srovnatelné s výsledky analýzy využívající dělení na slova). Každý organismus je v analýze reprezentován řadou hodnot vyjadřujících přítomnost či nepřítomnost určitého slova (substringu) v jeho genetické sekvenci, nebo rozdílem ve frekvenci daného slova oproti dalším sekvencím. Takto reprezentované sekvence DNA jsou následně použity ve shlukové analýze, která podobnosti textů zobrazuje v grafu hierarchického shlukování (dendrogramu).

V návaznosti na Damerau-Levenshteinovu vzdálenost dodejme, že Bag-of-words model byl využit ke klasifikaci dat biomedicínského inženýrství, jako jsou výsledky

vyšetření EEG nebo EKG (Wang et al., 2013). Jeho další aplikace slouží k detekci infekce v organismu na základě zastoupení různých tříd molekul T-receptorů (Lovato, 2015).

VYUŽITÍ KVANTITATIVNÍCH VLASTNOSTÍ TEXTU V POPISU GENETICKÉHO TEXTU

Lingvistické metody jsou v analýze genetických textů využívány v souvislosti s řadou dalších výzkumů, nejen pro účely taxonomické. Dlouhou tradici mají např. výzkumy Zipfova zákona (Zipf, 1949), který formuluje vztah mezi počtem výskytů slova a jeho distribucí v textu (obecněji v jazyce). Příkladem je jeho využití pro predikci funkce nekódující DNA (Niyogi & Berwick, 1995; Tsonis et al., 1997; Havlin et al., 1995; Mantegna et al., 1995) nebo k popisu hierarchie genetického kódu (ve smyslu skladby rovin jako je fonetická, lexikální a větná, viz Matlach & Faltýnek, 2016, Faltýnek et al., 2019). O projevech tohoto zákona v textu se dlouhodobě vedou diskuse, protože každý text, který je produktem libovolného systematického procesu, zákon projevuje: může se jednat o šifru, tzv. *monkey typing* — text vzniklý náhodnými úhozy do klávesnice, zápis činnosti nějakého technického zařízení (k tomu viz např. otázku projevů zákona v náhodném textu Ferrer-i-Cancho & Elvevåg, 2010; shrnutí současného stavu výzkumu Zipfova zákona a jeho výhledy předkládá Piantadosi, 2014).

Novější lingvistickou paralelu v souvislosti s DNA představuje analýza projevů Menzerath-Altmannova zákona v genetickém textu (Altmann, 1980). Verbální formulace zákona zní takto: čím větší je jazykový konstrukt, tím větší jsou jeho konstituenty ve své průměrné délce; jedná se např. o velikost věty ve vztahu ke slovům, slova ke slabikám apod. To ukazují např. Ferrer-i-Cancho et al. (2013); Hřebíček (2002, 2007) ukazuje, že Zipfův zákon je variantou Menzerath-Altmannova zákona. Studie se zaměřují především na vztah velikosti a počtu chromozomů v genomu (Baixeries et al., 2013; Hernández-Fernández et al., 2011), velikosti exonů a jejich počtu v genu (Li, 2012; Nikolaou, 2014), další výzkumy studují vztah velikosti proteinových domén a velikosti proteinu (Shahzad et al., 2015). Vliv segmentace textu na výsledky analýzy projevů Menzerath-Altmannova zákona ukazují Benešová, Faltýnek a Zámečník (2015) a Benešová (2014), v oblasti výzkumu genetického textu s tímto souvisí výběr srovnávaných jednotek (exon—intron, sekundární struktury apod.) a velikost n-gramů vstupujících do analýzy. Projevy těchto zákonů jsou většinou vykládány jako ekonomizační ve smyslu zákona nejmenšího úsilí (způsoby vysvětlení viz Piantadosi, 2014).

Jako příklad lingvistické analýzy DNA slouží další výzkumy: komplexitu genetického textu (pravděpodobnost výskytu za sebou jdoucích částí řetězce) zkoumá Popovová, Segal a Trifonov (1996). Trifonov také upozorňuje na možnosti zkoumání raných stádií vzniku genetického kódu (Trifonov et al. 2001), za pozornost stojí též jeho práce s Berezovským zaměřená na popis struktury proteinů analogicky ke struktuře jazykové (Trifonov & Berezovski, 2002). Zemková zkoumá využití lingvistických metod v genomice v širokém rozsahu např. na genetických textech parazitů (Zemková, 2016; Zemková et al., 2014). Ukazuje, že tyto texty mají vyšší míru opaku-

jících se struktur. Systematický přehled použití lingvistických metod v molekulární biologii předkládá Bolshoy (2003, 2010). Kvalitativní strukturou genomů a proteomů, např. přítomností textových motivů v rozsáhlých vzorcích proteinů, se zabýval Ohno (1992). Analogii mezi přirozenými jazyky a genetickým kódem, nikoliv ale na základě analýzy textu, popisuje také Ji (1997, 1999), Pattee a Kull (2009), Barbieri (2006, 2008, 2015), Favareau (2009) a Sharov (2010).

V souvislosti s českou lingvistikou stojí za zmínku, že analogii přirozeného jazyka a genetického kódu představil Jakobson (1971), o němž píšeme výše. Hovoří o bázích DNA jako o fonémech, o tripletech jako slovech a genech jako větách, mluví ale také o synonymii tripletů, fonologických opozicích bází, pružné stabilitě genetického kódu atd., což působilo v diskusi s biology určité rozpaky (Jacob et al., 1968). Na problémy spojené s analogizováním jazyka a genetického kódu upozorňuje Markoš a Faltýnek (2011) — viz tradiční pojetí bází jako písmen, genů jako vět. Na základě experimentu založeném na mapování Zipfova zákona na mRNA diskutuje Matlach a Faltýnek (2016), základní předpoklady znakových vlastností genetického kódu ukazují Lacková, Faltýnek a Matlach (2017).

MODELOVÁ ANALÝZA PŘÍBUZNOSTI DRUHŮ RODU BRASSICA ZA VYUŽITÍ METODY BAG-OF-WORDS

V závěrečné části textu chceme představit modelovou analýzu fylogenetických vztahů pomocí lingvistické metody Bag-of-words popsané výše. Naším předpokladem je, že podobnost (a tedy blízkost) mezi texty přirozeného jazyka nebo mezi texty DNA (sekvencemi) lze srovnávat na podobném principu, tj. na základě složek (slov, bází) obsažených v jednotce (věta/text, sekvence).

Věnovat se budeme taxonomizaci čeledi Brassicaceae (brukvovité). Současná taxonomie zahrnuje 3700 rostlin, dále hovoří o 338 rodech a 25 tribech (tribus je novodobý taxon zařazený mezi čeleď a rod) této čeledi (Al-Shehbaz et al., 2006; Bailey et al., 2006). Čeleď Brassicaceae, jmenovitě rostliny rodu *Brassica*, zahrnují řadu kulturně využívaných plodin, které jsou také v souvislosti s jejich dlouhodobým šlechtěním morfologicky velmi odlišné. Jsou pěstovány pro list (*B. oleracea*), otevřený a tvořící hlávky, kořen (*B. rapa*), i semena (*B. napus*). V souvislosti s odlišnými způsoby taxonomizace (morfologický/fenotypický × molekulární; viz výše) se tak čeleď Brassicaceae jeví být jako vhodný reprezentativní materiál k demonstrování metod molekulární fylogenetiky, tzn. metod zohledňujících podobnost genetických markerů, a nikoliv morfologických znaků.

Pro analýzu jsme vybrali významné kulturní plodiny rodu *Brassica* — *B. oleracea* (brukev zelná), *B. oleracea var. capitata* (hlávkové zelí), *B. oleracea var. alboglabra* (čínská brokolice), *B. oleracea var. botrytis* (květák), *B. rapa* (brukev řepák vodnice), *B. rapa var. chinensis* (čínské zelí), *B. rapa var. pekinensis* (pekingské zelí), *B. rapa var. oleifera* (brukev řepák olejný), *B. juncea* (brukev sítinovitá), *B. nigra* (brukev černá), *B. carinata* (hořčice habešská), *B. napus* (brukev řepka). Pro vybrané druhy a variety jsme našli vzorky (viz Tabulku 1) v genetické bance NCBI (National Center for Biotechnology Information) — jednalo se o ITS marker (konkrétně nekódující úsek

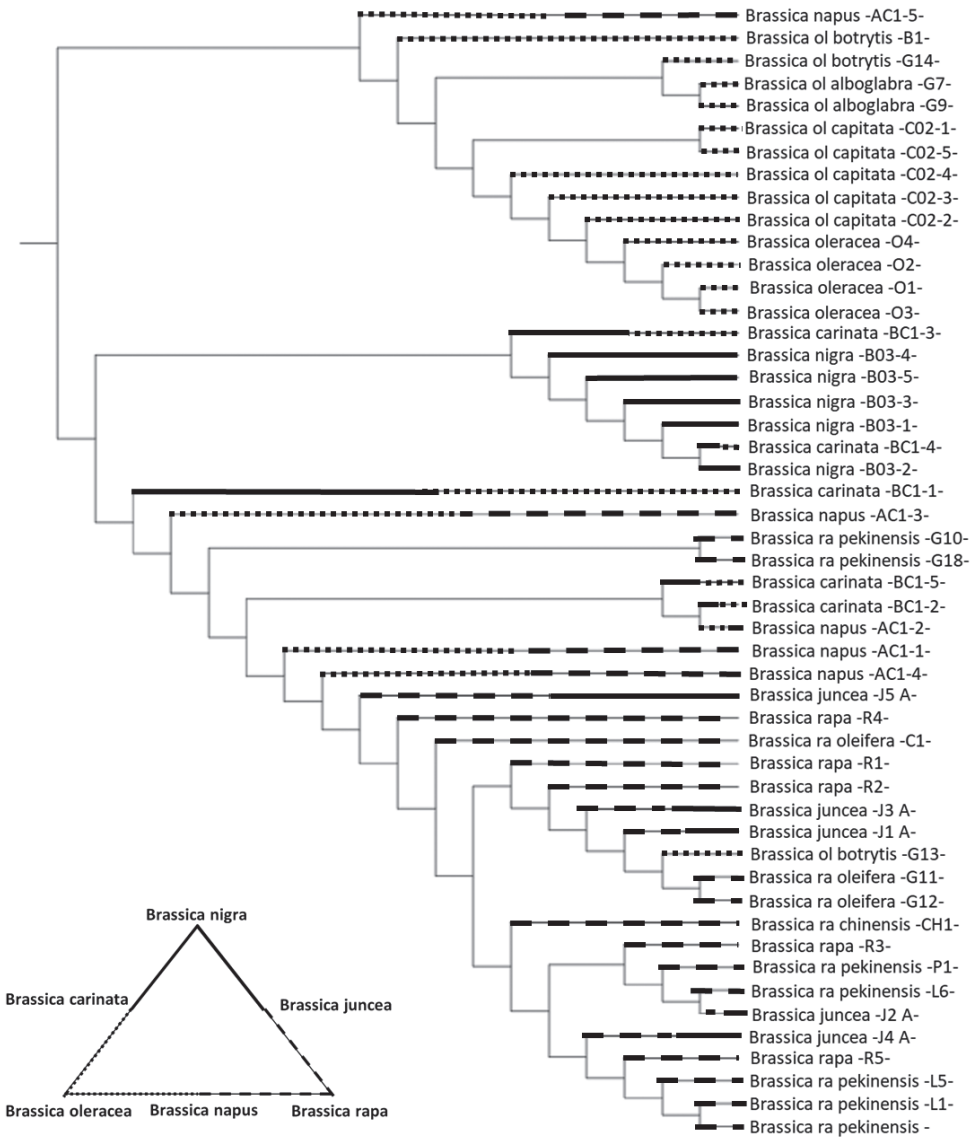
ribozomální RNA ITS1 5.8S ITS2), který je pro fylogenetické studie využíván pro svou variabilitu i mezi blízce příbuznými druhy. Brukve patří mezi tzv. allopolyloidní rostliny (dochází u nich ke znásobení chromozomových sad, které pochází od více druhů), proto je každý vybraný zástupce reprezentován několika vzorky.

Druh/Vzorek	**1**	**2**	**3**	**4**	**5**
Brassica carinata	DQ003688.1	DQ003689.1	DQ003690.1	DQ003691.1	DQ003692.1
Brassica juncea	MG923961.1	MG923962.1	MG923963.1	MG923964.1	MG923965.1
Brassica napus	DQ003657.1	DQ003658.1	DQ003659.1	DQ003660.1	DQ003661.1
Brassica nigra	DQ003644.1	DQ003645.1	DQ003646.1	DQ003647.1	DQ003648.1
Brassica ol. alboglabra	GQ891870.1	GQ891871.1			
Brassica ol. botrytis	AF128099.1	GQ891875.1	GQ891876.1		
Brassica ol. capitata	DQ003650.1	DQ003651.1	DQ003652.1	DQ003653.1	DQ003654.1
Brassica oleracea	MG923981.1	MG923982.1	MG923983.1	MG923984.1	
Brassica ra. chinensis	AF128095.1	GQ202246.1	GQ202249.1	GQ202250.1	
Brassica ra. oleifera	GQ268061.1	GQ891873.1	GQ891874.1		
Brassica ra. pekinensis	GQ891872.1	GQ891880.1	GQ202251.1	AF128096.1	
Brassica rapa	MG923985.1	MG923986.1	MG923987.1	MG923988.1	MG923989.1

TABULKA 1: Seznam použitých vzorků pro jednotlivé druhy a variety rodu *Brassica*.

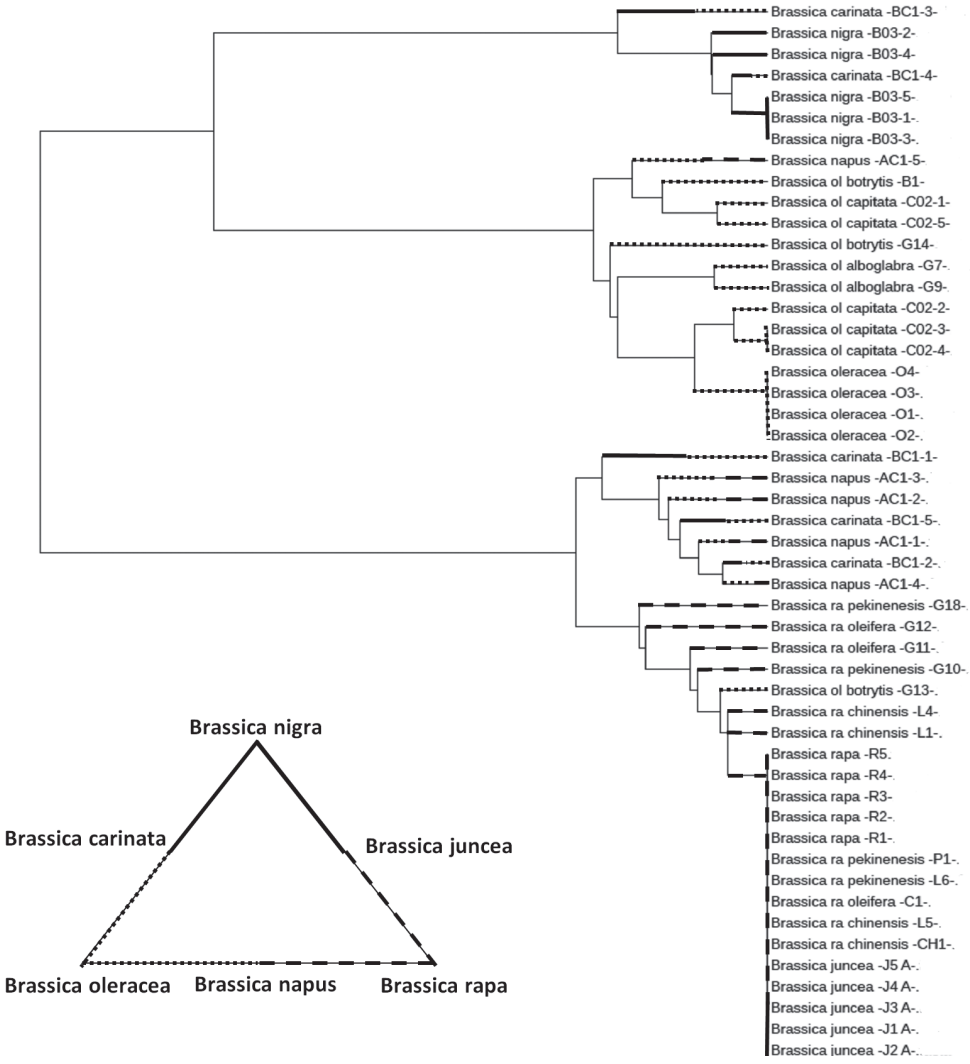
Pro srovnání a zhodnocení vhodnosti metody Bag-of-words pro fylogenetické studie jsme plodiny rodu *Brassica* nejprve podrobili standardizované bioinformatické metodě Multiple Sequence Alignment, která porovnává genetické sekvence na základě delecí, insercí, substitucí a transformací nukleotidů/aminokyselin. Použili jsme program webPRANK (https://www.ebi.ac.uk/goldman-srv/webprank/), který ke zhodnocení genetické podobnosti hledá v sekvencích fylogenetické vzory. Výsledky (viz Graf 1) jsou zobrazeny pomocí kladogramu a dále graficky rozlišeny vzhledem ke genetickým vztahům hlavních druhových skupin rodu *Brassica* — *Brassica olerace*, *Brassica rapa*, *Brassica nigra* a křížením vzniklé *Brassica napus* (*B. oleracea* + *B. rapa*), *Brassica carinata* (*B. oleracea* + *B. nigra*) a *Brassica juncea* (*B. rapa* + *B. nigra*).

Samotnou analýzu Bag-of-words jsme provedli následovně: Sekvence jsme rozdělili na 3-gramy, tedy subsekvence o délce tří nukleotidů (v bioinformatice užíván pojem k-mer). Subsekvence tripletů jsme dále použili jako „slova" v Bag-of-words analýze. Na základě zastoupení těchto „slov" a jejich frekvence v jednotlivých sekvencích jsme pak měřili podobnost sekvencí. Jejich srovnání spočívá ve sloučení seznamů „slov" srovnávaných „genetických textů" a reprezentací jednotlivých „textů" číselnou řadou odpovídající zastoupení jednotlivých „slov" v tzv. globálním slovníku (0 — v textu se slovo nevyskytuje, 1 — v textu se slovo vyskytuje jednou atd). Výsledné číselné řady představují vlastnosti textů vstupující do hierarchického shlukování. Rozdíly mezi jednotlivými vlastnostmi srovnávaných textů/objektů jsou reprezentovány jako vzdálenosti objektů, které jsou v grafu vyznačeny pomocí počtu větvení a délky větví, na nichž se objekty nachází. K srovnání sekvencí jsme využili shlukovací metodu ward.D2 a pro clustering eukleidovskou vzdálenost.



GRAF 1: Multiple Sequence Alingment vybraných zástupců rodu Brassica pomocí programu webPRANK.

Ve výše uvedeném grafu (viz Graf 2) můžeme pozorovat, že Bag-of-words model 3-gramů ITS markeru velmi dobře vyjadřuje příbuzenské vztahy druhů, jejich variet a hybridů rodu *Brassica*. Stejně jako při Multiple Sequence Alignment pozorujeme oddělení zástupců *B. oleracea* od zástupců *B. rapa* na samostatných větvích. Z dějin agrikultury zjišťujeme (viz Sadowski & Kole, 2011), že hlávkové zelí, květák a čínská brokolice mají společného předka, divokou brukev zelnou, kterou nacházíme na



GRAF 2: Bag-of-words model zástupců rodu *Brassica* za využití 3-gramů nukleotidů; clustering pomocí shlukovací metody ward.D2 a euklidovské vzdálenosti.

větvi spolu s varietami této druhové skupiny. Stejným způsobem můžeme identifikovat společný původ čínského zelí, pekingského zelí a brukve řepáku olejného, původ těchto variet se váže k vodnici, kterou v této skupině také nalézáme. V druhové skupině s *B. rapa* nacházíme také druh *B. juncea*, která je křížencem *B. rapa* a *B. nigra*. Vzorky hybridu *B. napus* se nachází na větvi mezi svými progenitory *B. oleracea* a *B. rapa* (pouze jeden se nachází na větvi s *B. oleracea*). Vzorky *B. carinata* nacházíme rozdělené mezi větve jejích progenitorů, část z nich je přímo na větvi s *B. nigra*, zbylé jsou spolu s *B. napus* blíže *B. oleracea*.

Když porovnáme výsledky bioinformatické metody Multiple Sequence Alignment a lingvistické metody Bag-of-words, zjistíme že vyhodnocují genetickou podobnost zkoumaných druhů velmi podobným způsobem — shlukování zástupců do větví i délka jednotlivých větví jsou takřka totožné. V každém případě můžeme při porovnání metod hovořit o shodném vyhodnocení genetické podobnosti poukazující na genetickou příbuznost zástupců rodu *Brassica* (podrobněji o využití metody k mapování mezidruhových vztahů rostlin čeledi Brassicaceae Owsianková et al., 2018).

ZÁVĚR

V článku jsme shrnuli historický vývoj vztahu biologie a lingvistiky, zaměřili jsme se na kvantitativní lingvistické metody využívané v analýze genetického textu a blíže představili metody využívané v molekulární fylogenetice. Na závěr jsme na modelové analýze kulturních plodin čeledi Brassiceacae čtenáři představili fylogenetickou analýzu pomocí lingvistické metody Bag-of-words a výsledky porovnali se standardizovanou bioinformatickou metodou Multiple Sequence Alignment za využití programu webPRANK — upozornili jsme tak na využitelnost Bag-of-words při rekonstrukci příbuznosti organismů.

LITERATURA:

Al-Shehbaz, A. I., Beilstein, M. A., Kellog, E. A., & Traffanstedt, T. (2006). Systematics and phylogeny of the Brassicaceae (Cruciferae): An overview. *Plant Syst. Evol.*, *259*, 89–120.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, *2*, 1–10.

Augustyn, P. (2013). What connects biolinguistics and biosemiotics? *Biolinguistics*, *7*, 96–111.

Bailey, C. D., Koch, M. A., Mayer, M., Mummenhoff, K., O'Kane, S. L., Warwick, S. I., Windham, M. D., & Al-Shehbaz, I. A. (2006). Toward a global phylogeny of the Brassicaceae. *Mol Biol Evol.*, *23*, 2142–2160.

Baixeries, J., Hernández-Fernández, A., Forns, N., & Ferrer-i-Cancho, R. (2013). The parameters of Menzerath-Altmann law in genomes. *Journal of Quantitative Linguistics*, *20*(2), 94–104.

Barbieri, M. (2006). *Organické kódy*. Praha: Academia.

Barbieri, M. (2008). Life is semiosis: The biosemiotic view of nature. *Cosmos and History: The Journal of Natural and Social Philosophy*, *4*, 29–52.

Barbieri, M. (2015). *Code Biology: A New Science of Life*. Dordrecht: Springer.

Benešová, M. (Ed.). (2014). *Menzerath-Altmann Law Applied*. Olomouc: Univerzita Palackého v Olomouci.

Benešová, M., Faltýnek, D., & Zámečník, H. L. (2015). Menzerath-Altmann law in differently segmented texts. In M. Benešová, J. Mačutek & A. Tuzzi (Eds.), *Recent Contributions to Quantitative Linguistics*. New Yourk: De Gruyter Mouton.

Berg, S. L. (1926). *Nomogenesis or Evolution Determined by Law*. London: Constable.

Berwick, R. G., & Chomsky, N. (2011). The biolinguistic program: The current state of its development. In A. M. di Sciullo & C. Boeckx (Eds.), *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of Human Language Faculty* (s. 19–41). Oxford: Oxford University Press.

Berwick, R. C., & Chomsky, N. (2016). *Why Only Us? Language and Evolution*. Cambridge: The MIT Press.

Bolshoy, A. (2003). DNA sequence analysis linguistic tools: Contrast vocabularies, compositional spectra and linguistic complexity. *Applied bioinformatics*, *2*, 103–112.

Bolshoy, A., Volkovich, Z., Kirzhner, V., & Barzily, Z. (2010). *Genome Clustering from Linguistic Models to Classification of Genetic Texts*. Berlin: Springer.

Brinkman, F. S. L., & Leipe, D. D. (2001). Phylogenetic analysis. In A. D. Baxevanis & B. F. F. Ouelette (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (s. 323–358). NY: John Wiley & Sons.

Cvrčková, F. (2006). *Úvod do praktické bioinformatiky*. Praha: Academia.

Čadková, L. (2015). Do they speak language? *Biosemiotics*, *8*(1), 9–27.

Čepl, J., Pátková, I., Blahůstková, A., Cvrčková, F., & Markoš, A. (2010). Patterning of mutually interacting bacterial bodies: Close contacts and airborne signals. *BMC Microbiol.*, *12*, 110–139.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

Dupuis, C. (1984). Willi Hennig's impact on taxonomic thought. *Annual Review of Ecology and Systematics*, *15*, 1–24.

Embleton, S. (2000). Lexicostatistics/ Glottochronology: From Swadesh to Sankoff to Starostin to future horizons. In C. Renfrew, A. McMahon & L. Trask (Eds.), *Time Depth in Historical Linguistics* (s. 143–165). Cambridge: McDonald Institute for Archaeological Research.

Faltýnek, D. (2017). *Co je nového v lingvistice*. Praha: Nová beseda.

Faltýnek, D., Matlach, V., & Lacková, L. (2019). Bases are not letters: On the analogy between the genetic code and natural language by sequence analysis. *Biosemiotics*, *12*(2), 289–304.

Fangerau, H., Geisler, H., Halling, T., & Martin, W. (Eds.). (2013). *Classification and Evolution in Biology, Linguistics and the History of Science: Concepts — Methods — Visualization*. Stuttgart: Steiner.

Faverau, D. (2009). *Essential Readings in Biosemiotics: Anthology and Commentary*. Dordrecht: Springer.

Ferrer-i-Cancho, R., & Elvevag, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE*, *5*, e9411.

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Belenguix, G., & Baixeries, J. (2013). The challenges of statistical patterns of language: The case of Menzerath's law in genomes. *Complexity*, *18*(3), 11–17.

Flegr, J. (2005). *Evoluční biologie*. Praha: Academia.

Forster, P., & Renfrew, C. (2006). *Phylogenetic Methods and the Prehistory of Languages*. University of Cambridge: McDonald Institute Press.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*, 435–439.

Grollemund, R., & Hombert, J.-M. (2012). Use of plant names for the classification of the Bantu languages of Gabon. In B. Connell & N. Rolle (Eds.), *Selected Proceedings of the 41st Annual Conference on African Linguistics* (s. 150–163). Somerville: Cascadilla Proceedings Project.

Havlin, S., Buldyrev, S. V., Goldberger, A. L., Mantegna, R. N., Peng, C. K., Simons, M., & Stanley, H. E. (1995). Statistical and linguistic features of DNA sequences. *Fractals*, *3*(2), 269–284.

Hennig, W. (1966). *Phylogenetic Systematics*. Urbana: University of Illinois Press.

Hennig, W. (1975). Cladistic analysis or cladistic classification? A reply to Ernst Mayr. *Systematic Zoology*, *24*(2), 244–256.

Hernández-Fernández, A., Baixeries, J., Forns, N., & Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy*, *13*(8), 1465–1480.

Hockett, C. (1982). The origin of speech. In W. S.-Y. Wang (Ed.): *Human Communication: Language and Its Psychobiological Bases: Readings from Scientific American* (s. 4–12). San Francisco: W. H. Freeman.

Hřebíček, L. (2002). *Vyprávění o lingvistických experimentech s textem*. Praha: Academia.

Hřebíček, L. (2007). Sémantické slapy v textových strukturách. *Slovo a slovesnost*, *68*(2), 83–90.

Jacob, F., Jakobson, R., Lévi-Strauss, C. & L'Héritier, P. (1968). Vivre et parler. *Les lettres françaises*, 1221–2.

Jakobson, R. (1932). *Zur Struktur des russischen Verbums*. Praha: PLK.

Jakobson, R. (1971). Linguistics in relation to other sciences. In R. Jakobson (Ed.), *Selected Writings II. Word and Language* (s. 655–696). The Hague: Mounton.

Ji, S. (1997). Isomorphism between cell and human languages: Molecular biological, bioinformatic and linguistic implications. *Biosystems*, *44*(1), 17–39.

Ji, S. (1999). The linguistics of DNA: Words, sentences, grammar, phonetics, and semantics. *Annals of the New York Academy of Sciences*, *870*, 411–417.

Katz, G. (2008). The hypothesis of a genetic protolanguage: An epistemological investigation. *Biosemiotics*, *1*, 57–73.

Kořenský, J. (2008). Teleologie jako jeden ze základních pojmů Pražského lingvistického kroužku? *Slovo a slovesnost*, *69*, 44–48.

Kosta, P., & Krivochen, D. (2012). Some thoughts on language Diversity, UG and the importance of language typology: Scrambling and non-monotonic merge of adjuncts and specifiers in Czech and German. *Zeitschrift für Slawistik*, *57*, 377–407.

Králík, J. (1976). Sovětský přínos k matematickým modelům proměny slovníku v čase. *Slovo a slovesnost*, *37*(1), 51–56.

Lacková, L., Faltýnek, D., & Matlach, V. (2017). Arbitrariness is not enough: Towards a functional approach to the genetic code. *Theory Biosci*, *136*(3–4), 187–191.

Lamprecht, A., Šlosar, D. & Bauer, J. (1986). *Historická mluvnice češtiny*. Praha: SPN.

Leška, O. (1986). Poznámky k teleologickému pojetí jazyka. In J. Nekvapil & O. Šoltys (Ed.), *Linguistica XVI* (s. 63–93). Praha: Interní tisk ÚJČ ČSAV.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics*, *10*(8), 707–710.

Li, W. (2012). Menzerath's law at the gene-exon level in the human genome. *Complexity*, *17*(4), 49–53.

Linnæi, C. (1758). *Systema naturæ per regna tria naturae: Secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Holmiæ: Laurentius Salvius.

Lovato, P. (2015). *Bag of words approaches for Bioinformatic*. Disertační práce, Verona: University of Verona.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M., & Stanley, H. E. (1995). Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E*, *52*, 2939–2950.

Markoš, A. (2003). *Tajemství hladiny*. Praha: Dokořán.

Markoš, A., & Faltýnek, D. (2011). Language metaphors of life. *Biosemiotics*, *2*(4), 171–200.

Markoš, A., & Švorcová, J. (2009). Recorded versus organic memory: Interaction of two worlds as demonstrated by the chromatin dynamics. *Biosemiotics*, *2*, 131–149.

Mayr, E. (1974). Cladistic analysis or cladistic classification? *Zeitschrift für Zoologische Systematik und Evolutionsforschung*, *12*, 94–128.

Matlach, V., & Faltýnek, D. (2016). Báze nejsou písmena. *Studie z aplikované lingvistiky*, *7*(1), 20–38.

Němec, I. (1989). Principy jazykového vývoje a historie češtiny. *Slovo a Slovesnost*, *50*, 81–96.

Nikolaou, C. (2014). Menzerath-Altmann law in mammalian exons reflects the dynamics of gene structure evolution. *Comput Biol Chem*, *53*, 134–143.

Niyogi, P., & Berwick, R. C. (1995). A note on Zipf's law, natural languages, and noncoding DNA regions. *A. I. Memo*, No. 1530, C.B.C.L. Paper No. 118.

Ohno, S. (1992). Of palindromes and peptides. *Genomics*, *90*, 342–345.

Owsianková, H., Faltýnek, D., & Kučera, O. (2018). Genetic analysis of cabbages and

related cultivated plants using the bag-of-words model. *Linguistic Frontiers*, *1*(2), 122–132.

Page, R. D. M., & Holmes, E. C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science.

Pattee, H., & Kull, K. (2009). A biosemiotic conversation: Between physics and semiotics. *Sign Systems Studies*, *37*(1/2), 311–331.

Peng, F., Schuurmans, D., Keselj, V., & Wang S. (2003). Language independent authorship attribution using character level language models. *10th Conference of the European Chapter of the Association for Computational Linguistics, EACL*.

Petr, J. (1982). Darwinovo pojetí jazyka a myšlení: Příspěvek k dějinám filozofie jazyka. *Slovo a slovesnost*, *43*, 177–199.

Petr, J. (1988). Ke Gebauerovu pojetí jazyka. *Slovo a slovesnost*, *49*(1), 3–29.

Piantadosi, S. (2014). Zipf's law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130.

Popov, O., Segal, D. M., & Trifonov, E. N. (1996). Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems*, *38*(1), 65–74.

Rabinovich, E., Ordan, N., & Winter, S. (2017). Found in translation: Reconstructing phylogenetic language trees from translations. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 530–540.

Rádl, E. (2006). *Dějiny biologických teorií novověku I.–II.* Praha: Academia.

Raible, W. (2001). Linguistics and genetics: Systematic parallels. In M. Haspelmath, E. König, W. Oesterreicher & W. Raible (Eds.): *Language Typology and Language Universals: An International Handbook* (s. 103–123). New York: Walter de Gruyter.

Rexová, K., Frynta, D., & Zrzavý, J. (2003). Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics*, *19*, 120–127.

Rexová, K., Bastin, Y., & Frynta, D. (2006). Cladistic analysis of Bantu languages: A new tree based on combined lexical and grammatical data. *Naturwissenschaften*, *93*, 189–194.

Romportl, S. (1989). Jazyk a etologický kód populace. *Slovo a slovesnost*, *50*(4), 257–269.

Romportl, S. (1994). *O biologické povaze přirozeného jazyka*. Boskovice: Albert.

Romportl, S. (1997). Genealogický strom (Příspěvek k metodologii evoluční jazykovědy). *Sborník prací Filozofické fakulty brněnské univerzity Studia minora Facultatis Philosophicae Universitatis Brunensis*, *45*, 5–17.

Romportl, S. (1999). Descendenční analýza. *Sborník prací Filozofické fakulty brněnské univerzity Studia minora Facultatis Philosophicae Universitatis Brunensis*, *47*, 27–32.

Ross, R. M., Greenhill, S. J., & Atkinson, Q. D. (2013). Population structure and cultural geography of a folktale in Europe. *Proceedings of the Royal Society B: Biological Sciences*, 280 (1756).

Sadowski, J., & Kole, Ch. (2011). *Genetics, Genomics and Breeding of Vegetable Brassicas*. Boca Raton: CRC Press.

Savický, N. (1991). O některých méně známých pramenech Tezí Pražského lingvistického kroužku. *Slovo a slovesnost*, *52*, 196–198.

Searls, D. B. (2002). The language of genes. *Nature*, *420*, 211–217.

Serva, M., & Petroni, I. F. (2007). Indo-European languages tree by Levenshtein distance. *Europhysics Letters*, *81*, 680–685.

Shahzad, K., Mittenthal, J. E., & Caetano-Anollés, G. (2015). The organization of domains in proteins obeys Menzerath-Altmann's law of language. *BMC Systems Biology*, *9*(44), 1–13.

Sharov, A. (2010). Functional information: Towards synthesis of biosemiotics and cybernetics. *Entropy*, *12*(5), 1050–1070.

Schleicher, A. (1869). *Darwinism Tested by the Science of Language*. London: J. C. Hotten.

Skalička, V. (1967). O kontinuitě slov. *Slovo a slovesnost*, *28*(4), 355–359.

Stace, C. A. (1989). *Plant taxonomy and biosystematics*. Cambridge: Cambridge University Press.

Sutton, L. A., Papadopoulos, G., Hadzidimitriou, A., Papadopoulos, S.,

Kostareli, E., Rosenquist, R., Tzovaras, D., & Stamatopoulos, K. (2014). An entity evolving into a community: Defining the common ancestor and evolutionary trajectory of chronic lymphocytic leukemia stereotyped subset #4. *Molecular Medicine*, *20*(1), 720–728.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of American Philosophical Society*, *96*, 452–463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, *21*, 121–137.

Syllaba, T. (1983). *Jan Gebauer na pražské univerzitě*. Praha: Univerzita Karlova.

Šoltys, O. (1991). Kulturní kontext Pražského lingvistického kroužku. *Slovo a slovesnost*, *52*, 198–201.

Toldo, R., Castelani, U., & Fusiello, A. (2009). A *bag of words* approach for 3D object categorization. In A. Gagalowicz & W. Philips (Eds.), *Computer vision/computer graphics Collaboration techniques: 4th International Conference, MIRAGE 2009* (s. 116–127). Berlin: Springer.

Trifonov, E. N., & Berezovsky, I. N. (2002). Proteomic code. *Molecular Biology*, *36*, 239–243.

Trifonov, E. N., Kirzhner, A., Kirzhner, V. M., & Berezovsky, I. N. (2001). Distinct stages of protein evolution as suggested by protein sequence analysis. *J. Mol. Evol*, *53*, 394–401.

Tsonis, A. A., Elsner, J. B., & Panagiotis, A. T. (1997). Is DNA a language? *J. Theor. Biol.*, *184*, 25–29.

Vachek, J. (1968). *Dynamika fonologického systému současné spisovné češtiny*. Praha: Academia.

Vachek, J. (1972). *Z klasického období pražské školy 1925–1945*. Praha: Academia.

Wang, J., Liu, P., She F. H., M., Nahavandi, S., & Kouzani, A. (2013). Bag-of-words representation for biomedical time series classification. *Biomedical Signal Processing and Control, 8*(6), 634–644.

Wheeler, Q. D. (2004). Taxonomic triage and the poverty of phylogeny. In H. C. J. Godfray & S. Knapp (Eds.): *Taxonomy for the Twenty-first Century*. *Philosophical Transactions of the Royal Society*, *359*, 571–583.

Zemková, M., Trifonov, E. N., & Zahradník, D. (2014). One common structural feature of "words" in protein sequences and human texts. *J Biomol Struct Dyn*, *32*(7), 1085–1091.

Zemková, M. (2016). *Lingvistické přístupy v genomice a lingvistická metafora v biologii.* Disertační práce, Praha: Přírodovědecká fakulta, Univerzita Karlova v Praze.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge: Addison-Wesley Press.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52.

Hana Owsianková
Katedra obecné lingvistiky, Filozofická fakulta Univerzity Palackého v Olomouci
<hana.owsiankova@upol.cz>

Ondřej Kučera
Katedra asijských studií, Filozofická fakulta Univerzity Palackého v Olomouci
<ondrej.kucera@upol.cz>

Dan Faltýnek
Katedra obecné lingvistiky, Filozofická fakulta Univerzity Palackého v Olomouci
<dan.faltynek@upol.cz>