ČASOPIS PRO MODERNÍ FILOLOGII 103, 2021, Č. 2, S. 228–243

OPEN ACCESS

Víceslovné jednotky typické pro české akademické texty[1]

Dominika Kováříková (Praha) – Oleg Kovářík (Praha) – Lucie Lukešová (Praha)

MULTI-WORD UNITS IN CZECH ACADEMIC TEXTS

This paper introduces Akalex, a new online tool created to support vocabulary research into Czech academic texts. The Akalex database includes close on 60 000 n-grams — candidates for typical academic words or multiword units — and it can be readily searched and filtered according to several criteria. The n-grams were extracted from the SYN2015 corpus of written contemporary Czech, based on their prominent frequency in academic texts and shared occurrence in many different academic disciplines, distinguishing them from general vocabulary on one hand and specialized terminology on the other. Each n-gram in the database is also furnished with additional information, such as part-of-speech, distribution by disciplines, frequency etc., making it possible to search for e.g. collocations with a specific lexeme (such as adjectives combined with the word *výzkum* 'research' or verbs with a certain preposition).

The features of Akalex were put to the test in our case study covering 2-grams to 6-grams used in all 24 academic disciplines included in the SYN2015 corpus. Of nearly 900 candidates, 236 were manually chosen by two annotators as typical for academic texts. These were then further analysed and split into groups based on their semantic, functional and formal features. Among the most frequent were lexical bundles, collocations with content words and combinations of two verbs pointing to a frequent use of passives in academic texts etc.

KEYWORDS

academic texts, academic vocabulary, multi-word units, corpus research

KLÍČOVÁ SLOVA

akademické texty, akademická slovní zásoba, víceslovné jednotky, korpusový výzkum

DOI

https://doi.org/10.14712/23366591.2021.2.4

1. ÚVOD

Akademické texty slouží ke komunikaci mezi odborníky v rámci akademických disciplín, a to jak v odborných monografiích nebo ve vědeckých článcích, tak ve vysokoškolských učebnicích, ve výuce i ve studentských kvalifikačních pracích. Nagy

1 Tento článek vznikl v rámci programu Progres Q08 *Český národní korpus* uskutečňovaného na Filozofické fakultě Univerzity Karlovy. Zároveň byl podpořen také v rámci projektu Kreativita a adaptabilita jako předpoklad úspěchu Evropy v propojeném světě reg.č.: CZ.02.1.01/0.0/0.0/16_019/0000734 financovaného z Evropského fondu pro regionální rozvoj.

a Townsend (2012) přiléhavě popisují jazyk akademických textů jako psaný (nebo i mluvený) jazyk, který umožňuje komunikaci a uvažování o tématech konkrétního vědeckého oboru. Studenti vědeckých oborů se v rámci přípravy na akademickou dráhu musí učit pracovat s poměrně pevnou strukturou celého textu (zvláště ve vědeckých článcích), s neobvyklou podobou vět, v kterých jsou např. bohatě rozvíjeny nominální složky textu, ale také se specifickou slovní zásobou oborovou (termíny) i slovní zásobou společnou většině oborů — tedy s jednoslovnými a víceslovnými jednotkami typickými pro akademické texty.

V posledních dvaceti letech projevují badatelé velký zájem právě o slovní zásobu společnou velkému množství akademických textů z různých oborů. V roce 2000 byl publikován Academic Word List (Coxhead, 2000), seznam 570 akademických slov (substantiv, sloves a adjektiv), který vyvolal mnoho reakcí. Ty kritické seznamu vytýkaly mimo jiné fakt, že obsahuje pouze jednoslovné jednotky (Simpson-Vlach — Ellis, 2010; Gardner — Davies, 2014; Granger, 2017). Další výzkum se tedy začal zaměřovat i na tzv. akademické fráze nebo víceslovné jednotky, a vznikly tak seznamy akademických víceslovných jednotek nebo kombinované seznamy jednoslovných i víceslovných jednotek, např. *A New Academic Vocabulary List* (Gardner — Davies, 2014), *Academic Keyword List* (Paquot, 2010) a *Academic Phrasebank* (http://www.phrasebank.manchester.ac.uk/).

Seznamy akademických slov a frází se většinou soustřeďují na angličtinu jakožto současný univerzální jazyk vědy a obvykle jsou určeny pro studenty i výzkumníky z celého světa, pro něž angličtina není mateřským jazykem. Stejně tak ale mohou podobné seznamy sloužit i rodilým mluvčím, zvláště vysokoškolským studentům a začínajícím badatelům.[2] Právě to nás vedlo k sestavení takového seznamu i pro češtinu. Seznam akademických slov a frází může napomoci plynulému vyjadřování jak v kvalifikačních studentských pracích (Chen — Baker, 2010), tak i ve vysoce odborných textech, jako jsou vědecké články nebo monografie (Hyland, 2012).[3]

Jedním z hlavních problémů výzkumu akademického jazyka je variabilita vyhledávaných víceslovných jednotek: jejich délka může být v rozsahu dvou až šesti slov[4] (výjimečně snad i více), může jít o tzv. lexical bundles[5] (Biber — Barbieri, 2007), o termíny, které se s oslabenou terminologickou platností rozšířily do velkého množ-

2 Nutnost naučit se akademickému jazyku jakožto nesamozřejmé dovednosti vyjadřuje citát, který je často zmiňován v kontextu kurzů zaměřených na akademickou angličtinu (EAP): „Academic language is […] no one's mother tongue." (Bourdieu et al., 1994, s. 8).

3 V češtině je situace velmi odlišná od angličtiny, k níž se vyjadřuje Hyland (2012); anglické seznamy akademické slovní zásoby slouží především velkému množství nerodilých mluvčích. Český seznam by však mohl sloužit jako pomůcka i rodilým mluvčím při psaní odborných textů nejvyšší úrovně, pokud by obsahoval v rámci každého hesla i další informace, například seznamy typických kolokací nebo seznamy synonym.

4 Nelze vyloučit, že existují i delší víceslovné jednotky (typické pro akademické texty), šlo by však o zcela výjimečné případy.

5 Lexical bundles jsou často opakované (velmi běžné) neidiomatické sekvence slov, které nejsou strukturně ucelené (Biber — Barbieri, 2007); jde např. o spojení *to be able to, at the end of the*.

ství akademických disciplín, o časté kolokace typických akademických slov, o frekventované spojení akademického slova s předložkou nebo spojkou apod.

Pro výzkum českých akademických frází jsme vytvořili aplikaci **Akalex** (www.korpus.cz/akalex, Kováříková — Kovářík, 2021). Tuto aplikaci lze použít při výzkumu akademické slovní zásoby, ať už se jedná o jednotky jednoslovné, nebo víceslovné. Je k dispozici i ostatním badatelům, kteří se věnují problematice typické slovní zásoby akademických textů. Akalex má složku databázovou, která obsahuje snadno prohledávatelné seznamy n-gramů (slov jdoucích za sebou v různém počtu n) splňujících určité podmínky — zjednodušeně lze říct, že jde o n-gramy, které se vyskytují častěji v akademických textech než v referenčním korpusu. K těmto n-gramům jsou přiřazeny hodnoty rozličných kvantifikovatelných vlastností: jde o frekvenční a distribuční rysy a asociační míry. Druhou složkou aplikace Akalex je seznam typických akademických slov a frází vzniklý na základě dalšího výzkumu materiálu v databázové složce.

2. DATA A METODOLOGIE

Vytvoření databázové součásti aplikace Akalex probíhalo v několika krocích. Prvním krokem bylo vytvoření vhodných subkorpusů z aktuálního reprezentativního korpusu češtiny SYN2015: subkorpusu akademických textů a referenčního subkorpusu složeného z textů beletristických a publicistických. Druhým krokem bylo sestavení kompletního seznamu n-gramů požadovaného rozsahu (šlo o 1- až 6-gramy). V další fázi byly všem n-gramům přiřazeny rysy, pomocí nichž lze automaticky odlišit kandidáty na typické jednoslovné nebo víceslovné jednotky. Mezi tyto rysy patří především charakteristiky založené na frekvenci n-gramů a na jejich distribuci v korpusu (více v kap. 2.3). Výsledkem je databáze, která obsahuje jak samotné n-gramy, tak různé charakteristiky, pomocí nichž lze data prohledávat.

Funkčnost databáze Akalex byla ověřena případovou studií (viz kap. 4) zaměřenou na víceslovné jednotky, ve které jsme podrobně zkoumali všechny 2- až 6-gramy splňující dvě hlavní podmínky: 1. jejich frekvence je alespoň dvakrát vyšší v akademických textech než v referenčním korpusu a 2. vyskytly se ve všech 24 disciplínách rozlišovaných v korpusu SYN2015.

2.1 DATA Z KORPUSU SYN2015

Veškerý materiál pro výzkum je čerpán z aktuálního reprezentativního korpusu psaných textů SYN2015; v jeho rámci byly vytvořeny dva subkorpusy. Subkorpus nazvaný SCI je složen z akademických textů, které jsou v korpusu SYN2015 označeny zkratkou **SCI** (pro textový typ) a popiskem „odborná literatura“. Referenční subkorpus nazvaný **REF** obsahuje všechny beletristické a publicistické texty z korpusu SYN2015, to jest takové, které jsou v rámci skupin textových typů označeny zkratkami FIC („beletrie“) a NMG („publicistika“). Subkorpus SCI obsahuje 13 milionů tokenů a referenční subkorpus REF má rozsah 81 milionů tokenů (vč. interpunkce).

V subkorpusu SCI se rozlišuje 24 akademických disciplín, ve shodě s klasifikací oborů používanou Národní knihovnou ČR. Zařazení jednotlivých textů do oborů hraje zásadní roli ve vyhledávání jednoslovných i víceslovných jednotek typických pro akademické texty — informace o tom, v kolika akademických disciplínách se zkoumaný n-gram vyskytuje (tj. jaká je jeho distribuce v oborech), je rozhodující jak pro zařazení do databáze, tak pro další výzkum. V tabulce 1 jsou uvedeny počty tokenů v jednotlivých oborech v subkorpusu SCI, které lze najít pomocí strukturního atributu *genre* „žánr/oblast".

obor	počet tokenů	obor	počet tokenů
AGR: zemědělství, chovatelství	171 227	LAN: filologie	622 298
ANT: antropologie, etnografie	425 223	LAW: právo	563 209
ART: umění, architektura	489 628	MED: lékařství	708 980
BIO: biologie	1 133 593	MUS: hudba	676 962
CHE: chemie	278 837	PHI: filozofie, náboženství	689 801
ECO: ekonomie, obchod, logistika	212 669	PHY: fyzika	627 385
EDU: pedagogika	674 414	POL: politika, vojenství	729 145
GEO: geografie, geologie	685 251	PSY: psychologie	468 837
HIS: historie, biografie	691 094	REC: sport, rekreace, hobby	382 963
ICT: výpočetní technika	136 026	SOC: sociologie	790 198
INF: knihovnictví, informatika	490 324	TEC: technika	362 925
ITD: interdisciplinární	246 295	THE: divadlo, film, tanec	679 426

TABULKA 1. Počet tokenů v jednotlivých oborech v subkorpusu SCI (zdroj: SYN2015)

2.2 N-GRAMY PRO AKALEX

Na základě zjištění předchozích výzkumů není vhodné za lexikální jednotky typické pro akademické texty považovat pouze samostatná slova, nýbrž je třeba zahrnout i víceslovné jednotky různého typu (ať už víceslovné lexémy, typické kolokace, nebo např. spojení s typickou předložkou). Proto jsme pro databázi brali v úvahu kromě jednotlivých slov (unigramy) i n-gramy o dvou a více slovech.

Při práci s n-gramy jakéhokoli typu je vždy třeba učinit některá rozhodnutí, která s sebou nesou jak výhody, tak i nevýhody. Především je třeba zvolit, zda pro n-gramy používat lemmata coby základní slovníkové tvary, nebo konkrétní slovní tvary, a zda brát v úvahu interpunkci.

V rámci databázové části Akalexu jsme se rozhodli používat lemmata, a nikoli slovní tvary. Výhodou tohoto přístupu především u vysoce flektivní češtiny je, že n-gram získá spojením několika podob na frekvenci a vyjeví se tak lépe jeho typičnost. Nevýhodná je naopak nutnost používat předpřipravenou lingvistickou interpretaci

v podobě lemmatizace, která může být nespolehlivá nebo zavádějící. Dalším úskalím je poněkud složitější interpretace konkrétních n-gramů, jejichž typické tvary často nemůžeme odhadnout intuitivně — tuto nevýhodu jsme ale částečně eliminovali doplněním nejčastějšího tvaru příslušného n-gramu do databázového záznamu.

Vyhledané n-gramy nepřekračují hranici věty, ale samotná interpunkční znaménka označující konec věty (zvl. tečka) nejsou do n-gramů zahrnována. N-gramy neobsahují ani neukončující interpunkční znaménka. Týká se to zvláště čárek, jejichž odstraněním můžeme získat spojení se spojkami, např. *(vyplývat) že* nebo *(otázka) zda*.

Zvlášť v jazycích, jako je čeština, je také třeba mít na paměti, že volný slovosled může ztížit interpretaci výsledků založených na n-gramech. Tuto otázku jsme se však rozhodli při vytváření této databáze zanedbat.

Po přidělení hodnot vybraných rysů jednotlivým n-gramům (viz dále) jsme původní velice rozsáhlý seznam n-gramů dramaticky zkrátili — v databázi Akalex jsou obsaženy pouze takové n-gramy, které splňují dvě základní podmínky: vyšší frekvence oproti textům v referenčním korpusu a relativně vysoká distribuce v oborech (více v kap. 2.4).

2.3 POUŽITÉ RYSY

Na základě předchozích výzkumů slov a víceslovných jednotek v akademických textech (jak akademických frází, tak terminologie) jsme identifikovali několik rysů, které se osvědčily při automatické identifikaci těchto jednotek (Kováříková, 2017; Kováříková — Kovářík, 2019). Jde především o vlastnosti frekvenční a distribuční, v případě víceslovných jednotek (bigramů a trigramů) jsou k dispozici i asociační míry sloužící k identifikaci kolokací. Kromě toho jsme do databázové části zařadili i informaci o slovním druhu jednotlivých členů n-gramů a o nejfrekventovanějším tvaru daného n-gramu. Poslední dva údaje jsme využili především při lingvistické interpretaci v případové studii (kap. 4).

Přehled rysů využívaných v databázi Akalex:

Poměr frekvencí. Poměr frekvence v subkorpusu akademických textů SCI a v referenčním subkorpusu beletristických a publicistických textů ukazuje, kolikrát častější je daný n-gram v akademických textech oproti textům jiného typu. Pokud je tedy hodnota tohoto rysu např. 10, znamená to, že n-gram je 10x častější v akademických textech, je pro ně tedy typický (Kováříková, 2017, s. 38).

Distribuce je údaj o tom, v kolika z 24 oborů, které jsou k dispozici v subkorpusu SCI, se n-gram vyskytuje. Čím je distribuce vyšší, tím šířeji je n-gram používaný, a tím vyšší je pravděpodobnost, že se může jednat o akademickou frázi. Naopak nízká hodnota tohoto rysu ukazuje spíše na to, že se jedná o n-gram terminologický, tedy specifický pro určitou disciplínu nebo malé množství příbuzných disciplín — do seznamu obecných akademických frází bychom ho tedy primárně neřadili (tamtéž).

Disperze coby další distribuční charakteristika ukazuje, jak rovnoměrně je zkoumaný n-gram rozložený v jednotlivých oborech, které jsou k dispozici v subkorpusu SCI. Čím nižší je hodnota disperze, tím rovnoměrněji je n-gram rozložen v disciplínách, a tím vyšší je tedy pravděpodobnost, že jde o akademickou frázi. Disperze vhodně doplňuje údaj o distribuci. Může se totiž stát, že nějaký n-gram se vyskytne ve vysokém počtu disciplín, ovšem s velmi rozdílnou frekvencí — takový bigram bude sice mít vysokou distribuci, ale na druhou stranu i vysokou hodnotu disperze. Nás zajímají především n-gramy s vysokou distribucí a nízkou disperzí, tedy takové, které se vyskytují v mnoha oborech a objevují se v nich s podobnou frekvencí.

Asociační míra PMI (*pointwise mutual information*; k dispozici pouze pro bigramy a trigramy). Asociační míry jsou matematické postupy pro vyhledávání kolokací v korpusech — zjišťují, která slova se spolu vyskytují častěji, než odpovídá náhodnému souvýskytu (Church — Hanks, 1990).

Nejčastější tvar (v Akalexu pouze zkráceně „tvar") je slovní tvar daného n-gramu vyskytující se v akademických textech s nejvyšší frekvencí. Tento údaj je velmi užitečný především při interpretaci výsledků. V některých případech nám nejčastější tvar může napovědět, že daný n-gram je ve skutečnosti součástí delšího n-gramu (např. bigram *následující kapitola* má nejčastější tvar *následující kapitole*, je totiž součástí frekventovaného trigramu *v následující kapitole*), jindy je nejčastější tvar klíčem k porozumění n-gramu (např. bigramy *být vyjádřit, na příklad* nebo *v rámec* mají nejčastější tvary *je vyjádřen, na příkladu* a *v rámci*).

POS, slovní druh jednotlivých členů n-gramu. Určení slovního druhu všech součástí zkoumaných n-gramů může velmi dobře posloužit při vyhledávání v databázi. Můžeme například vyhledat všechna spojení slovesa s typickou předložkou nebo spojení substantiva s typickým přívlastkem (tj. předcházejícím adjektivem). Při analýze dat můžou být označené slovní druhy jedním z kritérií třídění (viz případovou studii v kap. 4).

2.4 KRITÉRIA PRO ZAHRNUTÍ N-GRAMŮ DO DATABÁZOVÉ SOUČÁSTI AKALEXU

Do databáze byly zahrnuty pouze takové n-gramy, které splňují dvě základní podmínky:

1. hodnota poměru frekvencí je alespoň 2, n-gram je tedy alespoň dvakrát častější v akademických textech než v referenčním korpusu beletristických a publicistických textů;
2. hodnota distribuce je alespoň 8, to znamená, že se n-gram vyskytuje minimálně ve třetině oborů rozeznávaných v subkorpusu SCI.

Tyto minimální hodnoty zajistí, že z obrovského počtu n-gramů, které lze extrahovat ze subkorpusu SCI, bude do databáze zahrnuto pouze zpracovatelné množství

n-gramů, u nichž je navíc vyšší pravděpodobnost, že nejde ani o obecně rozšířená slova nebo kombinace, ani o termíny (ty se vyskytují spíše v malém množství oborů).

3. DATABÁZOVÁ SOUČÁST AKALEXU: ZDROJ PRO ZKOUMÁNÍ AKADEMICKÝCH LEXIKÁLNÍCH JEDNOTEK

Databáze Akalex je určena k tomu, aby sloužila k výzkumu akademické slovní zásoby češtiny, a to jak jednotek jednoslovných, tak víceslovných, ale i k zachycení dalších jevů, jako je například typická valence nebo běžné kolokáty v akademické češtině. Obsahuje 1- až 6-gramy, které jsou alespoň dvakrát častější v akademických textech než v textech jiného typu a které se vyskytují alespoň v osmi akademických oborech. Tím je zajištěno, že databáze obsahuje zpracovatelné množství n-gramů, u nichž už existuje předpoklad, že by se mohlo jednat o akademickou lexikální jednotku nebo frázi.

V databázi je k dispozici celkem zhruba 57 tisíc n-gramů. Největší část, téměř dvě třetiny celého materiálu, tvoří bigramy. Počet různých typů n-gramů je uveden v tabulce 3.

3.1 INFORMACE OBSAŽENÉ V DATABÁZOVÉ ČÁSTI AKALEXU

Kromě samotných n-gramů jsou v databázi k dispozici informace o n-gramech, které mohou sloužit k vyhledávání, třídění nebo vytváření seznamů podle zadaných podmínek.

Mezi tyto informace patří nejčastější tvar celého n-gramu, slovní druh jednotlivých členů n-gramu, poměry frekvencí, distribuce v oborech, disperze v oborech a asociační míra PMI (pro bigramy a trigramy). Jednotlivé položky jsou podrobněji popsány v kap. 2.3. Jejich přehled je k dispozici v tabulce 2.

Rys nebo informace (pořadí dle databáze)	**Popis**
Tvar	nejčastější tvar n-gramu v akademických textech
POS1, 2, ...	slovní druh první, druhé (popř. další) složky n-gramu
Poměr frekvencí	kolikrát je n-gram častější v akademických textech než v jiných typech textů (beletrie, publicistika)
Distribuce	v kolika oborech se n-gram vyskytl alespoň jednou
Disperze	jak rovnoměrně je n-gram rozložený v oborech
PMI	jak silná je kolokace (jen pro bigramy a trigramy)

TABULKA 2. Informace o n-gramech dostupné v databázi Akalex

3.2 NASTAVENÍ PRAHOVÝCH HODNOT

Problémem při automatickém vyhledávání akademických lexikálních jednotek (a obecně při automatickém vyhledávání na základě nějakých vybraných rysů) je na-

stavení prahových hodnot.[6] Není jednoduché posoudit, zda by mezi akademické lexikální jednotky měly být zahrnuty pouze n-gramy s nejvyšší možnou distribucí (24), nebo n-gramy vyskytující se alespoň ve dvou třetinách oborů (16), případně v polovině oborů (12) atd. Podobně nelze jednoduše rozhodnout, zda lexikální jednotky typické pro akademické texty budou i mezi n-gramy dvakrát, nebo až třikrát či víckrát frekventovanější v akademických textech než v textech jiného typu.

V podstatě jde vždy o rozhodnutí, zda chceme najít velké množství akademických lexikálních jednotek, a spolu s nimi i velké množství šumu (zaměřujeme se na pokrytí — *recall*[7]), nebo jestli budeme v zadávání prahových hodnot přísnější, protože chceme výsledky bez šumu, tj. označené n-gramy budou skutečně akademické lexikální jednotky, ovšem s tím, že jich objevíme jen menší množství, mnohé zůstanou nerozpoznány (zaměření na přesnost — *precision*).

Pokud se např. rozhodneme, že nastavíme poměrně přísné podmínky, tedy že nás zajímají všechny n-gramy, které se vyskytují minimálně ve 20 oborech, s poměrem frekvencí alespoň 6 a s disperzí o hodnotě menší než 1, pak se vyhledá jen menší množství n-gramů, z nichž ale vysoké procento skutečně budou akademické lexikální jednotky. I v případě poměrně přísného nastavení je ale nutná ruční kontrola, jak je to u automatického vyhledávání obvyklé.

Pokud ale snížíme požadavky, např. vezmeme v úvahu i n-gramy s poměrem frekvencí alespoň 2, které se vyskytují v 16 oborech, výsledky sice bude náročnější zpracovat, protože budou obsahovat větší množství n-gramů, které je nutno zkontrolovat, ale na druhou stranu získáme některé platné akademické lexikální jednotky, které by nám jinak zůstaly skryty.

Rozhodnutí, jestli se soustředit na vyšší *precision*, nebo *recall*, záleží mimo jiné na výzkumném úkolu. Chceme-li vytvořit stručný seznam velmi rozšířených a spolehlivých akademických lexikálních jednotek, budeme nastavovat přísnější podmínky (*precision*). Pokud je naším záměrem vytvořit obsáhlý seznam jakožto základ slovníku akademických lexikálních jednotek, tedy budeme požadovat spíše stovky až tisíce jednotek (*recall*), můžeme snížit prahy jednotlivých charakteristik, ovšem čeká nás větší množství ruční práce, aby byl seznam spolehlivý.

Databáze Akalexu je vytvořena tak, aby si každý badatel mohl prahové hodnoty nastavit podle vlastních představ a konkrétního výzkumného úkolu.

3.3 VYHLEDÁVÁNÍ PODLE LEMMATU NEBO SLOVNÍHO DRUHU

V databázi lze vyhledávat i konkrétní lemmata nebo části lemmat. Tímto způsobem můžeme vyhledávat především typické kolokáty určitého slova, které jsou specifické pro akademické texty.

6 Jednou z možností, jak najít vhodné nastavení prahových hodnot, je zjišťování prostřednictvím strojového učení nebo data miningu, jaké kombinace hodnot různých rysů přinášejí nejlepší výsledky (více v Kováříková, 2017, a Kováříková — Kovářík, 2019).

7 *Precision* a *recall* jsou standardní evaluační statistické míry (viz např. Manning — Schütze, 2000).

Příkladem může být vyhledání slov pojících se se slovem *odvodit*. V záložce 2-gramy zadáme do rámečku pod nadpisem LEMMA1 sloveso *odvodit*. Výsledkem jsou čtyři bigramy s těmito nejčastějšími tvary: *odvodit z, odvozen od, odvodit že, odvodit i*. V tomto případě nás budou zajímat první tři bigramy, protože obsahují typickou předložku nebo typickou spojku. Čtvrtý bigram není z hlediska zkoumání akademických lexikálních jednotek zajímavý, což lze ověřit i vyhledáním tohoto bigramu v korpusu, např. v SYN2015. Pokud sloveso zadáme do rámečku pod nadpisem LEMMA2, budou výsledkem další čtyři bigramy: *je odvozen, lze odvodit, (je) možné odvodit, můžeme odvodit*. V tomto případě nás zajímají všechny čtyři bigramy. První napovídá, že sloveso *odvodit* se v akademických textech často vyskytuje v trpném rodě, další tři zase ukazují na modalitu, která je se slovesem v mnoha případech spojená.

Další možností je vyhledávat určité slovní druhy, například spojení slovesa se spojkou. Do rámečku pod nadpisem POS1 zadáme V, pod POS2 zadáme J (značky slovních druhů se shodují se značkami v korpusech řady SYN). Výsledků je v tomto případě pochopitelně daleko větší množství, zhruba 600 řádků. Prvních 8 položek v tomto seznamu má tyto nejčastější tvary: *definována jako, označuje jako, charakterizovat jako, chápat jako, jeví jako, je tudíž, vyplývá že, (se)*[8] *ukazuje že*. Sedm z těchto položek je pro nás zajímavých, jde o slovesa, která se pojí s typickou spojkou, v tomto případě *jako* nebo *že*. Položka *je tudíž* je pak pouhým spojením obecně vysoce frekventovaného slova (*být*) a spojky velmi frekventované v akademických textech (*tudíž*).

Obojí typ vyhledávání lze ještě zkombinovat například tak, že vyhledáme typická adjektiva (přívlastky) pro vybrané substantivum. Pro substantivum *analýza* tak získáme 6 předcházejících adjektiv, která se vyskytují alespoň v deseti disciplínách: *podrobná, detailní, kritická, další, hlubší, následná analýza*.

3.4 EXPORT DAT

Kompletní seznamy n-gramů obsažených v databázi společně s doplňujícími informacemi o nejčastějším tvaru, slovních druzích a frekvenčních a distribučních charakteristikách lze pomocí jediného tlačítka exportovat do formátu .csv, z kterého se snadno vytvoří např. excelová tabulka.

Stejně tak je možné exportovat pouze části seznamů, nějakým způsobem tříděné nebo omezené: například všechna adverbia následovaná slovesy, nebo všechny 4-gramy vyskytující se alespoň ve dvaceti oborech, nebo všechny adjektivní 1-gramy alespoň desetkrát častější v akademických textech než v jiných typech textů.

3.5 PROPOJENÍ S KORPUSEM SYN2015

Jednotlivé n-gramy v databázi Akalex jsou propojeny s korpusem SYN2015 — pokud tedy uživatel chce ověřit kontext, v jakém se daný n-gram skutečně používá v akademických textech, může tak učinit jediným klikem (tlačítko Zobrazit výskyty).

8 V případě, že se sloveso obvykle vyskytuje s reflexivním zájmenem *se* (ať už jde o reflexivní sloveso, nebo o pasivní tvar), přidáváme v této práci po ověření ve vyšších n-gramech v databázi Akalex zájmeno *se* do závorky při uvádění příkladů.

4 PŘÍPADOVÁ STUDIE

V případové studii, kterou jsme ověřovali funkčnost databáze Akalex, jsme se soustředili na omezené množství 2- až 6-gramů, konkrétně pouze na ty, které se vyskytují ve všech 24 oborech. Další parametry jsme neměnili ani jinak neomezovali.

S tímto nastavením jsme našli celkově **793** bigramů, **99** trigramů a **dva** tetragramy. Vyšší n-gramy se do tohoto užšího výběru nedostaly (vyskytují se v menším množství oborů). Všech 894 takto vyhledaných n-gramů jsme zkoumali podrobněji. Ručně jsme na základě shody mezi dvěma anotátorkami[9] vybrali n-gramy typické pro akademické texty. Ty jsme pak dále analyzovali, především jsme se pokoušeli najít skupiny formálně, sémanticky nebo jinak podobných n-gramů, abychom mohli posoudit, jaké typy jednotek a jaké jevy jsou typické pro akademické texty.

Typ n-gramu	Počet n-gramů v databázi	Počet n-gramů s distribucí D=24	Z toho počet víceslovných akademických jednotek
1-gram	9000	N/A	N/A
2-gram	36000	793	181 (23 %)
3-gram	10000	99	53 (54 %)
4-gram	2000	2	2 (100 %)
5-gram	150	0	0
6-gram	9	0	0
celkem	**57000**	**894**	**236 (26 %)**

TABULKA 3. Tabulka poskytuje informaci o počtu typů n-gramů v databázi Akalex, o počtu n-gramů vyskytujících se ve všech 24 disciplínách, a o počtu nalezených jednoslovných a víceslovných jednotek. Značka N/A znamená, že daná hodnota se v rámci výzkumu nezkoumala (u 1-gramů).

4.1 SEZNAM BIGRAMŮ, TRIGRAMŮ A TETRAGRAMŮ — KRITÉRIA VYŘAZOVÁNÍ

V rámci zkoumaných n-gramů jsme analyzovali necelých 900 jednotek. Z tohoto původního seznamu jsme z různých důvodů velké množství n-gramů vyloučili a vybrali jsme pouze 236 akademických frází různého typu.

Hlavním typem vyřazeným ze seznamu byla náhodná kombinace dvou nebo tří slov, která nevytvářela smysluplnou (ani funkční) frázi. Dalším velmi častým typem n-gramů, které nebyly dále analyzovány, byly kombinace dvou nebo tří vysoce frekventovaných slov, zvláště kombinace se slovy *být*, *který* nebo *v*, ale i kombinace s frekventovanými zájmeny, předložkami nebo spojkami typickými pro akademické texty (*týž, tento, vůči, avšak*).

Třetím nejčastějším důvodem pro vyřazení z dále analyzovaného seznamu byla účast nižšího n-gramu na n-gramu vyšším (tedy bigram je ve skutečnosti součástí uceleného trigramu apod., například bigram s nejčastějším tvarem *straně druhé* je ve

9 Spoluautorky článku, Dominika Kováříková a Lucie Lukešová.

skutečnosti součástí trigramu *na straně druhé*, trigram *vzhledem k tomu* se ve skutečnosti objevuje obvykle jako součást tetragramu *vzhledem k tomu že*).

Všech 236 vybraných n-gramů jsme dále analyzovali: sledovali jsme jak sémantické a funkční typy těchto frází, tak jejich formální rysy, zvláště příslušnost jednotlivých částí n-gramů ke slovním druhům.

4.2 SKUPINY AKADEMICKÝCH FRÁZÍ NA ZÁKLADĚ SÉMANTIKY A FUNKCE

Velkou část n-gramů, které vytvářejí nějakou smysluplnou jednotku, můžeme rozčlenit do skupin na základě podobností sémantických nebo funkčních. Příklady uvádíme v nejčastějším tvaru.

V našem materiálu byla velmi výrazně zastoupena kategorie vymezená funkčně. Jde o **lexical bundles** („slovní svazky"), tedy vysoce frekventované sekvence slov (často s velkým podílem funkčních slov), které jsou obvykle strukturně neúplné a slouží často jako diskurzní konektory (Biber — Barbieri, 2007). Do této skupiny patří především víceslovné předložky nebo spojky. Mezi *lexical bundles* patří např. n-gramy *v rámci, na principu, na úrovni, s výjimkou, a to i, v souvislosti s, do jisté míry, z tohoto hlediska, v souladu s, v závislosti na, vzhledem k tomu že*.

Druhá skupina, která je v n-gramech s nejvyšší distribucí zastoupena mnohem méně, je skupina **kolokací plnovýznamových slov** typických pro akademické texty, např. *další vývoj, různé typy, vyššímu stupni, celkového počtu*. Tuto skupinu považujeme za nejdůležitější pro výzkum typické slovní zásoby akademických textů. V n-gramech s nižší distribucí (které nesplnily podmínky této případové studie) se do ní totiž řadí oborově nespecifické víceslovné termíny. Jde o původní víceslovné termíny, které jsou sice původně zakotveny v některé z vědeckých disciplín, ale s poměrně vysokou frekvencí se používají (s oslabenou terminologickou platností) i v ostatních oborech. Jsou to často termíny spojené s výzkumem, s vědeckými metodami a s vědou obecně (ty původně pocházejí z oboru filozofie), případně termíny matematické a statistické: *vědecká teorie, empirický výzkum, teoretické východisko, statistické údaje, kvantitativní metoda, naměřené hodnoty, průměrná hodnota*.

Do této kategorie řadíme i typické **kolokace** jednoslovných původních termínů, jako příklad nám může posloužit termín *metoda;* v databázi Akalex (v 2-gramech s distribucí alespoň 9) pro ni můžeme vyhledat tyto kolokáty: *tradiční metody, použité metody, uvedené metody, specifické metody, využívat metody, metoda založená na, metody měření, metody práce, metoda spočívá v*.

Další tři typy, které lze v n-gramech frekventovaných v akademických textech rozpoznat, jsou v našem materiálu zastoupeny jen zcela okrajově, ač v n-gramech s nižší distribucí hrají důležitou roli. Tyto typy uvádíme pouze pro získání celkového přehledu o problematice, naprostá většina uvedených příkladů nesplňuje podmínku nejvyšší distribuce v oborech. Jde o fráze pro orientaci v textu, o fráze odkazující k odborné literatuře a o časoprostorové odkazy.

Pro akademické texty jsou typické víceslovné jednotky, které jsou zaměřeny na **organizaci textu a orientaci v něm**. Většinou se týkají částí textu, jako jsou kapitoly nebo oddíly. Případně se jejich pomocí odkazuje k něčemu, co už bylo zmíněno nebo co bude více rozvedeno v následujícím textu. Ve vybraném materiálu jsme nalezli

jedinou takovou frázi (*následující kapitola*), ale v n-gramech s nižší frekvencí je tento typ zastoupen bohatě: *jak již bylo zmíněno, výše uvedený, viz kapitola, viz dále, v tomto odstavci, v předchozím odstavci, následující oddíl, v závěru kapitoly*.

Dalším typem frází typických pro akademické texty se **odkazuje k odborné literatuře**. Zařazujeme sem frekventované n-gramy, které se vztahují buď ke konkrétním textům nebo autorům, nebo k seznamům literatury, nebo jde o poznámky o odborné literatuře obecně: *v odborné literatuře, v seznamu literatury, v zahraniční literatuře, v české odborné literatuře, odborné články, v monografii, autorský kolektiv, autoři uvádějí že, autor vychází z, odborné publikace*.

Posledním typem frází typických pro akademické texty, který je vymezitelný na základě sémantické podobnosti, jsou **časoprostorové odkazy**: *v druhé polovině 19. století, v první polovině 20. století, v osmdesátých letech, ve třicátých letech, po druhé světové válce, do první světové války, v současnosti, do současnosti, ve střední Evropě, v českých zemích, v českém prostředí, v Čechách a na Moravě, země západní Evropy, na území české republiky, v Severní a Jižní Americe, na přelomu*.

4.3 FORMÁLNĚ VYMEZITELNÉ SKUPINY

Podobně jako na základě sémantických a funkčních vlastností můžeme ve skupině analyzovaných n-gramů vyčlenit několik skupin víceslovných jednotek vyznačujících se morfologicko-syntaktickými nebo lexikálně-syntaktickými rysy, které jsou příznačné pro akademické texty. Tyto typy se obvykle nepřekrývají se skupinami z předešlé kapitoly.

Prvním specifickým typem jsou bigramy, v kterých se pojí sloveso *být* s příčestím trpným jiného slovesa. Trpný rod je v akademických textech více než třikrát častější než v referenčním korpusu beletrie a publicistiky. Některá slovesa na tom mají vyšší podíl než jiná, protože mají daleko vyšší tendenci vyskytovat se v akademických textech v trpném rodě: *bylo naznačeno, je vyjádřen, je odvozen, jsou označovány, jsou uvedeny, jsou uváděny, je popsán* apod.

Další skupinu tvoří spojení slovesa *být* s adjektivem nebo substantivem ve jmenném přísudku. Tato skupina má tři podtypy typické pro akademické texty:

— po spojení slovesa *být* s adjektivem následuje vedlejší věta nebo infinitiv: *je patrné že, je zřejmé že, je nezbytné, je obtížné, je podstatné, je nutné*;
— po spojení slovesa *být* s adjektivem následuje typicky substantivum v nominativu: *je charakteristická, jsou typické, jsou přítomny, je základní*;
— spojení slovesa *být* se substantivem v instrumentálu: *je důsledkem, je předmětem, je příkladem, je zdrojem, je výsledkem, je součástí*.

Do poslední skupiny, která je velmi početná, patří plnovýznamová slova, která se pojí s typickou předložkou, spojkou nebo spojovacím výrazem. Takováto spojení můžou být do seznamu akademických frází zařazena ze dvou důvodů. Prvním, méně častým, je daleko vyšší výskyt s danou předložkou nebo spojkou v akademických textech než v textech jiného typu (např. akademické *platit pro* vs. *platit za*, akademické *dělit na* vs. *dělit* bez předložky). Druhým důvodem je, že samo plnovýznamové slovo se výrazně

častěji vyskytuje v akademických textech než v referenčním korpusu: *odvodit z, dospět k, charakterizovat jako, kritéria pro, představu o, sestává z, rozdíl mezi, údaje o, definována jako, (za) předpokladu že, v případě kdy*.

4.4 KOMBINACE SLOVNÍCH DRUHŮ

Nejčastějšími kombinacemi slovních druhů (POS-gramy) jsou spojení dvou sloves (VV, prvním slovesem je pomocné *být*), slovesa s typickou předložkou (VR), předložky a substantiva (RN), substantiva s typickou předložkou (NR), slovesa a adjektiva (VA, pomocné sloveso *být*) nebo slovesa s typickou spojkou (VJ). Mezi trigramy jsou nejčastější kombinace předložka-substantivum-předložka (RNR) a předložka-zájmeno-substantivum (RPN). Příklady takových víceslovných jednotek jsou uvedeny v tabulce 4.

Některé z těchto kombinací se zcela nebo částečně překrývají s výše zmíněnými formálně vymezitelnými skupinami, konkrétně kombinace slovesa být s dalším slovesem (VV) odpovídají slovesům v trpném rodě, skupina plnovýznamových slov následovaných typickým funkčním slovem sestává téměř celá z kombinace slovesa či substantiva s předložkou nebo slovesa se spojkou (VR, NR a VJ). Kombinace předložky a substantiva, resp. předložky, substantiva a předložky (RN a RNR) odpovídají téměř výhradně víceslovným předložkám (zařazeným výše mezi lexical bundles, kap. 4.2).

kombinace slovních druhů (POS-gram)	**počet výskytů**	**příklady**
VV (sloveso-sloveso)	31	*bylo naznačeno, je vyjádřen, je odvozen, jsou označovány, jsou uvedeny, jsou uváděny, je popsán*
VR (sloveso-předložka)	30	*vázána na, založena na, odvodit z, dochází k, (se) používá pro, (se) dělí na, poukazuje na, omezen na, vyplývá z, vede k, platí pro*
RN (předložka--substantivum)	24	*na příkladu, z hlediska, v důsledku, na základě, na principu, v prostoru, pro potřeby, za účelem, ve formě, v kapitole*
NR (substantivum--předložka)	18	*poznatky o, vazby mezi, vztah mezi, souvislost mezi, vztahu k, základ pro, úvahy o, snaha o, vazby na*
VJ (sloveso-spojka)	10	*definována jako, (se) označuje jako, charakterizovat jako, chápat jako, (se) jeví jako, vyplývá že*
RNR (předložka--substantivum-předložka)	9	*ve vztahu k, v závislosti na, v souladu s, s ohledem na, v souvislosti s*
RPN (předložka-zájmeno--substantivum)	7	*v našem případě, v této souvislosti, z tohoto důvodu, ve své podstatě, v této oblasti*

TABULKA 4. Nejčastější kombinace slovních druhů v rámci bigramů a trigramů

5. SHRNUTÍ

V příspěvku jsme představili aplikaci Akalex, která je určena k výzkumům slovní zásoby typické pro akademické texty. V databázové části aplikace je k dispozici seznam téměř 60 tisíc n-gramů — kandidátů na typické akademické jednoslovné nebo víceslovné jednotky, který lze jednoduše prohledávat a třídit podle preferovaných kritérií, jako je například počet oborů, v kterých se n-gram vyskytl, slovní druh členů n-gramu nebo i výskyt konkrétního lemmatu. N-gramy byly do databáze zahrnuty na základě dvou obecných kritérií: vysoká frekvence v akademických textech oproti textům jiných typů a co nejrovnoměrnější výskyt ve velkém množství disciplín (všechny údaje jsou založeny na datech z korpusu SYN2015). Těmito rysy se odlišují od obecné slovní zásoby na jedné straně a od terminologie na straně druhé.

Každá položka představovaného seznamu obsahuje kromě samotného n-gramu i množství informací, které slouží jak k třídění, tak i k posouzení vyhledaných jevů. Pro různé typy výzkumů poslouží různá kritéria. Podle **slovních druhů** lze např. vyhledávat kolokace konkrétního slova, které jsou typické pro akademické texty: hledáme tak např. adjektiva či slovesa, která se pojí s lemmatem *výzkum*.

Poměr frekvencí v akademických a neakademických textech nám napoví, nakolik je daný n-gram charakteristický pro akademické texty, např. n-gram s hodnotou 10 je desetkrát častější v akademických textech ve srovnání s texty publicistickými a beletristickými (*zásadní význam, podstatným faktorem, analýza ukázala*).

Čím vyšší je hodnota **distribuce** v oborech, tím univerzálněji se daný n-gram v akademických textech používá. Údaj o distribuci je vhodné kombinovat s údajem o **disperzi** neboli rovnoměrnosti rozložení v oborech. Mezi velmi rovnoměrně rozložené n-gramy s nejvyšší distribucí patří například *z hlediska, v důsledku, jedním ze základních, v souvislosti s, vzhledem k tomu že*.

Funkčnost aplikace Akalex jsme ověřovali případovou studií. Soustředili jsme se na 2- až 6-gramy, které se vyskytly ve všech 24 akademických oborech dostupných v korpusu SYN2015. Z bezmála 900 kandidátů jsme vybrali 236 víceslovných jednotek typických pro akademické texty. Tyto víceslovné jednotky jsme dále analyzovali a hledali jsme početnější, a tedy pro akademické texty typické skupiny n-gramů, a to na základě sémantických, funkčních a formálních kritérií. Z těch nejhojněji zastoupených typů je třeba jmenovat především lexical bundles (zvl. víceslovné předložky a spojky), dále kolokace plnovýznamových slov, kombinace dvou sloves odhalující typičnost trpného rodu v akademických textech nebo spojení slovesa či substantiva s typickou spojkou nebo předložkou.

Úkolem do budoucna je jistě provést rozsáhlejší výzkum víceslovných jednotek typických pro akademické texty. Velmi důležité je především zaměřit se na jednotky, které se vyskytují i v nižším (ne však velmi nízkém) počtu oborů — lze očekávat, že mírné snížení požadavku na distribuci přinese velký počet zajímavých víceslovných jednotek, jejichž odhalení umožní lepší poznání slovní zásoby českých akademických textů. Dalším směrem, kam by se mohl budoucí výzkum ubírat, je také práce s jednoslovnými jednotkami typickými pro akademické texty (1-gramy v Akalexu). Na těchto základech by bylo možné po vzoru anglických seznamů zmíněných v úvodu

studie vytvořit seznam základní slovní zásoby českých akademických textů obsahující jak jednoslovné, tak víceslovné jednotky.

POUŽITÉ KORPUSY

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

LITERATURA

Biber, D. — Barbieri, F. (2007): Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26, s. 263–286.

Coxhead, A. (2000): A new Academic Word List. *TESOL Quarterly*, 34, s. 213–238.

Gardner, D. — Davies, M. (2014): A New Academic Vocabulary List. *Applied Linguistics*, 35, s. 305–327.

Granger, S. (2017): Academic phraseology: A key ingredient in successful L2 academic literacy. *Oslo Studies in English*, 9, s. 9–27.

http://www.phrasebank.manchester.ac.uk/ (cit. 25. 1. 2020)

Hyland, K. (2012): Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, s. 150–169.

Chen, Y.-H. — Baker, P. (2010): Lexical bundles in L1 and L2 student writing. *Language, learning and technology*, 14, s. 30–49.

Church, K.W. — Hanks, P. (1990): Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, s. 22–29.

Kováříková, D. (2017): *Kvantitativní charakteristiky termínů*. Praha: Nakladatelství Lidové noviny.

Kováříková, D. — Kovářík, O. (2019): Automatic Identification of Academic Phrases for Czech. *International Conference on Computational and Corpus-Based Phraseology*, Malaga: Springer, s. 227–238.

Kováříková, D. — Kovářík, O. (2021): *Akalex. Nástroj pro výzkum slovní zásoby akademické češtiny*. Praha: FFUK. Dostupné z www: https://www.korpus.cz/akalex.

Manning, C. D. — Schütze, H. (2000): *Foundations of Statistical Natural Language Processing*. Cambridge/London: The MIT Press, s. 267–271.

Morley, J.: *Academic Phrasebank*. Dostupné z:

Nagy, W. — Townsen, D. (2012): Words as Tools: Learning Academic Vocabulary as Language Acquisition. *Reading Research Quarterly*, 47, s. 91–108.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London & New-York: Continuum, s. 56–58.

Simpson-Vlach, R. — Ellis, N. (2010): An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31, s. 487–512.

Dominika Kováříková | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy | Panská 7, 110 00 Praha 1
ORCID ID: 0000-0002-4419-6901
dominika.kovarikova@ff.cuni.cz

Oleg Kovářík | Datamole, s. r. o. | Banskobystrická 11, 160 00, Praha 6
oleg.kovarik@gmail.com

Lucie Lukešová | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy | Panská 7, 110 00 Praha 1
ORCID ID: 0000-0003-1855-7141
lucie.lukesova@ff.cuni.cz