OPEN ACCESS

Projekt Lexikálně-sémantické databáze češtiny (LSD-Czech): uživatelský pojmový slovník a online databáze

Zora Obstová — Ondřej Tichý — Aleš Klégr

ABSTRACT:
The LSD Project: A Czech Online Onomasiological Dictionary and Database
The paper introduces the upcoming outputs of the Lexico-Semantic Database Project (LSD-Czech) supported by The Technology Agency of the Czech Republic (TA CR) [TL02000041]: a Czech online conceptual dictionary and a lexico-semantic database. It reviews the phases of the implementation of the project and its results to be made available by the end of 2022. It describes both the dictionary and the database (from which the dictionary is generated). While the dictionary is intended for the general user, the database will serve linguists and experts in natural language processing and language data application. The paper briefly discusses the field of onomasiological lexicography, focusing on what is an onomasiological dictionary, what dictionaries of this type exist for Czech, and introduces the two dictionaries on which the database is founded, Haller's *Český slovník věcný a synonymický* (1969–86) and Klégr's *Tezaurus jazyka českého* (2007). The focus of the paper is on the description of the functions and features of the dictionary, the steps whereby to search in it and the envisaged future improvements. Finally, it gives the specifications of the database and the next steps planned in its development.

KLÍČOVÁ SLOVA / KEY WORDS:
onomaziologická lexikografie, tezaurus, lexikálně-sémantická databáze, digitalizace, čeština
onomasiological lexicography, thesaurus, lexico-semantic database, digitization, Czech

1 ÚVODEM

Cílem článku je seznámit čtenáře s hlavními připravovanými výstupy projektu Lexikálně-sémantické databáze češtiny (LSD-Czech), podpořeného Technologickou agenturou České republiky [TL02000041]. Představíme zde tedy jednak digitální pojmový slovník určený běžnému uživateli, jednak digitální lexikálně-sémantickou databázi (z níž je slovník generován), která bude sloužit lingvistům a odborníkům v oblasti zpracovávání přirozeného jazyka a aplikace jazykových dat. Než však přejdeme k popisu obou výstupů, považujeme za důležité uvést zejména čtenáře z řad budoucích „laických" uživatelů databáze do — obvykle poněkud opomíjené — problematiky onomaziologické lexikografie. V oddílu 2 se tedy nejprve zaměříme na to, co je to pojmový slovník, jaké slovníky tohoto typu pro češtinu existují, ze kterých slovníků aplikace Lexikálně-sémantické databáze vznikla a co v ní uživatel najde. V oddílu 3 popíšeme průběh realizace projektu, jehož výsledkem oba výstupy jsou, a v částech následujících potom nastíníme konkrétní podobu digitálního slovníku a postup

vyhledávání v něm. Zmíníme také parametry lexikálně-sémantické databáze i další kroky, které jsou v souvislosti s oběma výstupy plánovány.

OPEN ACCESS

2 POJMOVÉ SLOVNÍKY A ČEŠTINA

2.1 ONOMAZIOLOGICKÉ SLOVNÍKY A JEJICH TYPY

Pojmový slovník (někdy označovaný také jako tematický, věcný či ideologický) spadá pod tzv. onomaziologické slovníky. Z typologického hlediska jde o slovníky, které třídí lexikální jednotky (slova, fráze, idiomy) nikoli podle jejich formy (jak to činí slovníky abecední), nýbrž podle jejich významu, pojmového obsahu. Zatímco abecedně uspořádané výkladové slovníky (patřící mezi slovníky sémaziologické) mají za cíl přiřadit slovům (formám) odpovídající význam, smyslem onomaziologických slovníků je ukázat, jakými slovy lze ten který význam, pojem vyjádřit, pojmenovat.

Onomaziologická lexikografie (srov. Svensén, 1993; Hartmann & James, 1998; van Sterkenburg, 2003) rozlišuje několik druhů těchto slovníků. U nás neznámější jsou slovníky synonym, které předkládají synonymní řady či hnízda řazená podle nejobecnějšího nebo nejčastějšího výrazu v daném hnízdě (pro snadné vyhledávání tvoří tyto reprezentativní výrazy záhlaví slovníku a jsou řazeny abecedně). Historicky existovaly slovníky synonym výkladové (vysvětlující rozdíly mezi synonymy), dnes převládají slovníky čistě výčtové. Variantou toho slovníku je slovník antonym, který často bývá součástí synonymních slovníků (antonymie je vlastně druh podobnosti, při níž má pouze klíčový sémantický rys opačnou hodnotu). Dalším typem onomaziologického slovníku jsou slovníky obrazové. Pojmy (většinou označující konkréta) jsou prezentovány formou obrázků (popř. fotografií), typicky uspořádaných do širších okruhů, tematických panelů (např. kuchyň, její součásti a kuchyňské náčiní, nemocnice, přístav, automobil, lidské tělo apod.). Mezi nejznámější moderní obrazové slovníky (v mnoha vydáních) patří *Bildwörterbuch* německého nakladatelství Duden.

Odlišným typem je tzv. opačný slovník (angl. *reverse* or *word-finding dictionary*). Nejde však o slovník se slovy řazenými podle abecedy od posledního písmena, rovněž označovaný jako reverzní (či slovník *a tergo*), ani o starší typ slovníků kolokací uváděných pod názvem *word finders*. Opačný slovník, jehož nejznámějším představitelem je *Reader's Digest Reverse Dictionary* (Kahn, 1989), je vlastně výkladový slovník naruby. Heslo začíná definicí (popř. definicí synonymem) a končí záhlavím (např. „about to happen IMMINENT, IMPENDING"; „go along with someone's wishes or ideas HUMOUR, INDULGE"). Navíc slovník obsahuje prvky obrazového slovníku (např. obrázek s popisem FULL-RIGGED SAILING SHIP) a pojmového slovníku (výčty příbuzných slov, např. FENCING TERMS, SCIENTIFIC INSTRUMENTS).

Konečně vlastní pojmový slovník v užším smyslu je slovník, který předkládá systém pojmů a ke každému pojmu je připojen výčet slov, jimiž je v daném jazyce označován, ať už formou substantiva, adjektiva, nebo slovesa. Pojmová hesla obsahují vlastně lexikální pole různého typu — hierarchická (zahrnující (ko)hyponyma, např. *les*, *bor*, *dubina*, *bučina*, a meronyma, např. *strom*, *větev*, *kořen*), lineární (bipolární, např. *samec–samice*, škály, např. *začátek*, *střed*, *konec*, nebo cykly, např. barvy,

OPEN ACCESS

roční období) a nestrukturované klastry (synonyma, případně i antonyma). Nejbohatší tradici mají pojmové slovníky anglické a německé.

2.2 ČESKÉ ONOMAZIOLOGICKÉ SLOVNÍKY

Pro češtinu vznikly všechny uvedené druhy onomaziologických slovníků s výjimkou slovníku reverzního (opačného). Patrně nejstarším specializovaným slovníkem synonym v češtině je *Stručný slovník českých synonym* J. Mašína a J. V. Bečky (1947), který následně Bečka vydal v rozšířené podobě jako *Slovník synonym a frazeologismů* (1977/1982). Později se k němu přidal *Slovník českých synonym* K. Paly a J. Všianského (1994/2008). Zatím posledním tištěným slovníkem tohoto typu je *Slovník českých synonym a antonym* (Lingea, 2007/2012). Nakladatelství Lingea zpřístupnilo i jeho online verzi (2.0) v rámci aplikace nechybujte.cz (obsahující i výkladový slovník, pravidla pravopisu a gramatiku). Existují i další méně propracované online aplikace (ABZ slovník českých synonym).

Za předchůdce obrazových slovníků lze považovat Komenského učebnici *Orbis pictus* (1658/1896). V cizojazyčné výuce se u nás tyto slovníky uplatňují i dnes. Po válce vyšel drobný, Dudenem inspirovaný Gallerův a Mrázkův *English in Pictures — Anglický obrazový slovník* (1947). Další se objevují od 90. let, např. *Obrázkový anglicko-český slovník* (Davies & Bezděková, 1993), *Velký obrazový tematický slovník česko-slovensko-anglicko-německý* (Archambaultová & Corbeil, 1999) a z něj odvozené menší slovníky. Vedle nich byla publikována celá řada takovýchto slovníků určených pro děti s cílem rozšiřovat jejich slovní zásobu (např. *Velký dětský obrázkový slovník*, Brauerová, 2018).

Pojmové slovníky vyšly v češtině dva, což je u „malého" jazyka, jakým je čeština, poměrně výjimečný jev a staví nás to na roveň nevelkému počtu jazyků, které mají své vlastní tezaury. Prvním z nich je Hallerův *Český slovník věcný a synonymický* (1969–86), který vychází z německého vzoru, a druhým je *Tezaurus jazyka českého* (Klégr, 2007), založený na Rogetově anglickém *Thesauru* (1852). Ani jeden z nich není v současnosti běžně ke koupi, dostupné jsou pouze v knihovnách a dotisky nelze očekávat. Jejich digitalizace v rámci zde představovaného projektu je tedy vedena mimo jiné i snahou prodloužit jejich život do 21. století a s jejich pomocí uchovat a dále rozvíjet bohatství českého lexika.

2.3 *ČESKÝ SLOVNÍK VĚCNÝ A SYNONYMICKÝ* (ČSVS)

Práce na *Českém slovníku věcném a synonymickém* (Haller, 1969–1986) byly započaty z podnětu překladatelské sekce Svazu československých spisovatelů a byly svěřeny Jiřímu Hallerovi. Předsedou redakční rady byl Vladimír Šmilauer. Slovník, pojatý především jako „jazyková pomůcka pro překladatele, spisovatele a novináře", ale také pro školu a širší veřejnost, si kladl za cíl „poskytnout ve věcném uspořádání co největší výběr synonymických výrazů jako možný repertoár vyjadřování i jako východisko k objevování nových vyjadřovacích možností českého jazyka" (Haller, 1969–1977, I. díl, Úvodní poznámky, s. V). Kromě toho však sledoval i cíl jazykovědný, totiž shromáždit a věcně utřídit českou slovní zásobu.

Jako základní osnova díla posloužil Halligův a von Wartburgův *Begriffssystem als Grundlage für die Lexikographie* (2. vydání z roku 1963), teoreticky důkladně propracované

OPEN ACCESS

klasifikační schéma, v němž nejsou uspořádána slova, nýbrž na jazykovém materiálu nezávislé pojmy, a které tedy může být dle názoru jeho autorů aplikováno na kterýkoli konkrétní jazyk (tamtéž, s. XXII).[1] Třebaže někteří badatelé považují takový záměr za utopický (srov. např. Wiegand, 2004, s. 68) a objevují se i tvrzení, že *Begriffssystem* nebyl nikdy prakticky realizován (Fischer, 2004, s. 46), je třeba připomenout, že na jeho základě vznikla řada specializovaných slovníků, zejména z románské oblasti.[2] Jediným pokusem uspořádat podle tohoto schématu veškerou slovní zásobu nějakého konkrétního jazyka však byl, alespoň pokud je nám známo, právě *Český slovník věcný a synonymický*.

Obrovský rozsah připravovaného slovníku, jeho složitá hierarchická struktura (až 5 úrovní) a encyklopedický charakter však způsobily, že po předčasné smrti Jiřího Hallera zůstalo dílo nedokončeno. Tři svazky, publikované v letech 1969, 1974 a 1977, obsahují tyto oddíly nejvyšší úrovně hierarchie: A. Vesmír, svět kolem nás; B. I. Člověk, bytost tělesná a B. II. Duševní stránka člověka; v roce 1986 byly doplněny abecedním rejstříkem. Nevydány (a z velké části nezpracovány) zůstaly oddíly B. III. Člověk, bytost společenská a C. Člověk a univerzum. Přesto však ČSVS, který čerpá lexikální materiál především z *Příručního slovníku jazyka českého* (1935–1957) a *Slovníku spisovného jazyka českého* (Havránek et al., 1960–71), obsahuje úctyhodné množství lexémů: na 1594 stranách je do 3193 hesel rozřazeno přes 400 000 jednoslovných i víceslovných lexikálních jednotek. Obsažené výrazy jsou doplněny kolokacemi, příkladovými větami, někdy i definicemi a stylisticky charakterizovány pomocí široké palety zkratek a značek.

Zatímco makrostruktura až na výjimky respektuje Halligovo a von Wartburgovo schéma, mikrostruktura, pro niž abstraktní a na jazyce nezávislý *Begriffssystem* žádné vodítko nenabízí, je značně proměnlivá, nekonzistentní a přizpůsobuje se konkrétnímu slovnímu materiálu.[3] Tato proměnlivost je vzhledem k povaze a rozsahu materiálu jistě pochopitelná a ospravedlnitelná, výrazně však ztěžuje digitální zpracování hesel v rámci zde popisovaného projektu a vyžádala si užití speciálního programu (viz níže) i zvýšeného úsilí při korekci naskenovaného textu.[4]

1 Aby byla nezávislost pojmů na konkrétním jazykovém materiálu garantována, vycházeli autoři při tvorbě své klasifikační struktury nikoli z rodné němčiny, nýbrž z francouzštiny (Hallig i von Wartburg byli romanisté).

2 Např. některé díly von Wartburgova *Francouzského etymologického slovníku* (von Wartburg, 1922–2003). Další slovníky i podrobnější informace o Halligově a von Wartburgově klasifikačním schématu uvádí Obstová (2020, s. 101–102).

3 To ostatně v Úvodních poznámkách deklaruje i sám Haller: „Ani ve zpracování hesel jsme neusilovali o nějakou uniformitu; každé heslo je uspořádáno podle svého vlastního rázu, daného příslušným slovním materiálem, jeho povahou i jeho rozsahem“ (Haller, 1969–77, I. díl, s. VII).

4 ČSVS vznikal v 70. letech minulého století, a je proto zřejmé, že ne všechny informace v něm obsažené jsou dnes zcela aktuální; to se týká nejen vlastního lexika, ale také např. stylových charakteristik. Aktualizace takto rozsáhlého díla by si vyžádala dlouhé roky práce, a je tudíž mimo možnosti realizovaného projektu. Jistě o ní však lze uvažovat do budoucna v rámci případného rozšiřování databáze (viz níže 4.1.3).

OPEN ACCESS

2.4 *TEZAURUS JAZYKA ČESKÉHO* (TJČ)

Východiskem pro *Tezaurus jazyka českého. Slovník českých slov a frází souznačných, blízkých a příbuzných* (Klégr, 2007) bylo dílo publikované před 170 lety v Anglii. Šlo o *Thesaurus of English Words and Phrases* (Roget, 1852), jehož autorem byl lékař a profesor fyziologie Peter Mark Roget (1779–1869), věnující se i řadě přírodovědných oborů. V 26 letech si pro svou potřebu vytvořil malý katalog slov, který používal při psaní. V důchodu se k této myšlence vrátil, rozpracoval ji a po čtyřech letech dílo dokončil. Za jeho života vyšel *Thesaurus* 28krát. Rodina Rogetů si podržela copyright a obsah *Thesauru* úspěšně tříbila a doplňovala až do roku 1952. Historie *Thesauru* prokázala jeho užitečnost a nesmírnou oblibu. Do současnosti se prodalo přes 32 milionů výtisků a z britské linie se později oddělily v nejrůznějších podobách americká a následně i australská verze. Na tomto slovníku lze ukázat, že podobné dílo nevzniká najednou (zatímco vydání z roku 1852 mělo kolem 15 000 slov, současná vydání obsahují přes 300 000 slov) a také že klíčem je nosná, časem prověřená koncepce. Ta u *Thesauru* zůstává po 170 letech bez podstatnějších modifikací navzdory změnám historicko-společenským, vědecko-technickým i jiným.

Podstatou *Thesauru* je klasifikační schéma, které Roget odvodil z přírodovědných taxonomií. Makrostrukturu tvoří několik hierarchicky uspořádaných rovin. Na nejobecnější rovině je obsah rozdělen do šesti tříd: *abstraktní vztahy*, *prostor*, *hmota*, *intelekt — užívání mysli* (s oddíly tvoření a sdílení myšlenek), *volní oblast — uplatnění vůle* (s oddíly volní jednání individuální a v sociálních skupinách) a *emoce*. Třídy jsou rozčleněny na 39 sekcí (například třída *hmota* na sekce *hmota obecně*, *anorganická* a *organická hmota*) a sekce na hlavy (vlastní hesla). Vedle tohoto vertikálního členění je makrostruktura uspořádána ještě horizontálně na úrovni hesel (pojmů). To znamená, že hesla jsou v rámci sekcí řazena za sebou do kontrastních (opozitních) dvojic (*shoda*, *neshoda*), případně trojic (*touha*, *lhostejnost*, *averze*), škál (*začátek*, *střed*, *konec*) či cyklů (barvy).

Mikrostruktura (vnitřní uspořádání) hesel spočívá v rozdělení slov spojených s daným pojmem podle slovního druhu na část substantivní, adjektivní, verbální a adverbiální. Jednotlivé slovní druhy jsou dále členěny na významové podskupiny (vlastnost, činnost, nositel vlastnosti, agens, patiens apod.), tedy podhesla. Tyto významové podskupiny představují lexikální pole ((ko)hyponym, meronym a synonym) spojená příslušným pojmem. Hesla jsou rovněž opatřena odkazy na hesla jiná. Součástí tištěného *Thesauru* je abecední rejstřík usnadňující vyhledávání. Nehledě na klasifikační schéma jsou hesla v tomto slovníku jednoduchá a jejich výhodou je jednotné a konzistentní zpracování, které značně ulehčilo digitalizaci.

Český *Tezaurus* implicitně převzal Rogetovu makrostrukturu, reflektovanou ve výběru a pořadí hesel, která vycházejí ze zkráceného vydání (Carney & Waite, 1985), a převzal i jejich mikrostrukturu. V českém *Tezaur*u jsou explicitně vyznačeny pouze nejvyšší a nejnižší rovina této hierarchie: třídy (I–VI) a hesla (1–885) a v rámci nich jsou pak kurzívou odlišena podhesla. Ostatní hierarchické roviny, (pod)sekce, jsou zachovány v pořadí hesel, tj. sled hesel odpovídající příslušné sekci je dodržen i v českém *Tezauru* (bylo by tedy možné zde sekce bez potíží zpětně doplnit). I když jsou sekce důležité pro logickou výstavbu slovníku a pořadí hesel, pro vyhledávání slov a orientaci ve slovníku zásadní nejsou. Oproti klasickému anglickému *Thesauru* se ten

OPEN ACCESS

český liší i menším počtem hesel, obsahuje 885 záhlaví (původní slovník měl přesně 1000 hesel, v moderní redakci 990). Neznamená to ovšem, že by se snížil počet pojmů v českém slovníku zachycených; došlo k tomu tak, že některá hesla byla sloučena (a naopak se zvýšil počet podhesel). Stejně je tomu i ve zkrácené britské verzi *Thesauru* (pojmová struktura jazyka je obtížně uchopitelná oblast, nejde o exaktní záležitost). Obsah českých hesel přitom nevznikal překladem anglického originálu, nýbrž vychází z významu (pod)záhlaví a příslušná lexikální pole byla sestavována podle situace v češtině. Také proto byla zvolena zkrácená verze, která poskytuje prostor pro tvorbu hesel nezávisle na angličtině. Účel Rogetova *Thesauru* vysvětluje jeho podtitul: *to facilitate the expression of ideas and assist in literary composition*. *Tezaurus* je tedy míněn jako praktická stylistická příručka rozvíjející přesné a bohaté vyjadřování, nikoli jako encyklopedie našich znalostí o světě, jak je tomu u ČSVS. Z toho také vyplývá, že oba slovníky se překrývají jen částečně.

3 PROJEKT LSD

3.1 CHARAKTERISTIKA A CÍLE PROJEKTU

Cílem projektu Lexikálně-sémantické databáze češtiny (LSD-Czech), podpořeného Technologickou agenturou České republiky [TL02000041], je vytvoření rozsáhlé elektronické databáze založené na vytěžení lexika a struktur dvou českých tištěných onomaziologických slovníků, popsaných výše: *Tezauru jazyka českého* (Klégr, 2007) a *Českého slovníku věcného a synonymického* (Haller, 1969–86). Projekt reaguje na potřeby dvou skupin uživatelů. Tou první je široká obec nelingvistů vytvářejících komunikáty v češtině, především spisovatelů, překladatelů, pracovníků médií, marketingu či státní správy, ale i učitelů, studentů[5] a veřejnosti. Současná slovní zásoba češtiny — na rozdíl od většiny velkých evropských jazyků — totiž dosud není souhrnně zpracována pro účely jazykové produkce, a školní výuka i běžná jazyková praxe postrádají konzistentnější oporu, která by umožnila rozvoj komunikačních dovedností a tvorbu významově precizních a výrazově bohatých textů. Onomaziologické slovníky napomáhají tvorbě kvalitních textů tím, že vedou uživatele od významu k formě. Vyhledávání ve věcně uspořádaných tištěných slovnících je však poměrně náročné; to je pravděpodobně jedním z důvodů, proč bývají — navzdory tomu, že jsou obvykle vybaveny abecedním rejstříkem — využívány méně často než slovníky sémaziologické. Zpřístupnění dvou českých tezaurů v digitální podobě bylo chápáno jako příležitost optimálně využít jejich obrovský potenciál: otevře uživatelům nové cesty k lexikálnímu bohatství češtiny a nabídne nové pohledy na české lexikum jako celek, jeho strukturu, členitost a provázanost.

Druhou skupinu uživatelů tvoří lingvisté a další odborníci zabývající se výzkumem češtiny, korpusovou lingvistikou, zpracováním přirozeného jazyka, strojovým zpracováním textů a strojovým učením. Smyslem databáze je poskytnout datovou

5 Databázi nepochybně využijí i cizinci osvojující si češtinu (víme například, že *Tezaurus jazyka českého* se uplatnil při univerzitní výuce češtiny pro cizince ve Finsku).

OPEN ACCESS

a strukturní základnu pro projekty sémantického značkování či strojového porozumění textu; zároveň bude možné porovnávat data vytěžená z obou onomaziologických slovníků s daty Českého národního korpusu. Do budoucna počítáme i s průběžnou aktualizací databáze, která tak umožní uchovat a zhodnotit výsledky desítek let práce českých lexikografů.

3.2 REALIZACE PROJEKTU

Projekt, na jehož realizaci se pod vedením Aleše Klégra podílejí dvě pracoviště, Filozofická fakulta Univerzity Karlovy a Ústav pro jazyk český AVČR, byl původně plánován na tři roky (únor 2019 — únor 2022); vzhledem k technickým problémům vzniklým zejména v důsledku koronavirové pandemie byl však prodloužen do konce roku 2022. Zpětnou vazbu k jednotlivým fázím projektu, které popíšeme níže, poskytuje pět aplikačních garantů zastupujících významné skupiny budoucích uživatelů databáze (uvedeno v abecedním pořadí): Česká asociace pro digitální humanitní vědy, z. s., Jednota tlumočníků a překladatelů, Obec překladatelů, Sjednocená organizace nevidomých a slabozrakých České republiky, z. s.[6] a Ústav formální a aplikované lingvistiky MFF UK. Realizace projektu byla rozčleněna do několika kroků, které se zčásti prolínají a zčásti na sebe navazují (tyto kroky byly detailněji popsány již jinde, viz Tichý et al., 2021).

1. V první fázi proběhla detailní analýza struktur zdrojových slovníků. Je třeba připomenout, že oba slovníky se — vzhledem k různé době svého vzniku (dělí je téměř čtyřicet let), rozdílně koncipované makrostruktuře a mikrostruktuře i tomu, že ČSVS zůstal nedokončen — překrývají jen částečně, a naopak se významně doplňují. Aby bylo možno propojit dva takto odlišné zdroje do jednoho vyššího celku a vytvořit software, který by umožnil vyhledávání v obou dílech zároveň a smysluplnou korelaci jejich kategorií i hesel, bylo třeba zmapovat sémantickou strukturu slovníků, porovnat zpracování jednotlivých hesel a v neposlední řadě také rozhodnout, které z informací — zejména z velmi obsažného ČSVS — zůstanou zachovány i v elektronické verzi, a které budou vypuštěny.

2. Zároveň s analýzou struktur probíhalo skenování obou papírových slovníků, automatické rozpoznání znaků (OCR) a poloautomatická desambiguace rozpoznaných typografických prvků.

3. Třetím krokem byla poloautomatická transformace digitalizovaných struktur a dat zdrojových slovníků do výsledné struktury ve formátu TEI-XML, navržené v prvním kroku, s tím, že původní struktury zdrojových slovníků zůstanou paralelně zachovány. V této fázi se ovšem ukázalo, že data ČSVS jsou natolik komplexní a nekonzistentní, že by zpracování pomocí původně plánovaných tradičnějších nástrojů (skriptování na základě pravidel) vedlo k nejistým výsledkům, a především by si vyžádalo velmi rozsáhlé manuální opravy. Rozhodli jsme se proto značkování slovníkových dat ČSVS zpřesnit a urychlit využitím technologie strojového učení: našim potřebám nejlépe vyhovoval projekt GROBID-Dictionaries (více informací viz

6 Webové rozhraní bude přizpůsobeno tak, aby odpovídalo i potřebám nevidomých a slabozrakých.

OPEN ACCESS

Khemakhem et al., 2017; Khemakhem et al., 2018). Tento nástroj ovšem nebyl navržen specificky pro onomaziologické slovníky, a proto musel být pro účely naší databáze ve spolupráci s autorem upraven. Po úpravách byl nástroj natrénován na několika desítkách ručně opravených stran. Díky této technologii jsme získali mnohem preciznější soubor dat, i ten však musel být podroben systematické manuální kontrole,[7] do níž byli zapojeni studenti a doktorandi FFUK.

4. Souběžně s činnostmi popsanými v bodech 2 a 3 probíhalo programování a návrh softwarových nástrojů. Tento souběh byl v souladu s harmonogramem projektu, značné problémy však způsobila data ČSVS, jejichž zpracování si vyžádalo delší čas a jejichž struktura se proměňovala ještě ve fázi manuálních oprav. Vzhledem k nekonzistenci slovníku jsme totiž opakovaně objevovali neočekávané strukturní prvky (ojedinělé typy podhesel, graficky odlišené doplňující specifické informace včetně např. taxonomických seznamů a tabulek ap.), s nimiž jsme se museli v plánované struktuře vyrovnávat. Takovéto proměnlivé datové schéma pochopitelně komplikovalo programování samotné online aplikace.

Technické řešení (podrobně popsáno v Tichý et al., 2021) bylo realizováno v programovacím jazyce JavaScript, a to jak na straně serveru, tak na straně klienta. Data jsou vzhledem k formátu TEI-XML uložena v databázi baseX, na kterou je pomocí API (Application Programming Interface, tj. rozhraní pro vzájemnou komunikaci aplikací) a xQuery napojena modulární struktura online aplikace. Přímo s databází komunikuje především modul Thesaurus, který provádí potřebné transformace dat a tvoří tak základní část aplikace. Díky API je přes něj možné čerpat data do aplikací třetích stran, zejména ale komunikuje s modulem zajišťujícím standardní uživatelské rozhraní. Modul uživatelského rozhraní je naprogramován tak, aby umožnil kompilaci na straně serveru. To zajistí dostatečnou flexibilitu a rychlost uživatelského rozhraní, ale i dobrou dostupnost například ze strany nástrojů pro indexaci.

5. Poslední krok představuje testování online aplikace, jehož se účastní řešitelé projektu, spolupracující studenti, zástupci aplikačních garantů a také veřejnost.

4 VÝSTUPY PROJEKTU

Základní výstup projektu Lexikálně-sémantická databáze češtiny má dvě konkrétní formy: digitální pojmový slovník pro veřejnost a online databázi určenou pro teoretický i aplikovaný lexikálně-sémantický výzkum prostřednictvím API pro lingvisty, specialisty v oblasti počítačového zpracování přirozeného jazyka (NLP) a další odborníky na aplikaci jazykových dat. Každá z těchto forem využití představuje odlišnou problematiku, která vyžaduje odlišný přístup a řešení, a proto budou popsány zvlášť.

7 Jak už bylo zmíněno výše v odd. 2.4, pozn. 4, data získaná z ČSVS nebyla nijak aktualizována; zachována je i původní ortografická podoba slov, která tak mnohdy neodpovídá současné ortografické normě. Aktualizace či doplňování slovníkových dat však může být zajímavou výzvou do budoucna, viz níže 4.1.3.

OPEN ACCESS

4.1 DIGITÁLNÍ SLOVNÍK PRO VEŘEJNOST

Tato aplikace bude volně přístupná na webové stránce www.najdislovo.cz (hostované na serverové infrastruktuře FF UK) a odkaz na ni bude i na webových stránkách Ústavu pro jazyk český, Českého národního korpusu i na stránkách všech výše zmíněných aplikačních garantů.

4.1.1 POPIS VYHLEDÁVÁNÍ V DIGITÁLNÍM SLOVNÍKU

Oba digitalizované slovníky (TJČ a ČSVS) zpracované v aplikaci LSD mají hodně společného, ale poněkud se liší nejen svou strukturou, ale zčásti i svým pojetím (viz výše 2.3 a 2.4). Oba nicméně slouží stejnému praktickému účelu. Jde o slovníky produkční, referenční příručky určené k psaní textů, jemuž napomáhají třemi způsoby. V první řadě uživateli poskytují výběr synonym ke zvýšení stylistické úrovně textu. Druhou funkci lze popsat jako mnemotechnickou: slovník může pomoci ve chvíli, kdy máme to správné slovo „na jazyku", ale nemůžeme si ho vybavit. A protože slovník obsahuje širší okruh slov, než jsou jen synonyma (např. pojmy hyperonymní, kohyponymní, hyponymní, kauzálně spojené atd.), stačí začít od prvního slova, které nás napadne, a dříve nebo později nás aplikace dovede k tomu, co hledáme. Tento slovník může být konečně i nástrojem tzv. „laterálního myšlení": díky tomu, že nás zavádí mezi slova spojená různými vztahy, dává nám možnost — na rozdíl od sekvenčního logického myšlení — kreativně vyjádřit daný obsah i jinak, z jiného úhlu, než jsme původně zamýšleli.

Na úvodní stránce aplikace najdeme především základní vyhledávací okno (*input box*, viz níže). Stránka ovšem umožňuje zobrazit i obecné informace o databázi a jejím fungování i o obou digitalizovaných slovnících. Ve spodní části úvodní strany jsou také zobrazeny hierarchické struktury, podle nichž jsou oba slovníky uspořádány. V případě českého *Tezauru*, jak již bylo zmíněno, je hierarchie zredukována (je implicitní) a vyhledávat podle ní nelze. Naopak u Hallerova slovníku může tato hierarchie uživateli vyhledávání konkrétních slov usnadnit: jednak je díky ní možné vyhledávání snadno omezit, resp. vybrat si z výsledků jen ty výrazy, které se týkají příslušné sémantické oblasti dané částí hierarchie (např. *koruna* v části hierarchie *peníze* vs. *rostlina*), jednak lze hierarchií přímo procházet a hledat slovo podobně, jako když hrajeme hru na hádání pojmů („Je to abstraktní, nebo konkrétní?", „Je to rostlina, nebo živočich?" atp.).

Vyhledávací okno je na úvodní straně jediné a je společné pro oba slovníky. Umožňuje fulltextové vyhledávání v obou slovnících zvlášť, případně i zároveň. Po zadání hledaného slova se zobrazí názvy a čísla hesel v obou slovnících, které dané slovo obsahují, seřazené podle relevance. V této fázi se uživatel může rozhodnout, zda bude dále pracovat pouze s jedním ze slovníků, anebo s oběma současně.

Ukažme si na konkrétním příkladě, jakým způsobem probíhá hledání například v *Tezauru*. Základním nástrojem aplikace je celotextové vyhledávání citlivé na tvar slova (vyhledávána jsou celá slova, aby seznam nebyl neúměrně dlouhý, uživatel má ale možnost vyhledávat jen části slov). Hledané slovo zapíšeme do vyhledávacího okna a aplikace najde všechny jeho výskyty v *Tezauru*. Jak již bylo řečeno, na rozdíl od prostých synonymických slovníků ovšem Lexikálně-sémantická databáze nenabídne jen synonyma daného slova, nýbrž všechna lexikální pole slov s podobným či

OPEN ACCESS

příbuzným významem, v nichž se slovo vyskytuje. Zpravidla bude hledané slovo součástí několika (či dokonce mnoha) lexikálních polí zároveň.

Přehled míst (adres), na kterých se slovo v *Tezauru* vyskytuje, se zobrazí v seznamovém okně (našeptávači), které se automaticky otevře a začne plnit výsledky během zadávání vyhledávaného slova. Výskyty jsou v něm seřazeny podle relevance. To znamená, že výše je řazena přesná shoda tvaru slova než jen shoda částečná, výše jsou řazeny výskyty, kdy je hledané slovo shodné s názvem části hierarchie či názvem hesla (viz následující příklad vyhledávání slova *pocit*), a dále jsou výskyty řazeny postupně, tak jak jsou seřazena hesla v tištěném slovníku. Položky či řádky v seznamovém okně jsou „adresy" míst, kde se slovo ve slovníku vyskytuje. Díky jednoduché struktuře českého *Tezauru* stačí k lokalizaci hledaného slova tři údaje: TŘÍDA (I–VI název třídy) — HESLO (1–885 název hesla) — PODHESLO. Vzhledem k tomu, že se název hesla a název prvního podhesla mohou shodovat, může adresa obsahovat stejný výraz. Zadáme-li třeba slovo *pocit*, objeví se na prvním místě v seznamu adresa: *VI Emoce > 730 Pocit > pocit*. A protože se slovo *pocit* v tomto hesle vyskytuje i v dalších podheslech, objeví se také následující adresy: *VI Emoce > 730 Pocit > kategorie citů*, *VI Emoce > 730 Pocit > soucit*, *VI Emoce > 730 Pocit > soucítit*. Po těchto primárních adresách (v nichž se hledané slovo a název hesla shodují) následuje seznam ostatních adres, např. *I Abstraktní vztahy > 99 Předešlost > předtucha*, *I Abstraktní vztahy > 99 Předešlost > předvídat*, *II Prostor > 281 Neklidnost > nervozita*, *III Hmota > 334 Tělesný pocit > tělesný pocit* a další. V každém podhesle, to jest na každé adrese, se hledané slovo může vyskytnout vícekrát. V takovém případě je počet výskytů indikován číslem v závorce za nalezenou adresou.

Po zvolení adresy, a tedy příslušného řádku, se v seznamovém okně otevře celé heslo (a tedy nejen příslušné podheslo), aby si hledající mohl prohlédnout všechna slova v hesle a vybrat si to, které potřebuje. Přestože podheslo významně zužuje oblast hledání, bude pro rychlejší orientaci hledané slovo (či slova) v (pod)hesle zvýrazněno (podsvíceno).

Uživatel má možnost vrátit se z otevřeného hesla zpět do seznamu adres v seznamovém okně a zvolit si další adresy spojené s hledaným slovem. Kromě toho si může prohlédnout i hesla předcházející a následující vzhledem k heslu aktuálně otevřenému a funkční budou i odkazy (umístěné většinou na konci hesel) k heslům nacházejícím se jinde ve slovníku. Slovník tak uživatele provádí složitou spletí pojmů a významů a inspiruje ho nejen k nalezení vhodného slova, ale mnohdy i k tomu, aby svou myšlenku formuloval jiným způsobem, než původně zamýšlel.

Zapsání nového slova do vyhledávacího okna zruší předchozí volbu a opět se otevře seznamové okno (našeptávač) s výčtem adres výskytů nově hledaného slova. Na konci každého hesla v *Tezauru* uživatel najde odkazy na hesla obsahově příbuzná, která si může vyhledat tak, že na ně klikne či je zadá do vyhledávacího okna.

Kliknutím na ikonu „domů" na liště se lze vrátit do výchozího stavu aplikace. Pomocí ikon (popis jejich funkce se objeví, nastaví-li se na ikonu kurzor) bude možné provádět další úkony, jako například měnit formát zobrazení hesel (standardní a kompaktní — pro zkušenější uživatele), měnit velikost písma a rozložení textu a přepínat mezi oběma slovníky.

Obdobným způsobem se provádí vyhledávání i v Hallerově ČSVS. Při hledání v ČSVS si uživatel navíc bude moci zobrazit i naskenovaný text původního slovníku

OPEN ACCESS

a porovnat digitalizovaná hesla s těmi původními, tištěnými, která jsou oproti digitalizované formě detailněji vnitřně rozrůzněna odlišnými typy typografie.

4.1.2 SOUČASNÝ STAV ZPRACOVÁNÍ

Na podzim roku 2022 jsou z podstatné části dokončeny první čtyři z pěti kroků vývoje popsaných v části 3. Vzhledem k již vícekrát zmíněné komplexitě a nekonzistenci dat ČSVS však není zcela dokončena jejich manuální kontrola. Přesto jsou data v aplikaci dostupná a využitelná pro uživatele, a to i vzhledem k tomu, že v aplikaci je možné v případě nejistoty tato data porovnat s naskenovanou verzí papírového slovníku. Předpokládáme, že ruční kontrola poběží ještě delší dobu a kvalita dat ČSVS se bude i za provozu postupně dále zlepšovat. Datová základna *Tezauru* je oproti tomu finální. Oba zdroje dat jsou průběžně ukládány také v úložišti LINDAT/CLARIAH-CZ, zajišťujícím jejich dlouhodobé uchování.

Softwarové zpracování je ve stádiu tzv. beta-verze. Je tedy funkčně kompletní, předpokládáme ale, že se podoba aplikace, a tedy i obslužného softwaru, může mírně změnit na základě poslední fáze projektu, testování.

4.1.3 PLÁNOVANÝ FINÁLNÍ STAV

Konečná podoba aplikace bude obsahovat i další sofistikované funkce. Aplikace například umožní v případě shody ve struktuře obou slovníků vyhledat a otevřít podobná hesla najednou. Uživatel si tedy bude moci porovnat obsah hesel obou slovníků, resp. doplnit si informace získané z jednoho slovníku pomocí slovníku druhého. Navíc bude mít kdykoliv možnost přepnout z prohledávání obou slovníků zároveň do práce pouze s jedním slovníkem, s tím, že druhý slovník ponechá otevřený. Bude si tak moci vyhledat v obou slovnících dvě různá, automaticky nepropojená hesla a ta následně porovnat.

Aplikace bude propojena s daty Českého národního korpusu; dovolí tedy uživateli — podobně jako při práci s *Internetovou jazykovou příručkou* (https://prirucka.ujc.cas.cz) — zobrazit frekvenční údaje k hledanému slovu či jeho kolokační profil. To pokládáme za důležité i vzhledem ke stáří a charakteru ČSVS.

Jedním z významných plánovaných cílů je také možnost průběžné aktualizace a doplňování slovníků v reakci na vývoj a inovaci českého lexika a na základě zpětné vazby od uživatelů. Součástí aplikace bude nejen editační režim pro správce, ale i možnost pro běžné čtenáře zasílat zpětnou vazbu ke konkrétním heslům. Zásadním úkolem bude rovněž zajistit optimální dostupnost aplikace a její propagaci. Jak bylo zmíněno výše, plánujeme umístit odkazy na stránky našich aplikačních garantů, očekáváme ale také (už jen kvůli hodnotě aplikace pro novináře a publicistiku) ohlas v médiích.

4.2 DATABÁZE URČENÁ PRO LINGVISTY A SPECIALISTY NA ZPRACOVÁNÍ JAZYKOVÝCH DAT

Předpokládáme, že samotná datová základna bude kromě běžných čtenářů, kteří k ní budou přistupovat přes standardní webové rozhraní, sloužit též specialistům zaměřeným na zpracování přirozeného jazyka a dalším aplikacím třetích stran. Pro tento typ

OPEN ACCESS

uživatelů bude určena jednak kompletní databáze ve formátu TEI-XML, zveřejněná v úložišti pro trvalé uchování dat LINDAT/CLARIAH-CZ, jednak programové rozhraní API, kterým bude možné čerpat data z online databáze přímo do aplikací třetích stran. Takový postup může být využit například při sémantickém značkování (databáze na vyžádání vrátí hierarchii či záhlaví hesla pro hledaný výraz), při rozšíření vyhledávání o slova sémanticky související (z databáze se dotaz uživatele doplní o všechna další slova ze shodného hesla) nebo pro bližší sémantickou specifikaci slova v textu (databáze identifikuje, které heslo obsahuje nejvíce slov z daného textu). O konkrétních způsobech využití budou nakonec rozhodovat sami výzkumníci a další uživatelé, podstatná je však plánovaná dostupnost dat jak díky úložišti, tak programovému API.

5. ZÁVĚR

Ve fázi testování, nejprve interního, našimi pracovníky a pracovníky aplikačních garantů, a následně veřejného, kdy bude aplikace otevřena veřejnosti s žádostí o zpětnou vazbu, očekáváme jak odladění všech chyb, tak i případné úpravy uživatelského rozhraní a API.

Naším cílem je, aby uživatelské rozhraní co nejlépe sloužilo všem výše zmíněným skupinám uživatelů a programové API umožnilo co nejsnazší využití v aplikacích třetích stran, například pro účely dalšího výzkumu. Je tedy pravděpodobné, že na základě zpětné vazby bude třeba provést některé dílčí úpravy, například v možnostech uzpůsobení uživatelského rozhraní, ve standardech zajišťujících kompatibilitu se zařízeními pro uživatele se speciálními potřebami či ve způsobech vytěžování dat programovým API.

Poděkování

Děkujeme oběma recenzentům za podnětné připomínky. Tento článek vznikl za podpory Technologické agentury České republiky v rámci programu ÉTA [TL02000041], dále byl podpořen projektem „Kreativita a adaptabilita jako předpoklad úspěchu Evropy v propojeném světě", reg. č.: CZ.02.1.01 /0.0/0.0/16_019/0000734, financovaným z Evropského fondu pro regionální rozvoj, a programem Cooperatio, vědní oblasti Lingvistika.

LITERATURA

Archambaultová, A., & Corbeil, J.-C. (1999). *Velký obrazový tematický slovník česko-slovensko-anglicko-německý*. Columbus.

Bečka, J. V. (1977/1982). *Slovník synonym a frazeologismů*. Novinář.

Brauerová, S. (2018). *Velký dětský obrázkový slovník*. LibroNet.

Carney, F., & Waite, M. (Eds.) (1986). *Pocket English Thesaurus*. Penguin.

Fischer, A. (2004). The notional structure of thesauruses. In C. Kay & J. Smith (Eds.), *Categorization in the History of English* (s. 41–58). John Benjamins. DOI

Galler, J., & Mrázek, J. (1947) *English in Pictures. Anglický obrazový slovník*. Nakl. J. Hrách.

Haller, J. (1969–86). *Český slovník věcný a synonymický 1–3, Rejstřík*. Státní pedagogické nakladatelství.

OPEN ACCESS

Hallig, R., & von Wartburg, W. (1952/1963). *Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas*. Akademie-Verlag.

Hartmann, R. R. K., & James, G. (1998). *Dictionary of Lexicography*. Routledge.

Havránek, B. et al. (1960–71). *Slovník spisovného jazyka českého*. ČSAV.

Internetová jazyková příručka (2008–2022). Ústav pro jazyk český AV ČR. Dostupné z https://prirucka.ujc.cas.cz/.

Kahn, J. E. (Ed.) (1989). *Reader's Digest Reverse Dictionary*. Reader's Digest Association Ltd.

Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. *Electronic Lexicography, eLex 2017*, Leiden, Netherlands. Dostupné z https://hal.archives-ouvertes.fr/hal-01508868v2.

Khemakhem, M., Herold, A., & Romary, L. (2018). Enhancing usability for automatically structuring digitised dictionaries. *GLOBALEX workshop at LREC 2018*, Miyazaki, Japan.

Klégr, A. (2007). *Tezaurus jazyka českého. Slovník českých slov a frází souznačných, blízkých a příbuzných*. Nakladatelství Lidové noviny.

Komenský, J. A. (1658/1896). *Orbis Pictus. Svět v obrazích*. Jaroslav Pospíšil.

Mašín, J., & Bečka, J. V. (1947). *Stručný slovník českých synonym*. Rudolf Mikuta.

Davies, H., & Bezděková, H. (1993). *Obrázkový anglicko-český slovník*. Jiří Fraus.

Obstová, Z. (2020). Zwei onomasiologische Wörterbücher als Basis für eine lexikalisch-semantische Datenbank des Tschechischen. In V. Kloudová, M. Šemelík, A. Racochová & T. Koptík (Eds.), *Spielräume der modernen linguistischen Forschung* (s. 92–111). Karolinum.

Pala, K., & Všianský, J. (1994/2008). *Slovník českých synonym*. Nakladatelství Lidové noviny.

Příruční slovník jazyka českého (1935–1957). ČSAV, SN.

Roget, P. M. (1852). *Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. Longman, Brown, Green, and Longmans.

Slovník českých synonym a antonym (2007/2012). Lingea.

Svensén, B. (1993). *Practical Lexicography. Principles and Methods of Dictionary-making*. Oxford University Press.

Tichý, O., Obstová, Z., & Klégr, A. (2021). A lexico-semantic database of Czech: An interim report. *Linguistica Pragensia*, *31*, 93–100. DOI

van Sterkenburg, P. (2003). Onomasiological specifications and a concise history of onomasiological dictionaries. In P. van Sterkenburg. (Ed.), *A Pratical Guide to Lexicography* (s. 127–153). John Benjamins. DOI

von Wartburg, W., & Keller, H.-E. (1922–1967). *Französisches Etymologisches Wörterbuch*. R. G. Zbinden.

Wiegand, H. E. (2004). Lexikographisch-historische Einführung. In F. Dornseiff, *Der deutsche Wortschatz nach Sachgruppen* (8. vyd., CD-ROM, bez paginace). De Gruyter. DOI

Zora Obstová | Ústav románských studií, Filozofická fakulta, Univerzita Karlova
<zora.obstova@ff.cuni.cz>

Ondřej Tichý | Ústav anglického jazyka a didaktiky, Filozofická fakulta, Univerzita Karlova
<ondrej.tichy@ff.cuni.cz>

Aleš Klégr | Filozofická fakulta, Univerzita Karlova
<ales.klegr@ff.cuni.cz>