VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÉ ZAŘAZOVÁNÍ NEZNÁMÝCH SLOV NA ZÁKLADĚ DERIVAČNÍCH VAZEB

AUTOMATIC CATEGORIZATION OF UNKNOWN WORDS BASED ON DERIVATIONAL RELATIONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **Bc. MARIE FALTUSOVÁ**
AUTHOR

VEDOUCÍ PRÁCE **doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2020

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2019/2020

Zadání diplomové práce

Studentka: **Faltusová Marie, Bc.**
Program: Informační technologie Obor: Bioinformatika a biocomputing
Název: **Automatické zařazování neznámých slov na základě derivačních vazeb**
Automatic Categorization of Unknown Words Based on Derivational Relations
Kategorie: Umělá inteligence
Zadání:

1. Seznamte se s dostupnými datovými zdroji a vědeckými výsledky v oblasti derivační morfologie češtiny.
2. Zpracujte slovníková data do podoby vzorů derivačních vztahů, identifikujte potenciální nové vazby mezi lemmaty morfologického slovníku a rozšiřte o ně současná data.
3. Navrhněte a implementujte systém pro automatické zařazování neznámých slov na základě existujících morfologických vazeb
4. Vyhodnoťte vytvořený systém na reálných datech a diskutujte možnosti zjednodušení manuálních kontrol.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle doporučení vedoucího

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/
Vedoucí práce: **Smrž Pavel, doc. RNDr., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2019
Datum odevzdání: 20. května 2020
Datum schválení: 1. listopadu 2019

Abstrakt

Tato diplomová práce se zabývá vytvořením systému pro automatické zařazování neznámých slov na základě derivačních vazeb. Pro tento účel byl systém navržen tak, aby z elektronických slovníkových dat získával derivační vazby a jejich rozborem z nich vytvářel slovotvorné modely. Na základě těchto znalostí je poté možné začleňovat nezařazená slova do stávajících hnízd, utvořených ze získaných vazeb, a jejich modelů, případně vytvářet nové. Čtenář bude postupně seznámen s důvody, které vedou k neustálé proměně či rozšiřování slovní zásoby, budou vysvětleny způsoby, jakými se odvozují slova v českém jazyce, a jak lze získat informace o změnách slov, vzniklých během derivačního procesu. Tento systém navazuje a rozšiřuje výzkum oblasti derivační morfologie v projektu morfologický analyzátor Výzkumné skupiny znalostních technologií, působící na Fakultě informačních technologií Vysokého učení technického v Brně.

Abstract

This master thesis deals with the construction of a system for automatic classification of unknown words based on derivation bonds. For this purpose, the system was designed to extract derivative links based on electronic dictionaries and to create word-forming models from them. Based on this knowledge, it is then possible to incorporate unclassified words into existing nests formed from the obtained bonds, and their models, or create new ones. The reader will be gradually acquainted with the reasons that lead to the continuous transformation or expansion of the lexicon, the ways in which the words in the Czech language are derived and how to obtain information about the changes caused by this derivation process. This system builds on and extends the research of the branch of morphology in the project of a morphological analyzer of the Research Group of Knowledge Technologies, working at the Faculty of Information Technology of the Brno University of Technology.

Klíčová slova

český jazyk, derivační morfologie, derivační vazby, derivace, lingvistika, slovotvorba, slovotvorný model, hnízdění, slovotvorná báze, formanty

Keywords

czech language, derivative morphology, derivative bonds, derivation, linguistics, word formation, word-formation model, nesting, word-formation base, formants

Citace

FALTUSOVÁ, Marie. *Automatické zařazování neznámých slov na základě derivačních vazeb*. Brno, 2020. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

Automatické zařazování neznámých slov na základě derivačních vazeb

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Marie Faltusová
3. června 2020

Poděkování

Na tomto místě bych chtěla poděkovat doc. RNDr. Pavlovi Smržovi, Ph.D. za vedení práce a odborné rady, potřebné pro vypracování tohoto projektu a za poskytnutí zdrojů dat, se kterými jsem mohla pracovat.

Obsah

Kapitola 1

Úvod

Dnešní doba se nese v duchu nezadržitelného technologického vývoje, kterým si člověk snaží zpříjemnit svůj život. Strojům je přenecháváno stále větší množství úkonů, dříve vykonávaných lidmi. Mezi bezesporné výhody zcela určitě patří jejich rychlost, přesnost a možnost provádět výpočty, které by člověk nedokázal. Právě tolik ceněná neomylnost nám ale připomíná, co počítači chybí – lidskost a emoce. Čím víc se obklopujeme počítači, tím víc jsme oddělení právě od této nenahraditelné stránky lidství. A tak se člověk snaží počítač alespoň trochu přiblížit sám sobě, aby se cítil vedle počítače přirozeněji. Vznikají stroje, které mají vypadat i chovat se jako člověk, inteligentní systémy, řešící složité situace z našich životů anebo po počítači vyžadujeme schopnost komunikace, jako by to byl člověkem.

Lidský způsob komunikace, pomocí jazyka, ale nelze omezit na přesná pravidla. Zdrojem jazyka je lidský mozek a neobsahuje pouze složku gramatiky, kterou se učíme jako děti po mnoho let. Obsahuje také potenciálně nekonečnou slovní zásobu, neustále se měnící a vyvíjející v čase. Narozdíl od počítače je člověk tvor společenský a působí na něj zároveň vnější vlivy prostředí, ve kterém žije. Úměrně rychlosti změn v jeho okolí, kterým je nucen se přizpůsobovat, vzniká potřeba být schopen podělit se o tyto dříve nepoznané skutečnosti s ostatními a vytváří se tak nová slovní zásoba. Způsoby, jak bude s novými slovy v jazyce naloženo, jsou různé, což platí zvláště u jazyků tak morfologicky bohatých, jako je čeština a budou rozebrány v druhé kapitole.

Tato práce je tvořena v rámci projektu morfologického analyzátoru Výzkumné skupiny znalostních technologií působící na Fakultě informačních technologií Vysokého učení technického v Brně. Zároveň také navazuje na mou bakalářskou práci Derivační morfologie češtiny na základě rozsáhlých korpusových dat. Seznámení s těmito daty, jejich přehled a současné zpracování bude uvedeno v kapitole třetí.

Hlavním cílem této práce je zkoumání změn, ke kterým dochází v českém jazyce během procesu derivace a odvození pravidel, na základě kterých by bylo možné odhadovat zařazení slov dosud pro systém neznámých. Druhotným záměrem je rozšířit funkcionalitu a data analyzátoru sloužící pro výzkum jazyka. Čtvrtá a pátá kapitola se věnují návrhu a implementaci systému, který tento záměr realizuje.

V šesté kapitole je následně celý systém vyhodnocen pomocí reálných dat, jsou diskutovány různé možnosti přístupu k datům, rozšíření systému a jejich případné výhody, nevýhody či dopady. Na závěr jsou rozebrána úskalí testování systému pracujícího s lingvistickými daty, nutnost manuální kontroly a možnosti jejího usnadnění.

Kapitola 2

Člověk, jazyk a slovní zásoba

Jazyk je jedním ze základních prostředků umožňující lidskou interakci. Je to systém, který se skládá z jednotek, pravidel, konvenčních norem a modelů, aby mohl sloužit svému účelu, kterým je primárně přenos informací. Tento proces nazýváme **komunikací**, a ta může mít mnoho forem a účastníků – od osobní komunikace mezi dvěma lidmi, až po komunikaci masovou, mediální, určenou pro miliony posluchačů. Základní typy komunikace jsou však pouze dva, a to orální (mluvená) nebo grafická (písemná).

Aby komunikace mohla plnit svůj cíl, musí být zprostředkována pomocí jazyka strukturovaným způsobem, srozumitelným a platným tak, aby bylo možné porozumění mezi všemi zúčastněnými stranami. Tento aspekt komunikace, při kterém dochází k převodu přijatých signálů standardizovaným způsobem, se nazývá kód. A neboť do komunikace pomocí přirozeného jazyka a dekódování sdělované informace spadá i zpracování gest, mimiky a jiných neverbálních signálů, stává se z lingvistiky poměrně obsáhlá věda [20].

Tato práce si však klade za cíl pracovat pouze s grafickou podobou jazyka v elektronické formě a zkoumat jeho proměnlivou povahu v souvislosti se slovní zásobou a mechanizmy derivační morfologie.

2.1 Přirozený jazyk pod vlivem společnosti

Přirozený jazyk je silně vázán na člověka, kterým je ovlivňován. Lidský vývoj ve všech svých aspektech způsobuje neustálou proměnu jazyka tím, že vytváří potřebu popisovat stále další a další nové skutečnosti nebo objekty. Mluvnická stavba jazyka, jeho pravidla, se nemění téměř vůbec nebo jen velmi pomalu, oproti tomu slovní zásoba se obměňuje velice rychle. Tím je vytvářena potřeba na tento vývoj reagovat a začleňovat nová slova do zavedených pravidel.

Přirozený jazyk je složitým systémem, komplikovaným z hlediska chápání, popisu i třídění. Na přirozený jazyk a vazby, kterými je ovlivňován, lze nahlížet mnoha způsoby, interně anebo externě [20].

Externí pohled

Externí pohled propojuje jazyk s vnějšími vlivy, které na něj působí, zejména emoce, smýšlení, společnost či etnikum, což je souhrnně nazýváno jako makrolingvistika. Jazyk je ve společnosti centrální – bez společnosti by nikdy neexistoval, a zároveň společnost nemůže existovat bez něj. Pro jednotlivce je nástrojem socializace, pomocí kterého je předáváno vědění

i zákony. Toto zcela zásadní propojení z hlediska lingvistiky zkoumá pomezní jazyková disciplína **sociolingvistika**.

Je možné uvést 3 body propojení jazyka a společnosti [20]:

1. Jazyk je společenská instituce, která se nezakládá na přirozeném vztahu vůči svému objektu, tzn. není nomenklaturou.

2. Jazykové znaky jsou vytvářeny lidským mozkem, mají psychickou podobu. Udržovány jsou ale společenským konsenzem, vzájemným potvrzováním a většinovou shodou, což vylučuje jakékoliv autoritativní vlivy jedince.

3. Jazyk je konvenční, formovaný zvyky a vývojem. Jako takový je tedy historickým jevem.

Jazyk český se, stejně jako všechny světové jazyky, historicky vyvíjel a byl ovlivňován různými milníky v dějinách národa. Jeho vznik, a tedy formování prvních pravidel jazyka, který dnes nazýváme čeština se datuje do období Velké Moravy. Důležitým milníkem byl příchod věrozvěstů Cyrila a Metoděje a zavedení staroslověnštiny, avšak největší vliv mělo období mezi 12. a 16. stoletím, kdy jazyk procházel vývojem z hlediska lexikálního a fonetického. Hlavním impulsem pro tento rozvoj jazyka byl fakt, že začal být hojně používán, čímž došlo i k významnému rozšíření slovní zásoby. Tímto byl zformován samostatný, běžně rozšířený jazyk s vlastní, rozvinutou mluvnickou stavbou.

I přes odsunutí jazyka do pozadí v období po bitvě na Bílé Hoře, byl český jazyk v 19. století opět obnoven a zmodernizován zjednodušením složitých vazeb a odstraněním zastaralých výrazů. Slovní zásoba češtiny byla v uplynulých staletích několikrát obohacena jinými jazyky na základě historických dějů, spjatých s územím českých zemí, kde byl tento jazyk používán. Došlo k začlenění slov zejména z latiny, románštiny a němčiny [7].

V novodobých dějinách češtinu ovlivňují zcela odlišné faktory. V době komunizmu podléhala slovní zásoba zejména politickým vlivům. Následně došlo k pádu režimu v roce 1989, při převratu nazývaném jako sametová revoluce. Tento politický děj měl na jazyk zásadní vliv, neboť bylo potřeba vybudovat nový právní systém, definovat nové pojmy a začlenit se tak do evropského kontextu. Změny proběhly i ve společenském životě, kde došlo k deideologizaci a projevily se zejména ve významové rovině lexikonu. Mnohá slova nabyla hanlivého významu nebo jej úplně pozbyla a staly se z nich archaizmy až historizmy (např. slovo *proletář*). Oproti tomu byl započat nový trend, kterým bylo vytváření nových slov a pronikání cizích výrazů do běžné mluvy. Ať se jedná o slova původem cizí nebo česká s novým významem, nazýváme je **neologizmy**.

Slova přejatá jsou výrazy přejímané z cizích jazyků. Zde je nejvýznamnější vliv angličtiny, obecně lze ale mluvit o přejímání z jakéhokoliv jazyka vyspělejších západních států, s nimiž byly nově navázány vztahy. Proto se tento typ slovní zásoby týká především oblastí jako je veřejný život, finančnictví, ekonomie, ekologie a politika. Jako příklad mohou složit slova *briefing* (z angličtiny), *sumit* (z angličtiny), *infrastruktura* (z latiny), *leasingový pronájem* (z angličtiny) atd.

Slovní zásoba byla také obohacena o **odvozeniny a složeniny**, kam lze zařadit výrazy jako *azylant, migrant, restrukturalizace* nebo *rekvalifikace*. A mimo jiné i **české jednotky**, kterými jsou buď slova zcela nová jako *bezrozporně, čerpadlář, nezcizitelný, obchodovatelnost, přezaměstnat* či *přerozdělit*, nebo došlo k obnovení některých stále platných výrazů z doby první republiky (*starosta, bezdomovec, přecenit*).

Je důležité zde zmínit, že došlo k vymezení pojmu **současný jazyk** nejen z hlediska lexikografického. Do této doby byla za současný jazyk považována ta verze češtiny, která

vycházela z dob národního obrození, respektive první republiky. Vysvětlením může být poznatek z pozorování historických vlivů na jazyk, kde je znát, že největších obměn a aktualizací jazyk dozná, když začne být (znovu) aktivně používán v literatuře. Tyto zlomy již v historii češtiny přišly několikrát – v 9. století s náboženskou literaturou, ve 14. století, což mělo za následek osamostatnění jazyka, a poté v době národního obrození na konci 19. století, kdy došlo k obnovení češtiny. Ve 20. století na tuto verzi jazyka navázala čeština používaná v období první republiky a vývojem publicistiky a beletrie došlo k lehké modernizaci jazyka, ale zbytek století se nesl ve znamení válek a následného komunistického režimu, čímž došlo ke zmrazení společenského života a tím i vývoji literatury. Proto dalším obnovením český jazyk prošel až po roce 1989 [6].

21. století se nese v duchu technologického pokroku a globalizace. Do jazyka se dostávají nové výrazy vycházející z potřeby pojmenovat dříve neexistující objekty a služby. Zároveň pronikání technologií mezi širokou veřejnost ovlivňuje výrazy používané v běžném jazyce a zapříčiňuje vznikání nových. Dochází k přejímání cizích slov a jejich případnému ohýbání podle pravidel českého jazyka. Dynamika dnešní doby se odráží i v tempu, jakým se mění slovní zásoba. Zejména hovorové výrazy podléhají trendům společnosti a mohou vznikat i zanikat neobvykle rychlým způsobem. Dochází také k posunu významů některých slov a vzniku nových slovních spojení s posunutým sémantickým výkladem. Tento trend doby je výzvou pro lingvistiku z hlediska stanovení hranice mezi spisovnými a hovorovými výrazy, určení významů těchto slov a formalizace jejich gramatických (zejména morfologických) vlastností.

Interní pohled

Mezi interní pohledy na přirozený jazyk patří vertikální pohled, který uplatňuje dělení jazyka tzv. „shora dolů“ od kolokací (spojení slov), přes lexikon (jednotlivá slova), morfologii flektivní a derivační, dělící slovo na jednotlivé morfémy, až po fonémy – elementární jednotky jazyka. Tyto procesy budou blíže popsány v kapitole 2.2.

Další typ interního pohledu na jazyk je horizontální, který zkoumá jazyk z hlediska variací jazyka. Ty mohou být geografického (dialekty a interdialekty) nebo funkčního původu. Jazyk je takto rozdělen na standardní, spisovnou verzi, která plní reprezentativní a formální funkci, mluvený jazyk, který odráží společnost a její běžné potřeby vyjadřování, profesní jazyk, anebo jazyky uzavřených skupin, hantýrku či argot.

U jazyka lze také pozorovat jeho „sbíhavost“ a souvztažnost jednotlivých jevů, což je koncentrický pohled na jazyk. Tyto vlastnosti se projevují v existenci tzv. jádra jazyka a okrajových neboli řídkých slov. Vzniká zde určitý vztah mezi národním jazykem (obecný jazyk celého národa) a idiolektem, což je jazyk každého jednotlivce. Metrikou tohoto jevu můžeme označit frekvenci jednotlivých slov, tedy jejich četnost výskytu, udávanou v číslech nebo procentuálním poměru. Tento jev je zmapován pomocí frekvenčních slovníků a je předmětem studia kvantitativní lingvistiky [20].

2.2 Slovotvorba z hlediska vnitřní lingvistiky

Vnitřní lingvistika se, na rozdíl od vnější lingvistiky a externího pohledu na přirozený jazyk, zabývá hlediskem obecných principů. Patří sem jazykovědné disciplíny jako gramatika, fonetika, syntax a také morfologie. Ta se dále dělí na flektivní morfologii (skloňování a časování) a slovotvorbu, nazývanou též jako derivační morfologii.

Slovotvorba se zabývá procesem tvoření jednotlivých nových slov (úzce spjato s neologií a neologizmy v jazyce). Ten je charakteristický podle použitého typu a počtu jednotlivých komponent, jejich pozicí, pořadím, druhem vzájemného vztahu a základu. Podle těchto rysů můžeme rozlišovat **slovotvorné modely**.

Nejčastějším slovotvorným procesem v češtině je derivace, zejména derivace substantiv, která se na těchto procesech zakládá, a kompozice. Nejedná se ale o jediné způsoby vytváření nových slov, strukturovaný přehled základních slovotvorných postupů má tuto podobu [2]:

1. derivace (slovotvorba odvozováním)
   (a) prefixace
   Slovotvorné prefixy (předpony) stojí před kořenem slova. Často slouží k odvozování sloves a k jejich perfektivizaci. Mezi slovotvorné předpony se řadí i předpona *ne-*.
   Např.: *výhoda – ne-výhoda, psát – na-psat.*
   (b) sufixace
   Sufixace je proces připojování slovotvorných sufixů (přípon) za kořen slova, u jmen se k nim připojuje také tvarotvorný morfém – deklinační koncovka a je možné připojovat i speciální případ sufixů, tzv. postfixy, které se přidávají až za deklinační koncovku.
   Např.: *plavat – plav-ec, kámen – kamen-ný, jaký-si.*
   (c) transflexe
   Transflexe spočívá v tvoření koncovkou nebo kmenotvorným sufixem. U sloves tímto procesem může docházet k imperfektivizaci.
   Např.: *malina – malin-í, fax – fax-ova-t, kousnout – kous-a-t.*
   (d) kombinované odvozování
   Jedná se o slovotvorbu pomocí více afixů, tedy připojováním prefixů a sufixů zároveň.
   Např.: *koleno – pod-kolen-k-a, moře – zá-moř-í.*
2. kompozice (slovotvorba skládáním slov)
   (a) tvoření nevlastních složenin
   Vzniká pouhým spojením dvojice slov do jednoho. Tímto způsobem může dojít i k derivaci prefixací, a to v případě, kdy se spojí jméno s předložkou a vnikají příslovečné spřežky.
   Např.: *zisku chtivý – ziskuchtivý, potichu, nalevo.*
   (b) tvoření vlastních složenin
   V tomto případě se členy složenin formálně liší od základových (fundujících) slov a bývají spojovány spojovacím morfémem.
   Např.: *sladký a kyselý – sladk-o-kyselý, velký a trh – vel-e-trh.*
   (c) kompozičně-derivační tvoření
   Dochází zde ke spojování slov (kompozici) a zároveň dalšímu slovotvornému procesu z kategorie derivace.
   Např.: *psát a romány – roman–o-pis-ec, měřit a čas – čas-o-mír-a.*

3. konverze (přechod slova k jinému slovnímu druhu bez použití afixace)

   K tomuto okrajovému způsobu tvoření slov v češtině dochází ve třech případech, kterými jsou změna slovního druhu neohebného slova, ustrnutí deklinačního paradigmatu (přechod ohebného slova k neohebnému) anebo změna deklinačního paradigmatu, která souvisí s univerbizací (popsáno níže).

   Např.: *ráno – podst. jméno i příslovce, vepřový – vepřové, vrchní číšník – vrchní.*

4. reflexivizace (vznik nových sloves přidáním zvratného zájmena)

   Přidáním zvratných zájmen se z nich stávají formanty v případě, kdy dojde pouze k přidání tohoto zájmena nebo spoluformanty v situacích, kdy zároveň dochází i k jiným (často derivačním) procesům.

   Např.: *povídat – povídat si, moudrý – umoudřit se.*

5. univerbizace (slovotvorba spojením více slov do jednoho)

   Jedná se o spojování ustálených, víceslovných spojení do jednoho slova. Využívá zejména sufixaci a vzniklá fundovaná slova často patří mezi hovorové výrazy.

   Např.: *základní škola – základka, řidičský průkaz – řidičák.*

6. abreviace (vznik zkratek z víceslovných pojmenování)

   Zkracování víceslovných pojmenování dochází k abreviaci. Tímto procesem se utváří zkratky či zkratková slova. Mohou být písmenné (použití počátečních písmen jména) anebo hláskové (použití počátečních hlásek jména). U některých zkratek dochází ke skloňování, jako by byly běžnými slovy.

   Např.: *Organizace spojených národů – OSN, Severočeské tukové závody – Setuza, České energetické závody – ČEZ – v Čezu.*

Obecně můžeme mluvit o **bázi a formantu**, kde báze je základ slova, neměnná část, použitá k dalšímu tvoření a nesoucí význam základového slova. Slovotvorná báze či základ je část slova společná pro slovo základové a slovo odvozené [2]. Zpravidla bývá bází kořen (v případě slovotvorby jde o slovotvorný kořen), který je společný pro slovo základové, fundující a zároveň pro všechna jeho slova fundovaná (derivace). Se slovotvorbou taktéž souvisí termín kmen, což je část slova, ke které se připojují tvaroslovné koncovky pádové, osobní, tvarové či infinitivní. Ve slovotvorbě se setkáváme s kmenem jednoduchým, který je totožný s kořenem (*žen-a*, *zdráv-a*) anebo odvozeným, jenž vznikl pomocí slovotvorných, odvozovacích sufixů (*uč-i-tel-k-(a)*, *měst-sk-(ý)*) [11]. Pokud během slovotvorného procesu došlo k prefixaci, reflexivizaci nebo kompozici může být slovotvorný základ celé slovo fundující, jinak bývá slovotvorným základem kořen či kmen slova [18].

Formant je část, ve které dochází při slovotvorném procesu ke změně. Slovotvorný formant je ta část slova, kterou se slovo odvozené od slova základového liší [2]. Souhrnně se tyto proměnlivé komponenty nazývají slovotvorné morfémy [20]. Morfém je nejmenší nedělitelná jednotka, nesoucí význam a podle své funkce se dělí na slovotvorné (podílející se na slovotvorbě), tvarotvorné (vyjadřující gramatické významy), kmenotvorné a interfixy (morfém konstrukčního, spojovacího typu nacházející se často mezi dvěma kořeny složených slov) [2]. Nekořenové morfémy se nazývají afixy a dále se dělí podle pozice vůči kořeni na prefixy (předpony), infixy (vpony), sufixy (přípony) a postfixy, což jsou derivační morfémy nacházející se na konci slova za gramatickými sufixy [1]. Prefixace a sufixace se souhrnně nazývají

afixací. Možný je ovšem i postup regresivní, kdy dochází k deprefixací či desufixací – tedy odtržením afixu (deafixaci) [20].

Součástí slovotvorné báze jsou také hláskové změny nebo také alternace, které jsou během tvorby slov v českém jazyce běžné. Vznikají na morfémovém švu (místo styku báze a formantu) v závislosti na sousedních fonémech a mění fonologickou skladbu formantu. Vždy dochází ke změně hlásek či hláskových skupin na alternace fonologicky blízké. Hlásky tak mohou v tomto procesu přibývat, měnit se a mizet [2]. Přejímání cizích slov se neřadí mezi slovotvorbu, ale slovo přejaté může být jejím předmětem, jak ilustruje dvojice slov *laser* a *laserový* [20].

Těmito postupy mezi slovy vzniká **slovotvorný vztah**, kdy jedno je základové (fundující) a z něj odvozené je utvořené (fundované). Je definovaný formální složkou (fundace, jedno slovo se zakládá na druhém) a významovou složkou (motivace, vzájemně odkazují na svůj význam). Na základě motivace určujeme i obecný slovotvorný význam, který nemusí být totožný s významem lexikálním (např. slovo *zelenina* neodkazuje na zelenou látku). Slova zakládající se na společném fundujícím slově tvoří slovotvorné řady a svazky. Skupina slov svázaných vzájemnými fundačními vztahy se stejným kořenem se nazývá slovotvorné hnízdo nebo také slovotvorná čeleď [2].

Aby bylo možné zkoumat tyto slovotvorné (derivační) vazby mezi slovy, je třeba provést **slovotvorný rozbor**. Ten spočívá v porovnání slova fundujícího a fundovaného, čímž získáme bázi (neměnná část, společná pro obě slova) a formant (proměnlivá část). Podle počtu základových slov a charakteru slovotvorného formantu můžeme určovat slovotvorný způsob a slovotvorný postup popisující použité typy slovotvorné úpravy. Můžeme snadno určit afixy i hláskové alternace, ke kterým případně došlo. Takto je možné rekurzivně pokračovat dále, a pokud bylo základové, fundující slovo samo utvořené, opakováním rozboru můžeme dospět až ke slovu neutvořenému (např. *odbarvovač – odbarvovat – odbarvit – barvit – barva*) [2].

2.3 Pohled neurolingvistiky na osvojování jazyka

Neurolingvistika je interdisciplinární obor kombinující neurologii a lingvistiku. Kromě těchto disciplín často zahrnuje také poznatky z psychiatrie, psychologie, fyziologie, foniatrie nebo logopedie. Své nejrozšířenější využití nachází zejména ve zkoumání poruch řeči, nicméně tento medicínsky zaměřený výzkum slouží lingvistům mimo jiné i k lokalizací jazykového systému v mozku. Výzkum v této oblasti je přínosný pro zkoumání samostatnosti jazykových aktivit jako je čtení, produkce a percepce řeči či psaní. V neposlední řadě je díky neurolingvistice možné přiblížit se pochopení principů osvojování jazyka člověkem [9].

Od narození probíhá v lidském mozku mnoho vývojových procesů a jedním z nich je učení se mateřskému jazyku. Tento úkol se skládá z mnoha částí odrážející komplexnost jazyka jako takového, neboť zahrnuje učení se fonémům, pravidlům syntaxe, propojení gramatiky se sémantikou a v neposlední řadě rozšiřování slovní zásoby.

Žádný stroj není schopen ovládnout během pár let mluvenou a psanou formu jazyka takovým způsobem, jako malé dítě, které se učí jazyk (obvykle mateřský jazyk) snáze než dospělí, studující další, cizí jazyky. Osvojování jazyka je v této životní fázi ulehčeno konkrétní neurální architekturou, která umožňuje aplikování kombinace mechanizmů vyvinutých evolucí k rychlému a efektivnímu zvládnutí komunikace. Taková architektura je definována strukturální propojitelností mezi konkrétními oblastmi mozku, jejichž vlastnosti jsou vhodné pro kódování přesně daných reprezentací okolního prostředí [4].

Popsaná organizace struktur mozku vysvětluje naměřenou citlivost malých dětí na organizaci vět, zejména pokud je podpořena prozodickými faktory řeči (např. tempo, tón hlasu, pauzy mezi slovy a větami, rytmus či hlasitost). To je přirozeným činitelem ve vnímání a dělení řeči na regiony či menší segmenty, což pomáhá dětem vstřebávat řečový tok. Získávání koherentních bloků a jejich analýza jsou považovány za hlavní mechanizmy pro extrakci slov z řečového proudu, následného pochopení jejich sémantického i syntaktického významu a možností tvoření dalších slov nebo spojení [4].

Na fonologické úrovni dítě začíná s menším repertoárem převážně základních zvuků. Některé fonémy jsou zkomoleny a splývají s jinými, což se projevuje nepřesnou výslovností některých slov. Až přibližně od 3 roku věku začíná být dítě schopno některé fonémy (zvuky) rozlišovat mezi sebou, načež následuje učení se jejich výslovnosti a pochopení významu ve slovech.

V případě syntaxe dochází k rozpoznávání jednodušších jednotek, které se následně kombinují do složitějších, větších struktur. Všechny tyto popsané procesy se vyznačují aditivním charakterem, tedy přechodem od jednoduššího, obecnějšího ke složitějšímu a přesnějšímu za předpokladu, že jsou vstřebány nové informace vedoucí k tomuto rozšíření znalostí. V pozdějším procesu vývoje ale dochází i k procesům subtrakce a reorganizace. S vývojem dítěte a množstvím zkušeností, které vstřebává, jsou irelevantní informace likvidovány, a ty potřebné jsou kódovány ekonomičtěji. Neplatí tedy zjednodušený předpoklad, že spolu korespondují neurony, synapse a mentální reprezentace, které jsou v nich kódovány.

Z hlediska lexikonu dochází nejen k rozšiřování, ale zejména k rozlišování mezi jemnými nuancemi významu. Jedno slovo může být v případě dítěte zpočátku označením pro velké množství objektů a postupně, jak se vyvíjí, začíná rozlišovat mezi jednotlivými koncepty a dochází k rozlišování mezi přesnými pojmy. Jako příklad lze uvést slovo *pták*, které ze začátku může označovat jakéhokoliv opeřence nebo objekt na obloze a až postupem času se dítě učí rozpoznávat jednotlivé druhy ptáků, od prostého dělení na kategorie (velký pták, malý pták) až po odborné názvosloví a jednotlivá jména.

Růst slovní zásoby u dětí není lineární. V počátcích dochází k vstřebávání nových slov jen velmi pomalu, až později dojde k výraznému skoku a utvoření většího slovníku. Tento jev je vysvětlován způsobem, jakým jsou kódována slova v jazyce. Jak již bylo naznačeno, dítě se začíná učit nejdříve napodobováním, což je prováděno aditivním způsobem a s velmi špatnou analýzou vnitřní struktury slova. Většinou dochází k rozpoznání, že slovo začíná na stejný foném nebo zvuk, ale již si neuvědomuje, která hláska je uprostřed slova. Děti tak rozpoznají např. podobnost mezi rýmy, ale nedokážou říct další slovo, které obsahuje uprostřed stejný foném. Tyto problémy mívají mnohé děti až do věku kolem 4 let.

P. W. Jusczyk[1] nastínil, že tato vlastnost se zdá být určující pro způsob práce s mentálním lexikonem. Jakmile jeho velikost přesáhne určitou hranici, ukazuje se, že mozek začne aplikovat optimalizaci uspořádání a vyhledávání, aby bylo možné najít slovo, kterým se chceme vyjádřit. Organizace mentálního lexikonu má fonologické uspořádání, což se projevuje uložením slov, začínajících na stejný foném, resp. zvuk, společně. Dokud dítě vnímá pouze základní fonetické rysy slov, má i jednodušší způsob uložení svého mentálního lexikonu, ale jakmile rozeznává jednotlivé fonologické jednotky, přístup k ukládání slov se mění. Postupuje tedy od počátečního stavu, kdy vstřebává spíše větší části slov (slabiky) až po znalost jednotlivých fonémů, pochopení strategie kódování a skládání slov z menších stavebních bloků (fonémy). Předpokládá se, že toto plné fonologické vědomí dítě získává až ve chvíli, kdy se začne učit číst, neboť abeceda dává slovům strukturu jednotlivých slabik a písmen.

[1] Autor tuto myšlenku uvedl ve své práci z roku 1997, nazvané *Infant speech perception and the development of the mental lexicon.*

Taková strategie usnadňuje učení nových slov, protože neznámé slovo je možné rozložit na menší, známé části, které si lze snáze zapamatovat.

Učení a vývoj mozku bývá často vnímán jako čistě biologický proces a je zanedbáván vliv okolního prostředí, v němž k tomuto dospívání dochází. Biologie ruku v ruce se zkušenostmi, se kterými se člověk během vývoje setkává, ovlivňují způsob, jakým se učí jazyk. Názorným příkladem je výše popsané učení se rozpoznávání fonémů, které není řízeno biologickým harmonogramem, jako spíš zkušeností s abecedním systémem. Je prokázáno, že negramotní lidé nemusí této úrovně rozlišování fonémů dosáhnout nikdy [3].

Výzkumy ukazují, že lidský mozek provádí dekompozici vnímaných slov na menší části, a to nejen z důvodu efektivního uložení v mozku, ale zároveň se takový přístup jeví i jako výhodná strategie při odhadování významu a zařazení neznámých slov. Lidský mozek je uzpůsoben ke vstřebávání řeči díky své unikátní architektuře, postupnému učení, vytváření nových pravidel z dostupných zkušeností a ukládání aktivní slovní zásoby.

2.4 Vědecké projekty v oblasti derivační morfologie českého jazyka

Čeština patří mezi morfologicky košaté jazyky, obsahující nespočetné množství formantů flektivního i derivačního charakteru. Právě propojování flektivní a derivační morfologie se díky projektům a studiím začíná ukazovat jako přínosné z lingvistického hlediska i z pohledu automatického zpracování češtiny. Toto spojení nicméně není v teoretickém popisu češtiny zakořeněno, a navíc se projevuje specificky, v závislosti na morfologické složitosti každého jazyka [22].

Nejkomplexnější popis derivačních procesů v českém jazyce vytvořil Miloš Dokulil. Ten definoval čtyři onomaziologické kategorie, a to kategorii substanční, odpovídající podstatným jménům, kategorii vlastností, odpovídající přídavným jménům, kategorii příznaků konstantních, která koresponduje s příslovci a kategorii příznaků dynamických odpovídající slovesům. Mezi těmito kategoriemi definoval tři typy posunů [22]:

1. Transpozici, která je chápána jako změna kategorie bez změny ve významu. Například *líný – lenost*, *pečlivý – pečlivost*.

2. Modifikaci, která se vyznačuje přidáním příznaku v rámci jedné kategorie. Jako příklad lze uvést *židle – židlička*.

3. Mutaci, jež reprezentuje změnu významu při přechodu v rámci jedné nebo do jiné kategorie. Například u dvojicí *bloudit – bludiště*, *galerie – galerista*.

Tyto kategorie jsou dále děleny na slovotvorné skupiny, které zahrnují slova odvozená od stejného slovního druhu se stejným slovotvorným významem, která se dále dělí na slovotvorné typy, jako např. slovotvorný typ deadjektivních substantiv s významem kvality končících na *-ost* nebo slovotvorný typ s příponou *-ství* atd. Dokulilovo dělení se stalo respektovaným a zároveň jediným přístupem pro popis derivací v českém jazyce.

Popis odvozovacích formantů bývá organizován podle výše popsaného dělení a vztah mezi derivací a flektivní morfologií bývá uváděn pouze vzácně. Ve slovotvorbě[2] bývá spatřována přechodová oblast mezi morfologií a slovníkem. Slovotvorné či lexikální prostředky jsou považovány za nemorfologické. Propojením těchto dvou lingvistických oblastí se zabývá

[2]Uvedeno v Dokulil a kol. (1986): *Mluvnice češtiny (1)* a Komárek a kol. (1986): *Mluvnice češtiny (2)*.

až B. Bednaříková[3] a K. Osolsobě[4]. V oblasti automatického zpracování češtiny již bývá toto propojení běžné [22].

Jedním z projektů zkoumajících derivační morfologii v českém jazyce je webový nástroj **Derivancze** [12], který umožňuje zadávání slov přímo do webového formuláře. Pro požadovaný výraz vypíše slovo základové a seznam jeho přímých derivací. Nástroj obsahuje přes 255 tisíc derivačních vztahů, které jsou sémanticky anotovány sadou 17 značek.

Data v Derivancze jsou reprezentována jako seznam dvojic *hledané slovo:výsledek* a celý nástroj je implementován jako minimální konečný stavový automat. Není v něm prováděna reálná analýza slov, pouze vyhledává všechny možné odpovědi pro zadané slovo. K tomu využívá předpřipravených dat a analýz z kompilační fáze programu, což umožňuje tomuto nástroji rychle interagovat s uživatelem a vstupem, který zadává.

Vazby jsou označované jako D-relace a reflektují sémantické souvislosti mezi slovy. Nástroj tedy nepracuje na základě určování slov základních, které mohou mít lehce odlišný význam, ale slova seskupuje podle jejich společného významu. To nástroji pomáhá vytvářet vztahy mezi slovy s ortografickými variantami (jako *socializmus* a *socialismus*) a rozlišovat synonyma (např. *normalizace*, nikoliv však vytvořit vztah ke slovu *normálnost*).

Autoři vycházejí původně z přístupu Miloše Dokulila a jeho dělení do kategorií. Dodržet ale přesné formální rozvržení dat tak, aby korespondovaly s kategoriemi, se ukázalo být příliš komplikované. Zejména proto, že teoretické zpracování vychází ze standardních, typizovaných příkladů a reálná slova aktivně používaného českého jazyka, získaná z korpusů odráží větší komplexnost jazyka [12].

Snaha o uchopení češtiny více formálnějším způsobem se projevuje v dalším rozsáhlém projektu nazvaném **DeriNet** [22]. Tato lexikální síť je navržena jako elektronická databáze českých plnovýznamových slov. Mezi tuto skupinu se řadí podstatná jména, přídavná jména, příslovce a slovesa. Ta jsou propojena orientovanými vazbami, které reprezentují slovotvorný vztah od slova základového, fundujícího směrem ke slovu odvozenému, fundovanému. Jednotlivá lemmata reprezentují uzly, a tak vzniká orientovaný stromový graf.

Konstrukce celé sítě vycházela z vybrané sady lemmat pocházejících z rozsáhlých korpusů češtiny, čímž autoři vybrali pouze lemmata aktuálně používaného jazyka. Cílem bylo identifikovat co největší množství derivačních vztahů za současného zachování konzistentního, ekonomického a lingvisticky adekvátního řešení. K identifikaci derivačních vazeb byly použity dostupné zdroje jazykových dat (např. slovník MorfFlex CZ, který je morfologickým seznamem označkovaných lemmat). Důraz byl kladen na lingvistickou správnost zachycovaných vztahů, a proto nové vazby vždy procházely manuální kontrolou. Verze DeriNet 1.1 obsahuje 969 tisíc lemmat a 718 tisíc hran[5].

Výzvou pro lingvisticky korektní reprezentace derivačních vztahů je značná nepravidelnost slovní zásoby českého jazyka. Zvláště komplikované jsou případy homonymie a polysémie. V síti jsou polysémní lemmata reprezentována jedním uzlem a všechny derivace jsou seskupeny pod tímto uzlem. Homonyma jsou naopak derivována z různých základových slov a mívají svou vlastní, odlišnou, sadu derivací a v síti mají být reprezentována jako samostatné uzly. Způsob rozlišování homonymie od polysémie je, avšak ve slovníku MorfFlex CZ příliš široký a vznikají nejasné hranice, rozlišující mezi homonymií a polysémií. Pro zajištění kompatibility mezi slovníkem a sítí DeriNet byla provedena nejprve automatická a poté manuální revize lemmat slovníku, která měla za cíl vymezit jádro homonym, lišících se derivačními i flektivními vlastnostmi.

[3] Zmíněno v publikaci: Bednaříková, B. (2009): *Slovo a jeho konverze.*

[4] Problematikou se zabývá v publikaci: Osolsobě, K. (2014): *Česká morfologie a korpusy.*

[5] Údaje jsou platné pro rok 2016.

Podobně lingvisticky nejednoznačný je také přístup k derivacím s negačním prefixem (a případnými dalšími prefixy, se kterými je negační předpona kombinována). Pokud by byl zachycen vztah jak k slovu základovému, které taktéž obsahuje negační prefix, tak i ke slovu základovému bez tohoto prefixu, vznikl by v síti cyklus, který není přípustný. Příkladem může být současný derivační vztah mezi slovy *mačkavost – nemačkavost – nemačkavý*. Obdobně fungují i některé další prefixy, jako derivační řetězce *biologie – biologický – biologicky* a *mikrobiologie – mikrobiologický – mikrobiologicky*, kde druhý řetězec může být považován za deriváty prvního. V tomto případě bylo nutné v síti jasně určit jednotný přístup k těmto typům vazeb [22].

Kapitola 3

Zdrojová data pro derivační morfologii

V rámci projektu morfologického analyzátoru[1] je k dispozici celkem 10 slovníků českého jazyka v elektronické podobě. Liší se jak tématem, tak uspořádáním a typem informací, které poskytují. Jejich specifika jsou popsána v této kapitole. V tabulce 3.1 jsou uvedeny všechny dostupné slovníky s jejich technickými parametry, jako je velikost zdrojového souboru slovníku, která přímo závisí jak na typu zdrojového souboru, tak na obsažnosti informací uváděných autory. Informace o počtu lemmat byla získána ze souborů typu `def`, dostupných u každého slovníku. Tento typ souboru obsahuje lemmata a jejich definice, získané zpracováním zdrojového kódu slovníku. Skutečný počet lemmat slovníku může být i vyšší, neboť lze dále získávat dalším zpracováním výrazy uvedené v definicích slovníku. Posledním údajem je, zdali slovník poskytuje informace o derivačních vazbách, které by bylo možné získat přímo, bez dalšího zpracování odvozováním.

Český etymologický slovník (cety)

Český etymologický slovník, vydaný roku 2008 v elektronické podobě nakladatelstvím LEDA, s. r. o., se zabývá původem uváděných lemmat, jejich odvozením a významem. Pro derivační morfologii je zajímavý zejména seznam souvisejících výrazů uváděný u většiny lemmat, ze kterého lze získat údaje nejen o derivacích, ale také o předponách, používaných pro dané lemma. Všechny tyto informace činí slovník velmi cenným zdrojem pro data potřebná k zmapování derivačních vazeb.

Školní slovník současné češtiny (czcz)

Slovník vydaný nakladatelstvím Lingea se soustředí na aktuální slovní zásobu, zejména nejběžnější slova, nová slova, odborné termíny, slova cizí i nespisovná [10]. Obsahuje velké množství lemmat spisovné češtiny a poukazuje na některé konkrétní typy derivací mezi nimi. Tímto je vhodný nejen pro získání vazeb daných slovníkem, ale díky svému rozsahu také pro následné rozšiřování o další vazby pomocí vytváření derivačních skupin.

[1] Stránka projektu je dostupná na: https://knot.fit.vutbr.cz/wiki/index.php/Ma

Slovník	Počet lemmat	Velikost slovníku [MB]	Podpora pro derivační morfologii
Český etymologický slovník	11 141	43,5	ano
Školní slovník současné češtiny	29 310	7,6	ano
HEURÉKA	56 081	225,1	ne
Slovník neologizmů 1	4 844	3,0	ne
Příruční slovník jazyka českého	83 926	101,3	ano
Slovník spisovného jazyka českého	178 243	53,4	ano
Akademický slovník cizích slov	57 885	154,0	ano
Slovník spisovné češtiny	50 847	152,9	ano
Technický slovník naučný	52 771	138,7	ne
Wiktionary	18 665	117,2	ne

Tabulka 3.1: Parametry slovníků dostupných v projektu morfologického analyzátoru.

HEURÉKA – univerzální encyklopedie (Heu)

Jedná se o vydání rozsáhlé všeobecné české elektronické encyklopedie nakladatelstvím LEDA, v rámci kterého byl vydán i odpovídající slovník pojmů [8]. Zabývá se především ustálenými víceslovnými výrazy, pojmenovanými entitami v češtině, zkratkami, jmény a odbornými pojmy. Slovník neposkytuje žádná data k derivační morfologii, jeho největší využití by bylo ve výzkumu zaměřeném např. na rozpoznávání pojmenovaných entit v textu[2].

Slovník neologizmů 1 (neologizmy)

Slovník neologizmů 1 vychází z databáze Neomat, která je sestavována od konce 20. století a byla tvořena jako neologický excerpční materiál pro lexikografické účely. Z této databáze postupně vznikly ještě další dva slovníky – Nová slova v češtině, Slovník neologizmů 2 a sborník statí Neologizmy v dnešní češtině [14].

Obsahem jsou české neologizmy a nové zkratky, uvádí jejich výslovnost, genitiv singuláru daného lemmatu a popisuje jejich význam. I přes to, že slovník obsahuje lemmata, které jsou derivacemi, tato vazba není ve slovníku uvedena, primárně tedy neposkytuje data k derivační morfologii.

Příruční slovník jazyka českého (psjc)

Příruční slovník jazyka českého je výkladovým slovníkem, klasifikovaným jako vědecký deskriptivní slovník. Vznikal v polovině 20. století, digitalizovaná forma byla zpřístupněna v roce 2007. Jeho hlavním účelem bylo pevně stanovit spisovná slova českého jazyka a vytvořit tak jasnou hranici mezi spisovností a nespisovností. Soustředí se především na dokonalejší lexikografický popis slov nebo použití citátů v definicích lemmat [15].

[2] Známé také pod anglickým ekvivalentem named-entity recognition (NER).

Slovník je cenným materiálem nejen z hlediska svého rozsahu, ale i pro derivační morfologii. Obsahuje položku *Odkaz*, kde uvádí lemma ve slovníku, se kterým daný výraz souvisí, čímž poskytuje spolehlivá data pro mapování derivačních vazeb, a to nejčastěji mezi substantivy a adjektivy, případně adjektivy a adverbii.

Slovník spisovného jazyka českého (ssjc)

Slovník spisovného jazyka českého vychází z lístkového lexikálního archivu, založeného při tvorbě Příručního slovníku jazyka českého. Narozdíl od něj se ale řadí mezi kodifikační slovníky, zachycuje více terminologii a význam vysvětluje v minimálních kontextech. Klade důraz na systémové zpracování slovní zásoby, podrobnou stylistickou charakteristiku slov a podporuje hnízdování, čímž uvádí vztahy mezi slovy cenné pro mapování derivačních vazeb. Elektronická verze byla zpřístupněna Ústavem pro jazyk český Akademie věd České republiky v roce 2011 [16].

Akademický slovník cizích slov (scs)

Slovník zachycuje slova, slovní spojení, zkratky a značky cizího původu, které se objevují v běžně používaném českém jazyce. Lemmata, řadící se mezi přejatá slova, doprovází definice popisující vlastnosti sémantické, pravopisné, výslovnostní a stylistické. Slovník vytvořil kolektiv lexikografů Ústavu pro jazyk český Akademie věd České republiky a jeho elektronickou podobu v roce 1999 vydalo nakladatelství LEDA, s. r. o. pod názvem Velký slovník cizích slov [13].

Díky své struktuře poskytuje velké množství derivačních vazeb, zejména mezi substantivy a adjektivy. Slovník obsahuje i verba, patřící do stejné derivační skupiny, avšak ta nejsou slovníkem s příslušnými lemmaty přímo propojena, proto musí být toto propojení předmětem jiného typu zpracování derivačních vazeb. Slovník je také vhodný pro mapování dublet obsažených lemmat.

Slovník spisovné češtiny (ssc)

Slovník spisovné češtiny je výkladový slovník českého jazyka reagující na novodobý vývoj jazyka. Byl vytvořen Ústavem pro jazyk český Akademie věd České republiky pod názvem Slovník spisovné češtiny pro školu a veřejnost a jeho elektronickou podobu vydalo nakladatelství LEDA v roce 2005 jako Slovník spisovné češtiny [17].

Obsahuje výrazy spisovné i hovorové, doprovázené rozsáhlým popisem významu, informací z oblasti flektivní a derivační morfologie. Slovní zásoba tohoto slovníku zasahuje do širokého spektra témat včetně jmen, zeměpisných údajů nebo odborných názvů. Stejně jako Akademický slovník cizích slov zahrnuje i dublety.

Technický slovník naučný (tsna)

Slovník obsahuje technické a jiné odborné výrazy, především je ale zaměřen na odborná slovní spojení, k nimž uvádí obor, v němž je daný výraz používán a související význam, popř. vysvětlení. Slovník neposkytuje žádná data vhodná pro derivační morfologii.

Wiktionary (wiktionary)

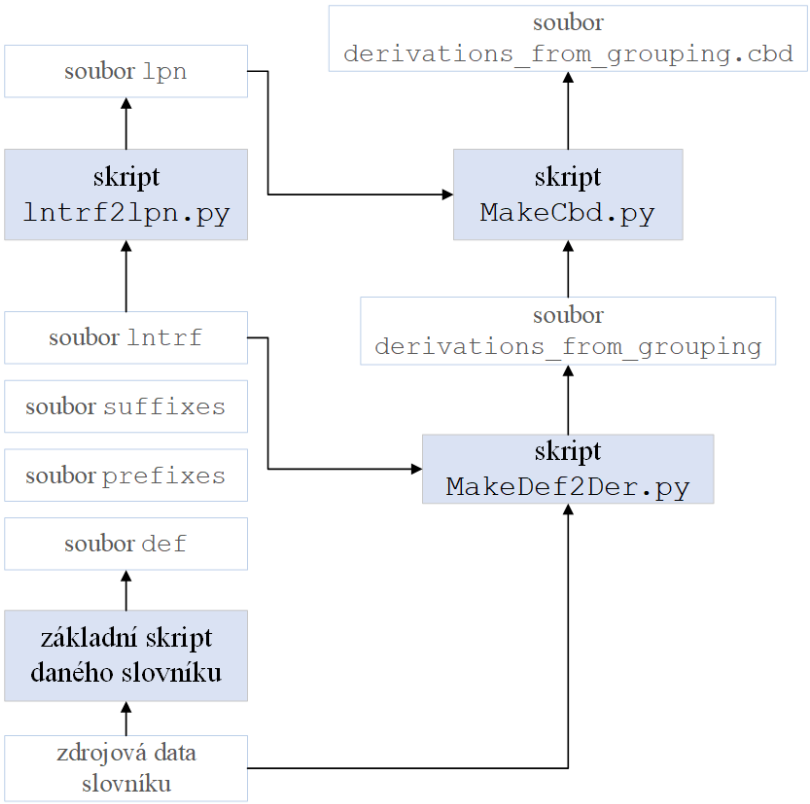
Slovník vycházející z otevřené encyklopedie Wikipedie je mnohojazyčný slovník obsahující překlady mezi jazyky, definice, výslovnost i etymologické údaje. Stejně jako Wikipedie,

i Wikislovník, v anglickém jazyce Wiktionary, je vytvářen otevřeným způsobem. Česká verze tohoto slovníku byla založena v roce 2004 a v současnosti obsahuje okolo 108 092 hesel, čímž se řadí mezi pátý největší Wikislovník na světě. V rámci projektu morfologického analyzátoru je k dispozici verze z roku 2015 [19].

Wiktionary je velmi obsáhlý z hlediska typů informací, které o daném výrazu poskytuje. Jedná se o výslovnost, překlad do cizích jazyků, dělení na slabiky, skloňování, význam, synonyma atd. Sekce *Související* obsahuje slova, která souvisí s daným výrazem. Nejedná se ale o data vhodná pro stanovení známých derivačních vazeb, neboť tato slova nejsou vždy derivacemi. Povaha vytváření tohoto slovníku se projevila ve faktu, že není udržována striktní lingvistická struktura. Tento fakt ovšem nevylučuje, že data mohou být využita při pozdějším zpracování.

3.1 Současné zpracování zdrojových dat

Na začátku kapitoly 3 byly popsány slovníky, jejich obsah a vhodnost pro získávání derivačních vazeb. Všechny zmíněné slovníky procházejí stejným způsobem zpracování, avšak kvůli jejich odlišné struktuře má každý na nejnižší úrovni svůj vlastní skript, který je stavěn na míru konkrétnímu slovníku. V této kapitole budou rozebrány soubory, které při základním zpracování každého slovníku vznikají a jsou spjaty se systémem pro automatické zařazování slov na základě derivačních vazeb. Obrázek 3.1 ilustruje, jak jsou postupně zdrojová data zpracována a získány níže popisované soubory, případně jak jsou použity dalšími skripty.



Obrázek 3.1: Postupné zpracování zdrojových dat slovníků.

Jedním z prvních souborů, které při zpracování slovníku vznikají, jsou soubory typu `def`. Jejich obsahem jsou základním skriptem upravená zdrojová data slovníku do čitelné podoby. Jejich formát (viz výpis 3.1) je *lemma* a na následujícím řádku její *slovníková definice* i se všemi informacemi, které slovník poskytuje. Původně tento soubor sloužil jako zdroj dat pro vytváření derivací, avšak formátováním definic při zpracování se ztratily některé derivační vazby, které slovník uváděl. Z tohoto důvodu je nyní pro získávání derivačních vazeb výchozím souborem zdroj slovníku.

```
mandlovník
 , -u m; bot. mandloň
mandlový
 příd. k mandle ; připomínající mandle : mandlové oči ve tvaru mandlí ;
 kuch. mandlová hmota z rozetřených sladkých mandlí a cukru ; chem.
 kyselina mandlová kyselina fenylglykolová, surovina pro výrobu léčiv ;
 bot. vrba mandlová mandlovka
```

Výpis 3.1: Výpis ze souboru `scs.def` pro ilustraci formátu dat typu `def`.

Dalšími soubory, které vznikají během zpracovávání jsou soubory s předložkami (soubor typu `prefixes`) a příponami (soubor typu `suffixes`), které slovník uvádí mezi ostatními lemmaty. Oba tyto soubory obsahují výčet předpon a přípon uváděných slovníkem. Pro ilustraci je uveden výpis 3.2 jako ukázka jednoho z typů formátů těchto souborů.

```
non-
  předp. Z lat. nōn {ne, nikoli} a odtud jako předp. i do fr. a angl. Srov.
   nonšalantní , nonsens , nonstop , nonkonformismus.
ob-
  předp. Z lat. ob- {ob-, proti-, za- ap.}. Někdy splývá s následující
   souhláskou (viz ofenziva , okupovat , oponovat ). K původu viz ob.
```

Výpis 3.2: Výpis ze souboru `cety.prefixes` pro ilustraci formátu dat typu `prefixes`.

Hlavním výsledkem celého zpracování zdrojových dat slovníku jsou nalezená slovníková hesla zpracovaná, pomocí informací v jejich definicích, do souboru typu `lntrf`. Ten poskytuje informace z hlediska flektivní morfologie ve formátu *lemma note tag*, kde *lemma* je slovníkový výraz, *note* je poznámka o stylu hesla (používá se jen pro zvláštní stylistické případy) a *tag* je značka[3], která nese informace o slovním druhu, rodu, čísle, pádu a o koncovce flektivního tvaru daného slova v tomto pádu.

```
taktometr  k1g[MI]nSc2[^:]*::u
taktoskop  k1g[MI]nSc2[^:]*::u
taktovat  k5eAaImF::
taky  k0::
takže  k8::
```

Výpis 3.3: Výpis ze souboru `psjc.lntrf` pro ilustraci formátu dat typu `lntrf`.

[3]Popis všech variant značek viz oficiální dokumentace projektu dostupná na http://knot.fit.vutbr.cz/wiki/index.php/Morfologick%C3%BD_slovn%C3%ADk_a_morfologick%C3%BD_analyz%C3%A1tor_pro_%C4%8De%C5%A1tinu.

Další úroveň zpracování představuje skript `lntrf2lpn.py`, společným pro všechny slovníky, jehož výsledkem je soubor typu `lpn`. Tento soubor zapisuje získaná data ve formátu *lemma#paradigm#note*, kde *lemma* je slovníkový výraz, *paradigm* je nalezený vzor a *note* je poznámka (není povinné). Vzorový výpis 3.4 ilustruje, jak jsou tato data zapsána v souboru.

```
pranýř#stroj#
pranýřek#vršek#
pranýřovat#účtovat#
pranéřovat#účtovat#wZ
```

Výpis 3.4: Výpis ze souboru `psjc.lpn` pro ilustraci formátu dat typu `lpn`.

Jak bylo zmíněno v úvodu, tato práce navazuje na výsledky mé bakalářské práce [5]. V té byly derivace zpracovány pouze v rámci Slovníku současné češtiny. Vznikl tak soubor s derivačními vazbami `ssc.derivations_from_grouping` (v bakalářské práci nazvaný jako `ssc.derivations_from_definitions`), založený na souboru `ssc.def`, kde jsou nalezené derivační vztahy ohodnoceny trojmístnou značkou reprezentující typ této vazby. Ukázka způsobu zápisu dat v souborech tohoto typu je uvedena ve výpisu 3.5.

```
finance -> finanční 1.0.8:ce:ční
finance -> finančník 1.9.9:ce:čník
finance -> finančnictví 1.1.1:ce:čnictví
finance -> financovat 5.0.2:e:ovat
finiš -> finišovat 5.0.2::ovat
finiš -> finišman 1.9.9::man
fintit -> fintit se 5.0.4:: se
firma -> firemní 2.0.5:ma:emní
```

Výpis 3.5: Výpis ze souboru `cety.derivations_from_grouping` pro ilustraci formátu souborů typu `derivations_from_grouping`.

Z něj je poté vytvářena jeho varianta, soubor typu `cbd` (viz výpis 3.6), který je propojením souborů typu `derivations_from_grouping` a `lpn`.

```
1.1.1:mědiryt#bez:mědirytina#slina#hT
1.1.1:menševizmus#komunismus:menševictví#stavení
1.1.1:oceloryt#návod:ocelorytina#slina#hT
1.1.1:předurčení#stavení:předurčenost#kost#hT
```

Výpis 3.6: Výpis ze souboru `ssc.derivations_from_grouping.cbd` pro ilustraci formátu dat typu `cbd`.

Kapitola 4

Návrh systému pro automatické zařazování neznámých slov

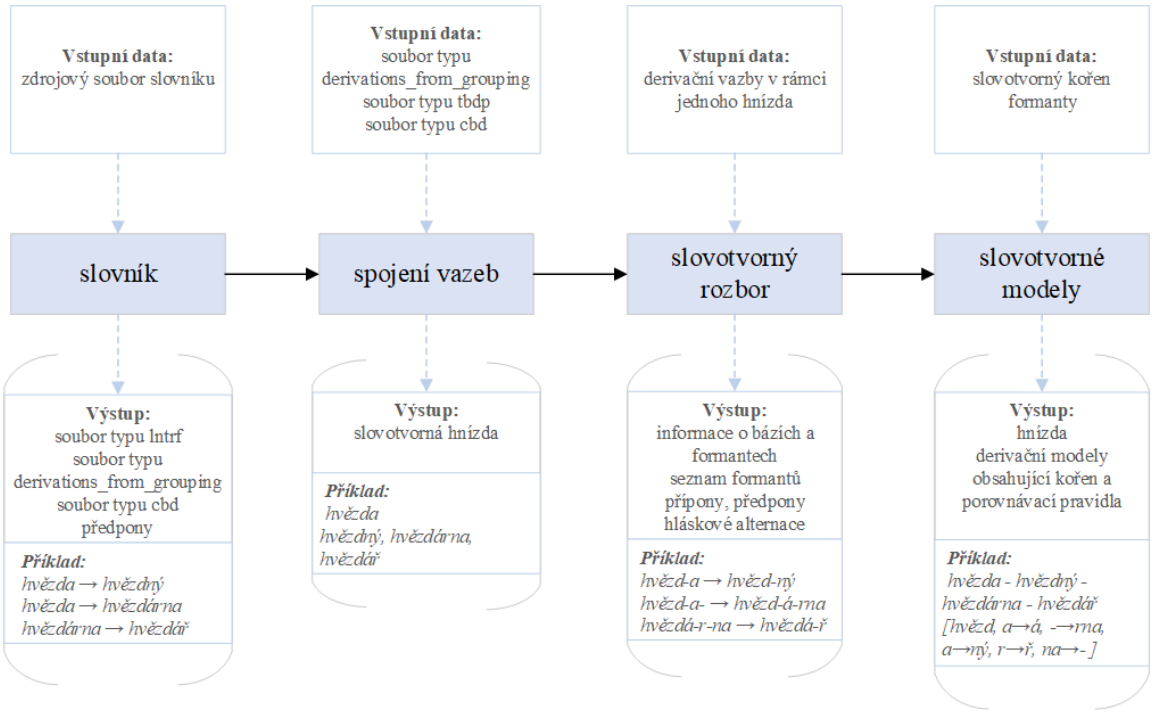
Systém je volně inspirován způsobem, jakým se lidský mozek učí řeči. Ač přesný postup, jakým k tomu dochází, není zcela objasněn, studie v oblasti neurolingvistiky odkrývají, že základ tvoří rozklad na menší části. Myšlenkou tohoto systému je nevytvářet umělá pravidla, definovaná v systému, ale nechat jazyk a lingvistická data tato pravidla utvářet tak, aby se systém učil ze slovní zásoby, se kterou se setkává.

Cílem tohoto systému je umět přiřadit novým slovům jejich derivační model, a tím určit jejich slovo fundující, díky čemuž je možné připojit jej do příslušné skupiny – slovotvorného hnízda. Aby bylo možné správně začleňovat slova, která zatím nejsou přiřazena k žádnému hnízdu, je třeba udělat slovotvorný rozbor a následně vytvořit rozhodovací logiku, která bude umožňovat tato slova správně zařazovat. Celý systém je proto pomyslně rozdělen na dvě části – první část pro utváření slovotvorných modelů (viz obrázek 4.1) a druhou část, která slouží k zařazování nových, zatím neznámých, slov do skupin.

Celý proces začíná zpracováním zdrojových dat slovníků pomocí jejich základního skriptu a vytvoření souborů s daty (rozebráno v kapitole 3.1). Tím získáme derivační dvojice zapsané v souborech typu `[slovník].derivations_from_grouping` podle vztahů mezi slovy uvedenými ve slovníku. Dalším možným vstupem jsou sobory typu `cbd`. Získání dat potřebných pro první část systému je přípravným krokem před samotným vytvářením modelů a klasifikací.

První část systému nad těmito daty zpracuje slovotvorný rozbor vzájemným porovnáváním slov v získaných vazbách. Tímto krokem se určí slovotvorná báze a formanty. Neboť se v českém jazyce vyskytují hojně výjimky a nepravidelnosti, je třeba, aby systém pracoval i s hláskovými alternacemi, které jsou součástí derivačního procesu (popsáno v kapitole 2.2). Český jazyk vytváří hláskové alternace pouze mezi některými hláskami (tzn. nemůže docházet k proměnám hlásky na jakékoliv ostatní), což je informace, se kterou systém může pracovat.

Následně budou získaná data z jednotlivých dvojic spojena v rámci hnízda, do kterého tyto dvojice patří. Hnízda vznikají spojováním derivačních vazeb pomocí společných slov, čímž dojde k seskupení dat, jež slovník takto nepřímo propojuje. Účelem je utvořit pravidla derivace, která jsou určena nalezenými formanty, proměnlivými částmi slov. Ta jsou spjatá s konkrétním hnízdem, z něhož vzešla. Každé hnízdo také sdílí nejmenší určenou slovotvornou bázi (ve většině případů se jedná přímo o slovotvorný kořen), společnou pro lemma



Obrázek 4.1: První část systému vytvářející slovotvorné modely.

a všechny jeho derivace. Vznikem hnízda, k němu náležících derivačních pravidel a báze, resp. kořene, se utvoří slovotvorný model.

Vytvořením slovotvorných modelů získáme informace o tom, jaké derivační změny mohou v daném hnízdě nastat, což je vstupní informací pro druhou část systému – klasifikátor. Způsob fungování byl inspirací pro pojmenování systému s klasifikátorem DeMoClas z anglického „derivation models-based clasificator“.

Pokud na vstup systému přichází nové slovo, které není zařazené do žádného z hnízd, vyvstává potřeba jej klasifikovat do jednoho z existujících hnízd. Pomocí porovnání slova s bázemi hnízd jsou vyhledány přípustné modely. Následně je provedena analýza, nakolik dané slovo odpovídá pravidlům vybraného modelu. Pokud získáme odpovídající podobnost, pak je slovo zařazeno do slovotvorného hnízda.

V případě, že je výsledek klasifikátoru správný, je možné nově klasifikované slovo zařadit do stávajícího hnízda. Díky tomu nastane nejen rozšíření dat, ale také úprava formantů hnízda (v některých případech může dojít i k zpřesnění báze), což vede k rozšíření derivačního modelu. Nebo=t jsou data uchovávána v souborech, jsou trvalého charakteru a jejich rozšiřováním a aktualizací se celý systém učí a zdokonaluje.

Když slovo neuspěje ani v jednom z navržených modelů, spadne do skupiny jedináčků. Tato skupina je opakovaně kontrolována, zdali není možné některé ze slov zařadit do hnízda. Vhodná chvíle pro pokus opětovného zařazení nastává, pokud byla do systému přidána nová data, rozšiřující stávající hnízda, čímž mohlo dojít k úpravě modelů. Těmito daty může být jako nový zdroj derivačních vazeb, tak nově klasifikovaná slova.

Kapitola 5

Implementace systému DeMoClas

Pro implementaci systému byl zvolen jazyk Python verze 3. Tento programovací jazyk je vhodný pro práci s textovými daty. Zároveň jsem využila datových struktur, které tento jazyk nabízí a jejich funkcí pro reprezentaci složitých dat, obsahujících různé složky, které je potřeba uchovávat. Volbu jazyka určila také nutnost navázat na implementaci ostatní funkcionality a zachování jednoty v rámci projektu morfologického analyzátoru.

Celý systém se skládá z několika částí, jak bylo popsáno v předchozí kapitole 4. Tyto části jsou spouštěny podle aktuální potřeby uživatele. První část je spuštěna pro vygenerování dat, případně pokud chceme data aktualizovat o nově zařazená slova klasifikátorem, pak jsou volány funkce implementované v této části. Samotný klasifikátor je poté spouštěn v případě potřeby klasifikovat nová slova. Vychází z dat vygenerovaných v první části systému, na základě kterých klasifikuje neznámá slova. Jeho výsledkem je informace o navržených aktualizacích a rozšířeních stávajících hnízd, případně aktualizovaný seznam v danou chvíli nezařaditelných slov, jedináčků. Poslední částí je samotný nástroj pro aktualizaci podle výsledku klasifikátoru.

5.1 Sjednocení zdrojů dat

Pro systém je naprosto zásadní mít k dispozici co největší možné množství derivačních dvojic získaných ze zdrojů obsahujících derivační vazby. Z toho důvodu došlo ke sjednocování a spojování dostupných zdrojů v projektu morfologického analyzátoru. Jak bylo rozebráno v kapitole 3 — v projektu morfologického analyzátoru jsou k dispozici slovníky poskytujících informace o derivačních vazbách.

V mé bakalářské práci byl vytvořen skript `MakeDer2Def.py`, který získával derivační vazby z definic slovníku *ssc* (soubor `ssc.def`). Při pozdější práci na projektu bylo zjištěno, že vhodnější pro získávání derivačních vazeb jsou přímo zdrojové soubory slovníků, neboť soubory typu `def` jsou uzpůsobeny pro přehledný výpis definic jednotlivých lemmat, avšak v některých případech dochází k zamlčení nebo porušení derivační vazby, kterou slovník uvádí a tato informace nepromítne do výsledných dat.

Z tohoto důvodu byl skript `MakeDer2Def.py` přepracován a pro některé slovníky byla vytvořena speciálně upravená verze, odpovídající typu zápisu dat ve slovníku. Nově tedy získává derivační vazby i z některých dalších slovníků (konkrétně *czcz*, *cety*, *ssc*, *psjc* a *ssjc*) a využívá k tomuto účelu jejich zdrojové soubory v nezpracované podobě. Ty vypisuje do souborů `[slovník].derivations_from_grouping`, které jsou umístěny ve složkách daných slovníků.

V projektu morfologického analyzátoru již dříve, v rámci diplomové práce pana Ing. Stanislava Černého [21], vznikl systém čtyřmístných značek, jehož zjednodušením bylo vytvořeno zmíněné trojmístné značkování vazeb, použité při vytváření derivací ve slovníku *ssc*. Místo, kde je tato čtyřmístná značka použita, je soubor typu `tbdp`, který je typem dat shodný se souborem `cbd`. Aby byla zaručena vzájemná kompatibilita a použitelnost i těchto derivačních vazeb, byl v rámci přípravy dat pro tuto práci vytvořen skript `tbdp2cbd.py`, který konvertuje soubory `tbdp` se čtyřmístnými značkami na soubory `cbd` se značkami trojmístnými, čímž dojde ke sjednocení formátů značek v rámci analyzátoru.

Při načítání dat do systému DeMoClas je nejprve načten zdrojový soubor s derivačními dvojicemi slovníku *cety*. První načtený slovník udává prvotní rozvržení slov do hnízd, a proto byl vybrán Český etymologický slovník, který obsahuje ve svých datech bohatou zásobu derivací pro většinu svých lemmat. Vazba vždy obsahuje bázi, slovo základní (slovníkové lemma) a derivaci, slovo od ní odvozené. Další načítané dvojice poté prochází kontrolou, zdali se báze nebo její derivace již nenacházejí v jednom z existujících hnízd. Pokud tomu tak je, dojde k připojení této dvojice do daného hnízda. V opačném případě je vytvořeno nové. Každé hnízdo reprezentuje jedna společná báze a její derivace. Díky použitému přístupu vzniká spojování vazeb do obsáhlejších skupin a není vytvářeno velké množství hnízd o malém počtu slov. Následují data z ostatních slovníků a jako poslední jsou načteny derivační vazby uvedené v souboru `czech.tbdp`, které doplní zdrojová data slovníků.

5.2 Vytváření derivačních modelů

V této části systému dochází ke zpracování získaných hnízd. Výstupem jsou derivační modely, sloužící k následné klasifikaci nových, nezařazených slov do hnízd. Tuto část funkcionality zajišťuje skript `MakeDerivationModels.py`.

Zarovnávání slov pomocí skórovací matice

Pro zkoumání změn či pravidel, vlivem kterých dochází k odvozování slov, je potřeba nalézt vhodný způsob, jak daná slova porovnat. Jejich délky jsou často rozdílné, během procesu derivace v nich dochází k různým změnám (hláskovým alternacím apod.) a často i k přidání předpony u jednoho nebo obou slov. Správně zarovnat obě slova, tedy nalézt takovou vzájemnou pozici jednoho vůči druhému, ve které se nejvíce shodují, je klíčové pro formování derivačních modelů.

Řešení zarovnávání slov je volně inspirováno způsobem, jakým se zarovnává DNA. Na vstupu funkce je dvojice slov – báze a její derivace, u nichž potřebujeme zjistit slovotvornou bázi, neměnnou část, a formanty, tedy části, které se od původního lemmatu mění a utváří nové tvary – derivace. Aby bylo možné určit bázi, která bude nejblíže společnému kořeni slov a odlišit formanty, je třeba obě slova co nejlépe zarovnat a nalézt část, kterou mají společnou.

Obě slova jsou vzájemně porovnávána způsobem postupného překrývání vůči sobě. V každém kroku je vypočítáno skóre vyjadřující vzájemnou shodu či neshodu aktuálního zarovnání. To, které získá nejlepší skóre, je bráno jako nejvhodnější možné a je použito k získání báze a formantů. Pro výpočet skóre jsem vytvořila skórovací matici (viz tabulka 5.1), v níž je zaneseno ohodnocení shody jednotlivých písmen české abecedy.

Ohodnocení míry shody jednotlivých písmen se odvíjí od změn, typických pro český jazyk, kde standardně dochází ke zkracování nebo prodlužování hlásek či jejich alternacím

	a	**c**	**d**	**e**	**h**	**i**	**n**	**o**	**r**	**s**	**t**	**u**	**y**	**z**	*****
á	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-1/5
č	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-1/5
ď	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-1/5
é	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-1/5
ě	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-1/5
í	-	-	-	-	-	3	-	-	-	-	-	-	1	-	-1/5
ň	-	-	-	-	-	-	3	-	-	-	-	-	-	-	-1/5
o	1	-	-	-	-	-	-	5	-	-	-	3	-	-	-1/5
ó	-	-	-	-	-	-	-	3	-	-	-	-	-	-	-1/5
ř	-	-	-	-	-	-	-	-	3	-	-	-	-	-	-1/5
š	-	-	-	-	-	-	-	-	-	3	-	-	-	-	-1/5
ť	-	-	-	-	-	-	-	-	-	-	3	-	-	-	-1/5
u	-	-	-	-	-	-	-	3	-	-	-	5	-	-	-1/5
ú	-	-	-	-	-	-	-	-	-	-	-	3	-	-	-1/5
ů	-	-	-	-	-	-	-	1	-	-	-	3	-	-	-1/5
y	-	-	-	-	-	1	-	-	-	-	-	-	5	-	-1/5
ý	-	-	-	-	-	1	-	-	-	-	-	-	3	-	-1/5
z	-	-	-	-	-	-	-	-	-	3	-	-	-	5	-1/5
ž	-	-	-	-	1	-	-	-	-	-	-	-	-	3	-1/5

Tabulka 5.1: Skórovací matice pro zarovnávání dvou slov v českém jazyce.

(např. *třást – třes – třaslavý*, *svit – svítit – světlo*), vložení samohlásek (např. *blud – bloudit*) a jiné (např. *běžet – běhat – ubíhat*, *abstrakce – abstrahovat*). Výpočet tedy zahrnuje varianty, kdy jsou aktuálně porovnávaná písmena zcela shodná, pak je ohodnocení 5 bodů. Pokud nejsou zcela shodná, přidělí se počet bodů příslušející dané dvojici písmen podle skórovací matice (1 až 3 body) a v případě, že se nejedná o přípustnou záměnu hlásek, je dvojice písmen brána jako neshodující se a dochází k odečtení jednoho bodu. Na celkové skóre aktuálního zarovnání obou slov se také aplikuje penalizace za písmena, která nejsou do porovnání započtena, jelikož se zrovna nepřekrývají s žádným písmenem druhého slova. Výše penalizace je -1 bod a byla do výpočtu zařazena z důvodů, které budou vysvětleny na následující ukázce výpočtu nejlepšího zarovnání lemmatu a derivace (tabulka 5.2).

V případě, kdy je alespoň jedno slovo kratší než 5 znaků, a navíc se jedná o slova, která se neshodují ve větší části písmen, je určení společné báze nesnadné. Na příkladu lze vidět, že pokud by bylo bráno v potaz pouze získané skóre, byla by za nejlepší zarovnání označena druhá možnost (skóre je 4), která je nesprávná. Tato chyba vzniká z matematického

Zarovnání slov vést a svůdce						Skóre	Penalizace	Celkové skóre zarovnání
	v	**é**	**s**	**t**				
				s	vůdce	-1	-8	-9
			s	v	ůdce	4	-6	-2
		s	v	ů	dce	-3	-4	-7
	s	v	ů	d	ce	-4	-2	-6
s	**v**	**ů**	**d**	**c**	**e**	**2**	**-2**	**0**
sv	ů	d	c	e		-4	-2	-6
svů	d	c	e			-3	-4	-7
svůd	c	e				-2	-6	-8
svůdc	e					-1	-8	-9

Tabulka 5.2: Ukázka výpočtu nejlepšího zarovnání lemma a její derivace.

důvodu, jelikož při výpočtu, kdy se překrývají pouze dvě písmena z každého slova a jeden porovnávaný pár je totožný (v příkladu písmeno *s*), vznikne zisk 5 bodů, z nějž už není možné odečíst víc než jeden bod, neboť více písmen již není zapojených. Je tedy potřeba vzít v potaz množství porovnávaných písmen z celkové délky slova, neboť paradoxně, čím menší je počet porovnávaných písmen, tím menší je celkový možný součet odečtených bodů a dochází k výběru zcela nevyhovujícího zarovnání.

V procesu zarovnávání lze také odhalit případy, kdy došlo během derivace k vložení hlásky (např. *žhnout – ožehnout*). U těchto dvojic dojde k zarovnání na společnou část slova, která následuje až za vloženou hláskou tak, aby došlo k oddělení celé přední části (*ž-hnout a ože-hnout*). Detekce vložené hlásky lze následně snadno provést porovnáním písmen před a za písmenem, které bylo vloženo, zarovnání o inserci posunout a rozšířit tak rozpoznanou společnou část. Vznikne zarovnání, které připouští různé varianty slovotvorné báze (*o-ž[e]*hnout*).

Hledání formantů

Ve chvíli, kdy známe nejlepší možné zarovnání báze a derivace, je možné jejich vzájemným porovnáním určit formanty. Mezi formanty patří předpony, přípony, vpony a případné hláskové změny uvnitř slova. Je nutné porovnávat písmena z obou slov na těch pozicích, které si odpovídají po zarovnání. Jako první dojde k rozpoznání a odtržení předpony, pokud ji některé ze slov obsahuje. Předponu signalizuje zejména pozice slov vůči sobě. Pokud je výsledná pozice zarovnání větší než nula, pak předponu obsahuje báze (v zarovnání skončila pomyslně předsunuta před derivaci), v opačném případě je výsledná pozice zarovnání záporná a předponu obsahuje derivace. I v takové situaci je ale třeba brát v potaz případy, kdy jsou předponovaná obě slova, jen jedna z předpon je delší než druhá. Další variantou je případ, kdy jsou obě slova předponovaná, ale předpony mají stejnou délku. Pak je výsledná pozice zarovnání rovna nule, ale počáteční písmena slov se neshodují, neboť obsahují předpony. Z tohoto důvodu nelze spoléhat pouze na výslednou pozici zarovnání, ale zároveň

musí dojít k nalezení prvního společného písmene, což zpravidla odpovídá prvnímu písmenu společného kmene či kořene obou slov. Ve výjimečných případech může dojít i k situaci, kdy jsou obě slova předponovaná a předpony začínají na stejné písmeno (např. *položit – podložit*). U těchto dvojic dojde k označení rozdílného písmene předpony jako za hláskovou změnu a celá slova jsou brána jako jedna slovotvorná báze (vznikne slovotvorná báze *po[d]*ložit*, bez rozpoznání předpon). Ke správnému rozkladu slov na bázi a formanty (mělo by být *pod-lož-it*, resp. *po-lož-it*) dojde až během určování společného kmene celého hnízda. Tento proces bude popsán v následující podkapitole.

Po odtržení předpony dochází k identifikaci společného základu slov – slovotvorné báze, která představuje kmen společný pro bázi i derivaci. V některých případech může dojít až k separaci samotného kořene, ale není to podmínkou ani pravidlem. Jelikož se jedná o část, kterou mají obě slova společnou, vzniká vyčleněním shodné části slov. Jak bylo již naznačeno, v českém jazyce někdy dochází v rámci slovotvorných bází ke vkládání hlásek do slov nebo k jejich změnám, proto pokud během zarovnání dojde k detekci vložené hlásky, je tento jev zaznamenán jako varianta kmene v rámci daného hnízda. Systém rozlišuje mezi vložením hlásky a jejich alternacemi způsobem zápisu, ke kterému využívá notaci regulárních výrazů. Pro vložené hlásky je použita notace hranatých závorek s hvězdičkou, neboť daná hláska se ve slově může a nemusí vyskytovat. Jako příklady lze uvést slova *žhnout – ožehnout* (slovotvorná báze je *ž[e]*hnout*) nebo *smluvit – smlouva* (slovotvorná báze je *ml[o]*uv*). Hláskové alternace jsou zaznamenány zápisem, který odráží způsob fungování alternací v procesu derivace a tím je záměna jedné hlásky za jinou. Notace využívající hranaté závorky se symbolem plus tedy určuje, že jedna z variant musí být ve slově přítomna. Příkladem mohou být změny hlásek *ů* a *o* ve skupině *dolík -– důlek – údolí – dolovat – důlní*, které spojuje slovotvorná báze *d[ůo]+l* nebo *dům — domek – domovník*, kde se slovotvornou bází stává *d[ůo]+m*. Aby nedocházelo k falešně pozitivním detekcím alternací u nesouvisejících slov, provádí se kromě porovnávání okolních písmen i kontrola se skórovací maticí, využívanou při zarovnávání slov, která obsahuje kladné číselné skóre pro dvojice písmen, jejichž záměny se v českém jazyce vyskytují a jsou tím pádem povoleny detekovat jako alternace. Matice tímto získává více druhů využití a slouží zároveň jako zdroj povolených záměn hlásek.

Určením slovotvorné báze na závěr vyplynou i případné sufixy. Tímto procesem je postupně vytvořen seznam formantů, které byly určeny díky rozboru daných dvou slov. I u formantů je využita notace regulárních výrazů, aby bylo možné rozlišit, které formanty jsou předponami, příponami a alternacemi, případně vloženými hláskami. Tento způsob zápisu byl cíleně zvolen pro další práci klasifikátoru s formanty a bázemi, což bude blíže popsáno v kapitole 5.3. Mezi tím, co prefixy jsou symbolizovány počátečním znakem „^ “ , který je v regulárním výrazu umisťuje na začátek slova a po nich musí následovat další písmena (báze), sufixy musí být předcházeny nenulovým počtem písmen a jsou ukončeny symbolem „$“, který udává jejich pevné umístění na konec slova. Výstup této analýzy (výpis 5.1) je tedy báze, seznam jejích derivací, seznam slovotvorných bází, který vznikl porovnáváním jednotlivých dvojic báze-derivace hnízda a seznam získaných formantů (prefixy a sufixy).

```
žhnout
['ožehnout', 'ožehavý', 'žhoucí']
['ž[e]*hnout', 'ž[e]*h', 'žh[n]*ou']
['^o', 'avý$', 'nout$', 't$', 'cí$']
```

Výpis 5.1: Ilustrativní výpis získaných dat po analýze formantů v rámci jednoho hnízda.

Díky určení nejdelší společné slovotvorné báze je možné provést kontrolu dat. Pokud je společná slovotvorná báze kratší než 3 písmena (včetně možných alternací), pak je derivace z tohoto hnízda vyčleněna a formanty vzniklé porovnáváním báze s touto derivací jsou zahozeny. Důvodů, proč vzniká tato neshoda mezi bází a některými derivacemi udanými slovníkem je více. Některé slovníky mezi derivacemi uvádějí z hlediska sémantiky související, avšak z hlediska morfologie zcela odlišné výrazy (např. vazby, které uvádí Český etymologický slovník: *vést – uvést – svůdce – svůdný*). Další aspektem je i velká nepravidelnost češtiny jako jazyka, proto v některých případech vzniká mezi bází a derivací příliš velký rozdíl na to, aby byla určena společná slovotvorná báze – a to i při aplikování možnosti hláskové alternace či vložení hlásky.

V takovém případě dojde k rozdělení hnízda na 2 a více různých hnízd, definovaných svými specifickými sadami slovotvorných bází a formantů. Jako příklad takové skupiny může být hnízdo *číhat – počíhat si – vyčíhat – očíhnout – čihadlo – čižba*, kde poslední zmíněné slovo bude odtrženo do samostatné skupiny. Nově vzniklá hnízda, skládající se z vyčleněných slov původního hnízda, jsou ihned zpracována stejným způsobem, jako bylo výše popsáno a zapsána do souboru czech.nests se všemi náležitostmi. V systému jsem se rozhodla pro spíše striktní přístup – radši hnízda rozdělovat, případně pokud nevznikne nové hnízdo (vyčleněné slovo je pouze jedno nebo ani mezi vyčleněnými slovy není nalezena společná slovotvorná báze), slova zařadit do seznamu jedináčků (dále vysvětleno v kapitole 5.3). Díky tomuto přístupu nevznikají formanty a slovotvorné báze, které jsou chybné a vytvořená pravidla se vyznačují vysokou měrou spolehlivosti. Jelikož se jedná o systém, který se neustále učí, je výhodnější si klasifikaci slov, u nichž není dostatečná jistota správnosti, nechat na dobu, až bude mít systém dostatek informací, aby je spolehlivě zařadil.

Získání kořene hnízda

Poté, co jsou získány formanty a slovotvorné báze porovnáním všech dostupných derivací pro danou bázi, je třeba najít nejkratší slovotvorný základ (v mnohých případech dojde až k získání kořene) společný pro toto hnízdo.

Za výchozí je označena nejkratší nalezená slovotvorná báze. Díky porovnání výchozí báze s ostatními slovotvornými bázemi dostáváme další formanty, zejména infixy, které vyplývají nejen z porovnání mezi bází a derivacemi, ale také se zde projeví rozdíly mezi jednotlivými derivacemi. Tímto krokem se ustanoví hlavní kořen, který bude reprezentovat celé hnízdo v derivačním modelu a zároveň se zkompletuje seznam možných pravidel skládání slov, náležících do tohoto derivačního hnízda (výpis 5.2). Infixy, které tímto dalším porovnáváním získáme, bývají zpravidla vloženy mezi kořen a sufix jsou odlišeny od ostatních afixů způsobem zápisu regulárním výrazem, který umožňuje jejich umístění mezi kořen a sufix. Zakončení infixů je dvojicí znaků „.*“. Infixy mají v tomto ohledu volnější způsob umístění, neboť musí následovat za nenulovým počtem písmen (obvykle kořenem) a zpravidla jsou následovány sufixem.

```
žhnout
['ožehnout', 'ožehavý', 'žhoucí']
ž[e]*h
['^o', 'ou.*', 'nou.*', 'cí$', 'avý$', '.+t$', '.+nout$']
```

Výpis 5.2: Výpis výsledné podoby kořene a formantů pro hnízdo báze *žhnout*.

Proces určování kořene je také důležitý z hlediska informací, které se během něj sdílí. Proces analýzy se zde posouvá od zkoumání dvou slov ke zkoumání celého hnízda a jeho celkových souvislostí. Je zde k dispozici širší kontext a porovnání s dalšími nalezenými slovotvornými bázemi v daném hnízdě, což umožňuje odhalit formanty, které není možné určit pouhým srovnáním dvou slov. Příkladem je hnízdo báze *položit* (výpis 5.3), jehož problematika byla nastíněna během popisu hledání formantů mezi bází a derivací, kde nebylo možné správně odhadnout předponu, příponu ani špatně detekované vložení hlásky.

```
položit
['uložit', 'složit', 'vyložit', 'naložit', 'přiložit', 'proložit', 'založit
', 'podložit', 'podložka', 'předložit', 'předložka']
lož
['^za', '^pro', '^po', '^pod', '^při', '^na', '^s', '^před', '^vy', '^u', '
it.*', 'ka$', 'it$']
```

Výpis 5.3: Ukázka důležitosti množství dat v rámci hnízd a kontextu, který do analýzy pravidel hnízda přinášejí.

K odtržení prefixu a sufixu dojde až porovnáním slovotvorné báze *po[d]*ložit* s ostatními (v rámci hnízda vznikne např. *ložit* či *lož*), a tím ke správnému určení prefixů *po-*, *pod-* a sufixu *-it*. Skript vždy počítá s možnými hláskovými variacemi a pokud dochází k jejich oddělení od slovotvorné báze či kořene, zapisuje do pravidel všechny validní možnosti.

Vytvoření derivačních modelů

Popsaným způsobem vzniknou derivační hnízda, která jsou vypsána do souboru `czech.nests`. Tento soubor obsahuje všechna dosud získaná hnízda ze zdrojových souborů slovníků (zdrojová data jsou popsána v kapitole 3). Hnízda jsou zapsána ve formátu (viz výpis 5.4):

číslo hnízda a grafický oddělovač
báze
derivace
kořen
seznam formantů

```
10 ----------------------------------
abstinent
['abstinence', 'abstinovat']
abstin
['en.*', 't$', 'ent$', 'ovat$', 'ce$']

11 ----------------------------------
abstraktní
['abstrakce', 'abstrahovat']
abstra
['tní$', 'hovat$', 'k.*', 'ktní$', 'ce$']

12 ----------------------------------
absurdní
```

```
['absurdita']
absurd
['ita$', 'ní$']
```

Výpis 5.4: Výpis ze souboru `czech.nests`.

Číslo hnízda je unikátní identifikátor, pomocí kterého se lze na konkrétní hnízdo odkazovat. Báze je lemma dané slovníkem a seznam derivací je seznam slov, které slovník uvádí v definici daného lemmatu. Následuje kořen získaný oddělením formantů, jak bylo popsáno v předchozích podkapitolách. Kořen nemusí být unikátní, i přes to, že stejný kořen může mít více hnízd vzájemně spolu nesouvisejících. Je tomu tak z vícero důvodů, k těm nejčastějším patří různá míra ořezání slovotvorných bází od formantů (míra, s jakou se báze přibližuje skutečnému kořeni daného hnízda souvisí s množstvím dat, respektive derivací, které jsou pro dané hnízdo k dispozici v danou chvíli) a výskyt homonym v českém jazyce, například *župan* (od slova *župa*) a *župan* (oblečení), kde obě hnízda mohou mít kořen *žup* v určité skladbě dostupných derivací pro každé z hnízd. Kořen tedy není možné použít jako identifikátor hnízda, na rozdíl od čísla. Jako poslední je vypsán seznam formantů, jejichž původní pozici ve slově lze rozlišit pomocí znaků regulárních výrazů.

Tento způsob zápisu je uživatelsky přívětivější, avšak pro následné automatizované zpracování příliš květnatý a složitý. Zároveň je třeba brát v potaz, že klasifikátor bude potřebovat v hnízdech vyhledávat, a jelikož čeština je jazyk velmi bohatý, musíme počítat s velkým množstvím dat. Optimalizace pro automatizované zpracování a vyhledávání slov v tomto případě hraje důležitou roli. Pro tento účel byl vytvořen soubor `democlas.index` (názorná ukázka ve výpisu 5.5), který je obdobou knižního rejstříku. V tomto souboru jsou vypsána všechna slova náležící do jednoho z hnízd, seřazená podle abecedy, zapsaná ve formátu:

slovo <tab><tab> číslo hnízda

Tento formát byl zvolen záměrně, neboť řádek začínající slovem, dva tabulátory jako oddělovač, a poté vypisovaná hodnota daného souboru je běžný způsob zápisu různých dat v projektu morfologického analyzátoru a vnímám jako vhodné udržovat určitou jednotu ve způsobu zápisu v rámci jednoho projektu. Zároveň je tento formát dobře čitelný a zpracovatelný.

```
alkohol  53
alkoholický  53
alkoholik  53
alkoholismus  53
alt  54
altán  55
altánek  55
alternace  57
alternativa  56
alternativní  56
alternovat  57
altistka  54
```

Výpis 5.5: Výpis ze souboru `democlas.index`.

Sestavení samotných derivačních modelů vychází z dat v souboru `czech.nests`. Modely jsou ale určeny k prohledávání a poskytování potřebných informací klasifikátoru, který z nich čerpá a vyhledává podle jejich pravidel nejpravděpodobnější hnízdo, do kterého slovo patří. Tomuto účelu je přizpůsoben i jejich formát, v jakém jsou vypsány (viz výpis 5.6). Každý řádek souboru `czech.dms` (zkratka z anglického „derivation models“) obsahuje následující informace:

číslo hnízda <tab><tab> kořen <tab><tab> formanty

```
484  crč  ['et$', 'ivý$']
485  ct[íi]+t  ['el$', 'elka$', '^po', '^u']
486  ctnost  ['ný$']
487  cuc  ['t$', 'at$', 'a.*', 'nout$', 'vý$']
488  cuc[e]*k  ['ovat$', 'ovitý$']
489  cud  ['ný$', '^ne', 'n.*', 'ost$', 'ný.*', 'a$', 'ý$']
490  cuch[t]*a  ['^po', '^roz', 't.*', 't$']
491  cuk  ['^za', 'nutí$', 'at$', 'nout$', 'at.*', '^u']
492  cuk  ['ina$', 'eta$']
493  cuk  ['ovar$', 'ářství$', 'natý$', 'ovím$', 'ovka$', 'árna$', 'árka$',
'natět$', 'ovinka$', 'r.*', 'ář$', 'r$', 'er.*', 'řenka$', 'ový$', 'árský$
', 'ovat$']
```

Výpis 5.6: Výpis ze souboru `czech.dms`.

Jedná se tedy o nejpodstatnější informace, které klasifikátor potřebuje a jsou vynechány například údaje o slovech, která do daného hnízda patří. Ty lze nalézt buď čtením samotných hnízd v souboru `czech.nestes` nebo pomocí rejstříku `democlas.index`. Díky tomuto dochází k vypuštění pro klasifikaci nepodstatných informací a k optimalizaci systému z hlediska objemu načítaných a zpracovávaných dat. Jako poslední jsou vypsány soubory `czech.prefixes` a `czech.sufixes`, které obsahují všechny nalezené prefixy a sufixy (bez infixů, protože u těch není zcela jasné, jestli jsou i sufixem). Tímto dojde k rozšíření již existujících dat morfologického analyzátoru a k vytvoření rozsáhlého souboru různých variací prefixů a sufixů v rámci českého jazyka.

5.3 Klasifikační jádro systému

Tato část systému provádí již samotné zařazování nových slov do vyhovujících hnízd, pokud takové hnízdo v systému existuje. Novým slovem se rozumí slovo, které ještě není v systému zařazeno do žádného z hnízd. Klasifikátor používá vytvořený rejstřík zařazených slov a derivační modely. Funkcionalitu nese skript `DeMoClas.py`, který jako svůj jediný argument přijímá soubor – seznam slov ke klasifikaci a jeho výstupem je číslo nejpravděpodobnějšího hnízda pro každé slovo. V případě, že žádné z hnízd dostatečně nesedí tvaru slova, je slovo zařazeno do seznamu jedináčků.

Vstupním souborem může být prakticky jakýkoliv soubor, který obsahuje slovo, které má být zařazeno, jako první údaj na řádku, oddělené libovolným bílým znakem od případného dalšího obsahu řádku. Může to být prostý seznam slov, kde na každém řádku je umístěno jedno slovo nebo i mnohé soubory dostupné v morfologickém analyzátoru. V tomto případě je dobré využít pro klasifikaci a obohacení znalostí systému například soubory typu `lntrf` (popis formátu souboru v kapitole 3.1), které obsahují lemmata dostupná ve všech slovnících

morfologického analyzátoru, a které byly pro určité slovníky rozšířeny o nově nalezená slova v průběhu zlepšování zpracování.

Pokud máme k dispozici informaci o slovním druhu slova (jako v souborech typu `lntrf`), pak je vstup rovnou vytřízen a pro účely derivační morfologie jsou ponechána ke klasifikaci pouze podstatná jména, přídavná jména, číslovky, slovesa a příslovce. Záměrně jsou vynechány slovní druhy, u nichž nemá smysl je zkoumat z pohledu derivační morfologie.

Klasifikace do hnízd

Načtená slova, určená ke klasifikaci nejprve prochází kontrolou, zdali už je systém nezná a nejsou zařazené do hnízda. Pro tento účel slouží soubor `democlas.index` – pokud se v něm slovo nachází, je již zařazeno (je k němu přiřazen číselný identifikátor příslušného hnízda, odpovídající záznamu v souborech `czech.dms` a `czech.nests`) a není potřeba provádět klasifikaci. Jelikož systém obsahuje velké množství dat, jejichž objem se předpokládatelně bude zvětšovat, byla pro vyhledávání v rejstříku implementována malá úprava. V rámci načítání rejstříku do klasifikátoru je tento seznam slov rozdělen do šesti skupin podle abecedy v tomto rozdělení: slova od *a* do *č*, *d* až *e*, *f* až *i*, *j* až *n*, *o* až *š* a *t* až *ž*. Díky tomu při vyhledávání slov systém hledá v menších částech, vybraných podle počátečního písmene hledaného slova. Toto je důležité nejen z hlediska toho, že počet slov přibývá, ale zároveň čím je systém zkušenější, tím více slov bude spíše nacházet v rejstříku než klasifikovat.

Pokud slovo v rejstříku nalezeno nebylo, dochází k vyhledávání vhodného hnízda pomocí derivačních modelů. Aby byla celá klasifikace co nejméně časově náročná, dojde v prvním kroku k vytipování vhodných potenciálních hnízd pomocí kořenů. Samotné kořeny v derivačních modelech reprezentují určitý vzorek, který se musí v daném slově vyskytovat. Slovotvorné báze často obsahují více variant, alternací, které jsou v předchozím zpracování zapsány ve formě regulárního výrazu. Porovnávání kořene se slovem se provádí tak, že je kořen dán jako vyhledávaný vzorek v daném slově. Pokud je kořen ve slově nalezen, model, který reprezentuje je zařazen do kandidátních hnízd, kam by mohlo slovo náležet.

Po získání seznamu vhodných modelů je třeba provést výběr nejpravděpodobnějšího. V případě, kdy je modelů nalezeno více, probíhá výpočet skóre, které odráží míru shody mezi složením slova a pravidly skládání formantů pro dané hnízdo, které derivační model reprezentuje. Výpočet celkového skóre ovlivňuje poměr písmen, která pokrývá kořen hnízda vůči délce slova a způsob, jakým korespondují formanty, zejména pak hláskové změny, se zkoumaným slovem. Pro tento účel jsou použita data, která nesou derivační model, a ta jsou zformována do regulárního výrazu s následující strukturou:

předpony/kořen/vpony/přípony

Pro názornou ukázku na reálných datech jsem vybrala hnízda, jejich derivační modely a vypsala v následujícím výpisu 5.7, jaká porovnávací pravidla jsou z nich vytvořena:

```
Hnízda:

5656 ----------------------------------
župa
['župní', 'župan']
žup
['a.*', 'n$', 'ní$', 'a$']
```

```
5657 ----------------------------------
župan
['župánek']
žup[aá]+n
['ek$']

Derivační modely:

5656  žup  ['a.*', 'n$', 'ní$', 'a$']
5657  žup[aá]+n  ['ek$']

Porovnávací pravidla:

()*(žup)(a)*(ní$|n$|a$)*
()*(žup[aá]+n)()*(ek$)*
```

Výpis 5.7: Výpis ze souboru `czech.dms`.

Pokud bychom měli na vstupu klasifikovat slova *župový* a *županový*, pak roli hraje rozdíl v infixu *an*, který je typický pro druhé zmíněné hnízdo, se sémantickým významem *župan* jako kus oblečení. Díky tomu je možné rozlišit, že slovo *župový* bude odvozené od slova *župa*, kdežto *županový* od slova *župan*, a to i přes to, že druhé z hnízd je ve chvíli klasifikace tvořeno pouze dvěma slovy a infix *an*, resp. *án* není osamostatněn od kořene. V případě slova *župan*, tedy při porovnání s prvním modelem, vznikne shoda s kořenem, infixem a sufixem *žup-a-n* a při porovnání s druhým modelem je shoda pouze s kořenem, tedy *župan*. I přes to, že slovo je oběma modely pokryto ve stejné míře, délka kořene pokrývající slovo má vyšší důležitost než postupné pokrytí formanty. Pak má systém tendenci upřednostňovat derivační modely s nejdelším možným odpovídajícím kořenem a dojde ke správnému zařazení do hnízda s bází *župan*. V případě slova *župový* vůbec nedojde k zařazení druhého hnízda do možných modelů, neboť jeho kořen není možné v tomto slově nalézt.

Po výpočtu skóre a vybírání nejlépe odpovídajícího modelu je třeba rozhodnout, jestli výše tohoto skóre je dostačující pro to, aby slovo bylo k danému hnízdu přiřazeno. Tento krok je zejména důležitý v případech, kdy je danému slovu nalezen pouze jeden odpovídající derivační model. V tomto případě totiž nedochází k porovnávání a vybírání nejlepšího modelu. I tak je ale nutné skóre vyhodnotit a zkontrolovat, že není příliš nízké, protože pak by v systému docházelo k zařazování zcela nesouvisejících slov do hnízd, která by tímto byla modifikována a jejich pravidla skládání formantů by byla poškozena. Největší tendenci k tomuto jevu mají hnízda, jejichž kořeny jsou velmi krátké (zpravidla 3 písmena) a jsou porovnávány s dlouhými slovy. Pravděpodobnost, že dojde v nějaké části klasifikovaného slova ke shodě tří písmen je vysoká. Následek takového zařazení ale je, že jsou vytvořeny zcela neodpovídající formanty, které když se promítnou do vytvořeného porovnávacího pravidla mohou jej natolik ovlivnit, že nedojde ani ke správnému zařazení slov, která by jinak do tohoto hnízda patřila.

Aktualizace dat systému

Systém v celkovém přístupu uplatňuje poměrně vysokou míru skepse, která je výhodnější z hlediska celé jeho konstrukce a fungování. Slova, kterým nebyl nalezen žádný odpovídající derivační model nebo u nejlepšího vybraného nedosáhla požadovaného minimálního

skóre jsou zařazena na seznam jedináčků. Jedná se o soubor `democlas.singletons`, který obsahuje tato (pro danou chvíli) neklasifikovatelná slova, která jsou sem zařazována z různých částí běhu systému. Pokaždé, když systém zpracuje nový zdroj derivačních vztahů nebo klasifikuje nová slova, čímž může dojít k modifikaci derivačních modelů či vzniku nových hnízd, znovu proběhne pokus o klasifikaci slov, která jsou zařazena mezi jedináčky. Systém se takto znovu vrací s novými informacemi a kontroluje, zdali není možné některá z těchto slov zařadit.

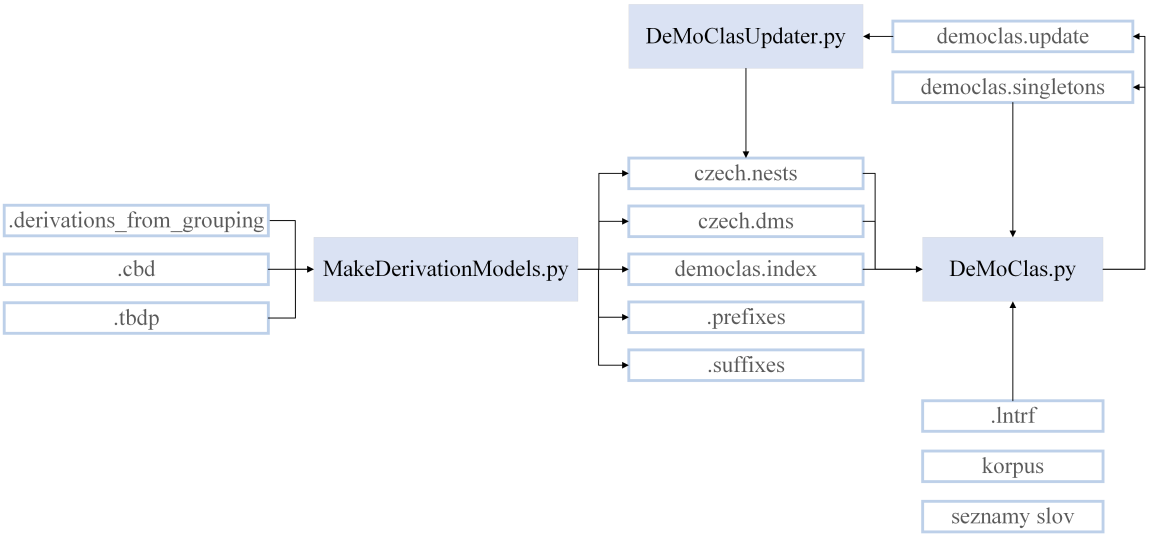
Zajímavým jevem je, pokud je systém konfrontován se slovy, jejichž hnízdo nebylo v žádném ze zdrojů pro derivační vztahy zaznamenáno. To mohou být buď slova patřící mezi archaizmy, odborné výrazy a podobné skupiny slov, která se nevyskytují v běžné slovní zásobě, anebo naopak neologizmy. Zejména díky korpusům mohou postupně do systému začít pronikat slova, která vznikla přirozeným vývojem jazyka (popsáno v kapitole 2.1) v nedávné době.

Po dokončení klasifikace je vypsán soubor `democlas.update`, který obsahuje návrhy systému na zařazení slov do konkrétního hnízda. Spuštěním aktualizace jsou znovu vypsána hnízda s novými daty, je náležitě upraven rejstřík slov v souboru `democlas.index` a derivační modely. Tuto část funkcionality zajišťuje skript `DeMoClasUpdater.py`. Díky oddělení procesu úpravy původních dat a jejich vypsáním do souboru získává uživatel přehled nad změnami systému. To usnadňuje testování, kontrolu správnosti a umožňuje nezávazné experimentování se systémem. Aktualizační nástroj dává uživateli velkou kontrolu nad vytvářenými daty, což je nezbytné pro systém, který vyžaduje určitou míru manuálních zásahů do výstupu klasifikátoru.

Kapitola 6

Vyhodnocení DeMoClas

Tato kapitola je věnována vyhodnocení systému DeMoClas z hlediska získaných dat, porovnání se slovy běžně používaného jazyka a možnostem zjednodušení mauální kontroly výsledků, která je v případě zkoumání jazyka nezbytná. Obrázek 6.1 ilustruje soubory, které do systému vstupují nebo jsou jím vytvářeny a skripty, které jsou jeho součástí. Systém je umístěn a spouštěn na serveru Minerva 1, který disponuje dvakrát procesorem Intel Xeon E5-2640 s frekvencí 2.50 GHz a 128 GB paměti RAM.



Obrázek 6.1: Přehled skriptů a souborů systému DeMoClas.

Prvním krokem systému je načtení dat. Tabulka 6.1 uvádí, z jakých zdrojů byla data načtena a celkový čas v sekundách, který zpracování dat zabralo. Dále je zde uveden celkový počet hnízd, který byl vytvořen v souboru `czech.nests` po přidání daného zdroje (odpovídá počtu modelů dostupných pro klasifikaci v souboru `czech.dms`), průměrný počet slov připadající na jedno hnízdo a počet dvojic, které byly ve zdrojovém souboru k dispozici.

Údaj průměrného počtu slov na hnízdo je z hlediska kvality hnízd důležitým ukazatelem. Během konstrukce systému a zkoumání vzniklých dat se ukázalo, že hnízda, která mají menší počet slov, mají i menší počet formantů, čímž se snižuje přesnost určeného slovotvorného kořene. Oproti tomu platí, že čím více slov je do hnízda zařazeno, tím lepší a přesnější je jeho model, což vede k přesnější klasifikaci. Můžeme vidět, jak postupné přidávání jednotlivých

Zdroj dat	Čas [s]	Počet hnízd	Počet slov	Průměrný počet slov na jedno hnízdo	Počet derivačních vazeb ve zdroji
cety	1.42	5 661	22 485	3,97	18 146
czcz	10.77	7 030	28 052	3,99	7 070
psjc	211.47	27 613	83 987	3,04	42 999
ssjc	454.16	37 496	111 875	2,98	34 456
ssc	522.65	38 947	116 319	2,98	4 628
tbdp	5829.29	85 147	346 676	4,07	332 108

Tabulka 6.1: Údaje ze zpracování zdrojových dat do hnízd pomocí skriptu `MakeDerivationModels.py`.

zdrojů v pořadí, které koresponduje s tabulkou, ovlivňovalo bohatost hnízd. Slovníky *cety* a *czcz* poskytují spíše širokou škálu derivací pro svá lemmata a díky tomu se utvářejí velká hnízda s bohatým seznamem formantů. Oproti tomu slovníky *ssjc* a *ssc*, které obsahují archaizmy, zeměpisné názvy, vlastní jména a odborné názvosloví, což jsou slova, která patří do spíše menších, velmi specifických hnízd, způsobují utváření velkého počtu menších hnízd. Jako výborný zdroj pro doplnění utvořených hnízd se prokázal soubor `czech.tbdp`, který zvedl průměrné množství slov na hnízdo o jedno.

Díky spojování vazeb a hnízd dohromady se často ztrácí sémantické rozdíly mezi slovy a klade se důraz na jejich čistou morfologickou podobnost. V tomto je klasifikátor DeMoClas podobný síti DeriNet, popisované v kapitole 2.4. Přidání sémantické roviny by zcela určitě přispělo k přesnosti klasifikace neznámých slov. I tento systém se potýká s problematikou homonymie a přidání informace o významu jednotlivých slov by přispělo k rozlišování mezi těmito slovy. Podle sémantického významu by pak musela být rozdělena i slovotvorná hnízda, čímž by se zvýšil jejich celkový počet a snížil průměrný počet slov na hnízdo. Pro klasifikátor by to mohlo znamenat zhoršení přesnosti derivačních modelů, neboť by se zmenšil počet určených slovotvorných formantů, avšak tato nevýhoda by byla kompenzována přidáním dalšího faktoru, podporujícího rozhodování. Aby rozšíření o sémantickou rovinu mělo pro systém skutečný přínos, bylo by potřeba, aby neznámá slova ke klasifikaci měla též určitý sémantický údaj. Toho je možné snadno docílit v případě klasifikace slovníkových lemmat, nikoliv však libovolných slov získaných např. z korpusů, které reflektují běžný jazyk.

Z hlediska teoretického členění derivačních vztahů podle Miloše Dokulila systém obsahuje všechny tři typy posunů mezi onomaziologickými kategoriemi – transpozici, modifikaci i mutaci, většinou v jednom hnízdě. Variabilitu derivačního procesu a bohatost českého jazyka lze zobrazit na množství slov, která mohou být spojena s jedním kořenem. Příklady sémanticky bohatých hnízd uvádím pro zajímavost ve výpisu 6.1.

```
vrch
['vrchní', 'vrchnost', 'vrchovatý', 'vrchovina', 'svrchní', 'povrch',
'povrchový', 'povrchní', 'vrchlík', 'vrchol', 'vrcholek', 'vrcholný',
'vrcholový', 'vrcholit', 'vyvrcholit']

výška
['výškový', 'výškař', 'výšina', 'navýšit', 'povýšit', 'povýšený',
'převýšit', 'vyvýšit', 'zvýšit', 'povýšeně', 'převýšení', 'povýšiti',
'povýšení', 'převýšiti', 'výšinka', 'výšinový', 'výšinský', 'výškařka',
'výškově', 'povýšenost', 'výšinný', 'zvýšit se', 'navýšení', 'vyvýšení',
'zvýšení', 'navýšivší', 'povýšivší', 'převýšivší', 'vyvýšivší', 'zvýšivší'
'povyšující', 'převyšující', 'zvyšující', 'navýšený', 'převýšený',
'vyvýšený', 'zvýšený', 'navýšen', 'povýšen', 'převýšen', 'vyvýšen',
'zvýšen', 'výškařův', 'výškařčin']

daleký
['dálka', 'dálkový', 'dálný', 'dálnice', 'dálava', 'oddálit', 'vzdálit',
'vzdálenost', 'opodál', 'zpovzdálí', 'dalekohled', 'dálnopis', 'daleko',
'dálkově', 'dálniční', 'oddalovat', 'vzdalovat', 'dalece',
'dalekohledový', 'opodálí', 'vzdálení', 'vzdáliti', 'vzdálenostní',
'vzdálený', 'vzdálí', 'dálnopisovat', 'dálnopisný', 'zpovzdáli',
'zpozdáli', 'zpozdálí', 'dálkař', 'dálkařka', 'dálno', 'vzdáleno',
'dálnopisování', 'oddalování', 'oddálení', 'vzdalování', 'oddálivší',
'vzdálivší', 'dálnopisující', 'oddalující', 'vzdalující', 'dálnopisovaný',
'oddalovaný', 'vzdalovaný', 'oddálený', 'dálnopisován', 'oddalován',
'vzdalován', 'vzdálen', 'oddálen', 'dálkařův', 'dálkařčin', 'dalekohledný'
'dálnopisový', 'dálnopisně', 'dálně']
```

Výpis 6.1: Ukázka variability některých kořenů.

6.1 Vyhodnocení klasifikace reálných dat

Pro vyhodnocení úspěšnosti klasifikátoru byl použit korpus, který Výzkumné skupině znalostních technologií poskytla v roce 2017 firma Seznam.cz. Z tohoto neotagovaného korpusu, obsahujícího slova používaná v každodenním jazyce, bylo vytřízeno 451 827 slov, které byly následně dány ke klasifikaci. Z výstupu klasifikátoru v souboru `democlas.update` bylo vybráno prvních 100 slov, kde klasifikátor navrhnul zařazení do jednoho z hnízd. Ze statistiky byly vynechány pokusy o zařazení zájmen, předložek či spojek, které korpus obsahoval a není možné je jednoduše vytřídit. V tomto reprezentativním vzorku klasifikátor dosáhnul úspěšnosti 69%, což znamená, že pro 31 slov ze sta by muselo být ručně upraveno číslo hnízda, kam byly navrženy k zařazení.

Klasifikátor s minimální chybovostí zařazuje typicky česká slova s délkou větší než tři písmena, zeměpisné názvy či odborné výrazy, které se již v nějaké formě v hnízdech vyskytují. Tato skutečnost vychází z povahy zdrojových dat získaných ze slovníků a ukazuje na důležitost rozmanitosti a množství zdrojových dat pro vytváření co nejbohatších hnízd z různých oblastí jazyka, což se pak promítá i do úspěšnosti klasifikace.

Zcela neúspěšný je klasifikátor v případě přejatých slov, zejména z angličtiny, která není schopen správně zařadit, ani rozpoznat, že hnízdo, které se jeví jako nejpodobnější se odvíjí

od zcela jiného kořene (např. slova *tablet*, *server*, *shop* atd.). Z obecného pohledu by bylo možné říct, že se zde střetávají báze a formanty původem z českého jazyka s úplně odlišnou sadou těchto lingvistických znaků, pocházejících z jazyka anglického. Do této skupiny platí i některé odborné výrazy jako slova *afázie* nebo *afix*. Řešením této situace je obohatit zdroje klasifikátoru o slovník přejatých slov, případně odborných výrazů, a vytvořit základní hnízda z připravených dat. Druhou možností je řadit tato slova do seznamu jedináčků a vyčkat, až se vytvoří požadované hnízdo s příchodem dalších, podobných dat.

Další problematickou skupinou jsou slova česká, ale velmi krátká (nejčastěji o délce tří nebo čtyř písmen), která je možné umístit do velkého množství hnízd. Tento problém vystihuje například dvojice *osa* – *osada*. Bez dalších údajů, zejména sémantického typu, není možné automatizovaně rozhodnout, zdali slovo *osada* není odvozeno od slova *osa*. Tento fakt spolu s homonymií a polysémií v jazyce ukazuje, že k jazyku není možné přistupovat pouze z jednoho pohledu, nýbrž se jedná o složitý systém, který je nutno uchopit ze širší perspektivy.

Zajímavostí může být zařazování zeměpisných názvů či jmen a výběr kandidátních hnízd, kam by slovo mohlo být zařazeno. Historicky mnoho názvů, zejména měst, vychází z jiných slov běžného jazyka a snaha klasifikátoru o zařazení takového názvu minimálně vypovídá o původu názvů v českém jazyce. Takovým příkladem mohou být např. *Hradec Králové* (klasifikátor se setkal se souvisejícím *královéhradecký*), *Jablonec*, *Havířov* či *Tábor*.

Celková statistika výstupu klasifikátoru je uvedena v tabulce 6.2. Podle rozložení výstupu můžeme vidět, že v systému je přes 77% slov, která jsou i v korpusu, což ukazuje na rozsah, ale i aktuálnost slovníků vzhledem k běžně používanému jazyku v 21. století. Zhruba 22% dat poté utváří slova, která klasifikátoru známá nejsou. Jedná se často o slova přejatá, cizí nebo na okraji spisovného jazyka. Z těchto slov se 70% snaží klasifikátor zařadit a navrhuje hnízdo, které považuje za nejpravděpodobnější. Zbývajících 30% rovnou zařazuje do seznamu jedináčků, čímž vyjadřuje, že pro toto slovo neexistuje aktuálně v systému hnízdo, které by mohlo být považováno za odpovídající.

Zdroj	Klasifikovaná slova [%]	Jedináčci [%]	Slova již zařazená [%]
korpus	15,99	6,59	77,48
cety	8,25	4,85	86,47
ssc	13,67	6,9	79,41

Tabulka 6.2: Výsledek klasifikace slov korpusu v porovnání se slovníky.

Testování nově vzniklých dat

Problematickou částí těchto dat je, že neexistuje automatický způsob, který by spolehlivě ověřil jejich správnost, pokud se jedná o nově klasifikovaná, neznámá slova. Jediným spolehlivým arbitrem je lidský mozek, který tuto dovednost vstřebává již od narození. Systém tedy vyžaduje velkou míru manuálních kontrol výsledků. Množství kontrol, objem kontrolovaných dat a celkovou náročnost takové kontroly lze různými způsoby snížit nebo alespoň zjednodušit.

Systém DeMoClas je implementován tak, aby umožnil uživateli kontrolu výsledků klasifikace dříve, než jsou tyto výsledky zaintegrovány do zdrojových dat klasifikátoru. To umožňuje výpis skriptu `DeMoClas.py` do souboru `democlas.update`, kde jsou uvedena rozhodnutí pro každé slovo zpracovávané klasifikátorem pro aktuálně vyhodnocený vstup. Díky tomu lze přehledně zkontrolovat výstupní data a upravit případné chyby. Teprve až uživatel bude spokojen s výstupem, spustí aktualizační nástroj `DeMoClasUpdater.py`, který vybraná data určená k aktualizaci začlení do zdroje.

Použití testovací sady a zpětné kontroly výstupu klasifikátoru s touto sadou nám může dát celkový přehled o úspěšnosti klasifikátoru. Tento výsledek je ovšem pouze orientační a neodráží různorodost dat, s jakými se může klasifikátor setkat. Výsledek při použití testovacího vzorku dat nemusí být směrodatný a velice záleží na jeho skladbě. Důvod lze vypozorovat z vyhodnocení klasifikace slov z korpusu porovnávaného se slovníkem *cety* a *ssc*. V případě korpusu se jednalo o slova běžné mluvy a bylo klasifikováno větší množství slov s větší úspěšností než u slovníků. Oba vybrané slovníky se vyznačují tím, že obsahují velmi specifická slova z oblasti odborných názvů, archaizmů až slangových či hovorových výrazů. Zároveň mají vysoké pokrytí slovotvornými vazbami a díky tomu je velké množství obecnějších slov již zařazeno do hnízda bez klasifikace. To způsobuje, že pokud jsou ke klasifikaci použity jejich soubory typu `lntrf`, zpracovávají se slova, která neměla udanou žádnou derivační vazbu a většinou patří do zmíněných okrajových skupin jazyka. Úspěšnost klasifikátoru se v případě obou slovníků pohybovala kolem 50%. Samotné slovníky jsou tedy blízké vhodné testovací sadě dat, za předpokladu, že by jejich derivační vazby byly vynechány ze zdrojových dat pro tvorbu hnízd a data by byla spojením lemmat z více slovníků, neboť každý z nich je jiný a zaměřený na různé skupiny slov v českém jazyce.

Kapitola 7

Závěr

Cílem této práce bylo zkoumat derivační proces v českém jazyce. Rozebrala jsem oblast derivační morfologie z hlediska vnitřní i vnější lingvistiky a zabývala jsem se způsobem, jakým dochází k učení jazyka u člověka. Na základě studia mechanizmu, kterého lidský mozek využívá, jsem se v práci rozhodla pro přístup, odrážející popsaný biologický proces, tedy nesnažit se tvořit předem daná pravidla, ale nechat tyto zákonitosti vznikat ze slov, která jsou součástí jazyka. Tento cíl byl splněn implementací klasifikačního systému DeMoClas, popsaného v této práci.

Z dostupných zdrojových dat slovníků a souborů, přístupných v rámci morfologického analyzátoru Výzkumné skupiny znalostních technologií, byly získány derivační vazby. Tyto vztahy byly uspořádány do slovotvorných hnízd, která představují skupiny morfologicky příbuzných slov. Následně došlo k jejich zpracování rozborem na slovotvorné báze a formanty, které mapují změny vznikající během derivačního procesu mezi slovem fundujícím a slovy fundovanými. Z nich jsou tvořeny derivační modely, představující soubor vlastností odvozovacího procesu vlastních danému hnízdu. Vytvořením této struktury byla naplněna vize zmíněná v závěru mé bakalářské práce.

Vzniklá data slouží jako vstupní informace pro klasifikátor, který zpracovává pro systém neznámá slova a zařazuje je do existujících slovotvorných hnízd. Ta jsou díky novým datům aktualizována a jejich derivační modely se rozšiřují. Slova, která není možné zařadit jsou uložena pro pozdější zpracování. Díky opakovanému procesu získávání nových dat se systém kontinuálně vyvíjí a učí se z informací, se kterými se setkává. Jeho přesnost se zvyšuje s množstvím dat, která již zpracovával a díky novým slovům se rozšiřuje i za hranice znalostí svých původních zdrojů, ze kterých na počátku vycházel. Neboť dochází ke zpracování velkého množství dat, byla provedena optimalizace zrychlující vyhledávání zařazených slov pomocí rejstříku.

Na závěr byl systém vyhodnocen na reálných datech, pocházejících z korpusu firmy Seznam.cz, a bylo popsáno, jakým způsobem výstup klasifikátoru ulehčuje manuální kontroly výsledků, které jsou vzhledem k povaze zpracovávaných dat nezbytné. Popsanými způsoby systém rozšiřuje funkcionalitu i data projektu morfologického analyzátoru.

Výsledky práce odráží, jak komplexní a bohatý český jazyk je. Získané poznatky o předponách, příponách a hláskových alternacích mapují chování češtiny a na tato data může navazovat libovolný lingvistický výzkum morfologických vlastností procesu odvozování. Systém může být rozšiřován i z hlediska zakomponování sémantiky. Mezi největší přínosy této iniciativy by patřilo zpřesnění zařazování nových slov, možnost seskupovat do hnízd slova nejen morfologicky, ale také sémanticky související a uvádět vztahy mezi hnízdy. Výhody a případné nevýhody tohoto kroku byly rozebrány v kapitole věnující se vyhodnocení práce.

Literatura

[1] Adam, R.: *Morfologie*. Nakladatelství Karolinum, 2015, ISBN 978-80-246-2800-4, dostupné také z: https://www.cupress.cuni.cz/ink2_stat/dload.jsp?prezMat=94493.

[2] Adam, R.; Beneš, M.; Bozděchová, I.; aj.: *Úvodní jazykový seminář: výklad a cvičení*. Nakladatelství Karolinum, 2014, ISBN 978-80-246-2633-8, dostupné také z: https://www.cupress.cuni.cz/ink2_stat/dload.jsp?prezMat=53866.

[3] Bishop, D. V. M.: How does the brain learn language? Insights from the study of children with and without language impairment. *Developmental Medicine & Child Neurology*, ročník 42(2), 2000: s. 133–142, ISSN 0012-1622. Dostupné také z: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-8749.2000.tb00058.x.

[4] Dehaene-Lambertz, G.: The human infant brain: A neural architecture able to learn language. *Psychonomic Bulletin & Review*, ročník 24, 2017: s. 48–55, ISSN 1531-5320. Dostupné také z: https://doi.org/10.3758/s13423-016-1156-9.

[5] Faltusová, M.: *Derivační morfologie češtiny na základě rozsáhlých korpusových dat*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, 2017, Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

[6] Filipec, J.: Naše současná společnost, slovní zásoba a slovníky. *Naše řeč*, ročník 75, 1992: s. 1–11, ISSN 0027-8203. Dostupné také z: http://nase-rec.ujc.cas.cz/archiv.php?art=7045.

[7] Havránek, B.: Přehled vývoje českého jazyka s hlediska marxistického. *Naše řeč*, ročník 35, 1951: s. 81–93, ISSN 0027-8203. Dostupné také z: http://nase-rec.ujc.cas.cz/archiv.php?art=4236.

[8] LEDA: *HEURÉKA - univerzální encyklopedie - elektronická verze pro PC*. [Online; navštíveno 14.01.2019]. URL https://www.leda.cz/Titul-detailni-info.php?i=88

[9] Lehečková, H.: Neurolingvistika: předmět, metody a historie. *Slovo a slovesnost*, ročník 45, 1984: s. 154–157, ISSN 2571-0885. Dostupné také z: http://sas.ujc.cas.cz/archiv.php?lang=en&art=2960.

[10] Lingea: *Školní slovník současné češtiny*. [Online; navštíveno 14.01.2019]. URL https://www.lingea.cz/skolni-slovnik-soucasne-cestiny.html

[11] Osolsobě, K.: *KMEN*. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. [Online; navštíveno 20.05.2020]. URL https://www.czechency.org/slovnik/KMEN

[12] Pala, K.; Smerk, P.: Derivancze — Derivational Analyzer of Czech. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ročník 9302, 2015: s. 515–523, ISSN 1611-3349. Dostupné také z: https://link-springer-com.ezproxy.lib.vutbr.cz/chapter/10.1007%2F978-3-319-24033-6_58.

[13] Ústav pro jazyk český Akademie věd České republiky: *Akademický slovník cizích slov*. [Online; navštíveno 12.01.2019].
URL http://www.ujc.cas.cz/sd/publikace/knizni/lexikologie-odd-publikace/lexikologie-slovniky/akademicky-slovnik-cizich-slov.html

[14] Ústav pro jazyk český Akademie věd České republiky: *Neomat*. [Online; navštíveno 12.01.2019].
URL http://www.ujc.cas.cz/elektronicke-slovniky-a-zdroje/Neomat.html

[15] Ústav pro jazyk český Akademie věd České republiky: *Příruční slovník jazyka českého (1935–1957)*. [Online; navštíveno 12.01.2019].
URL http://www.ujc.cas.cz/elektronicke-slovniky-a-zdroje/Prirucni_slovik_jazyka_ceskeho.html

[16] Ústav pro jazyk český Akademie věd České republiky: *Slovník spisovného jazyka českého*. [Online; navštíveno 12.01.2019].
URL http://lexiko.ujc.cas.cz/texts/ssjc.html

[17] Ústav pro jazyk český Akademie věd České republiky: *Slovník spisovné češtiny pro školu a veřejnost*. [Online; navštíveno 12.01.2019].
URL http://www.ujc.cas.cz/sd/publikace/knizni/lexikologie-odd-publikace/lexikologie-slovniky/slovnik-spisovne-cestiny-pro-skolu-a-verejnost.html

[18] Rusínová, Z.: *SLOVOTVORNÝ ZÁKLAD*. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. [Online; navštíveno 20.05.2020].
URL https://www.czechency.org/slovnik/SLOVOTVORNÝ%20ZÁKLAD

[19] Wikislovník: *Wikislovník:Hlavní strana*. [Online; navštíveno 12.01.2019].
URL https://cs.wiktionary.org/wiki/Wikislovn%C3%ADk:Hlavn%C3%AD_strana

[20] Čermák, F.: *Jazyk a jazykověda*. Nakladatelství Karolinum, 2011, ISBN 978-80-246-2360-3.

[21] Černý, S.: *Analýza slovotvorných vazeb v češtině*. Diplomová práce, Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, 2010, Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

[22] Ševčíková, M.; Žabokrtský, Z.; Vidra, J.; aj.: Lexikální síť DeriNet: elektronický zdroj pro výzkum derivace v češtině. *Časopis Pro Moderni Filologii*, ročník 98(1), 2016: s. 62–76, ISSN 0008-7386. Dostupné také z: https://search.proquest.com/docview/2224459013/fulltextPDF/152D18AAAE5647A5PQ/1.