
**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**


# DERIVAČNÍ MORFOLOGIE ČEŠTINY NA ZÁKLADĚ ROZSÁHLÝCH KORPUSOVÝCH DAT

**DERIVATIONAL MORPHOLOGY OF CZECH ON LARGE CORPUS DATA**

**BAKALÁŘSKÁ PRÁCE**
**BACHELOR'S THESIS**


**AUTOR PRÁCE** — **MARIE FALTUSOVÁ**
**AUTHOR**

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



**BRNO 2017**

**Vysoké učení technické v Brně - Fakulta informačních technologií**

Ústav počítačové grafiky a multimédií Akademický rok 2016/2017

# Zadání bakalářské práce

Řešitel: **Faltusová Marie**
Obor: Informační technologie
Téma: **Derivační morfologie češtiny na základě rozsáhlých korpusových dat**
**Derivational Morphology of Czech on Large Corpus Data**
Kategorie: Umělá inteligence

Pokyny:

1. Seznamte se s existujícími nástroji pro práci s derivativní morfologií a se zpracováním českého textu na morfologické a syntaktické úrovni.
2. Navrhněte a realizujte rozšíření stávajícího morfologického analyzátoru češtiny, který pokryje derivace velké části slov objevujících se v korpusech.
3. Implementujte podporu generování slovotvorných vazeb mezi základními tvary.
4. Vyhodnoťte rychlost a paměťové nároky aplikace.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle doporučení vedoucího

Pro udělení zápočtu za první semestr je požadováno:

- Prototyp řešení

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Smrž Pavel, doc. RNDr., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 17. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

doc. Dr. Ing. Jan Černocký
*vedoucí ústavu*

# Abstrakt

Tématem této práce je zkoumání slovotvorby v českém jazyce. Hlavním cílem je vytvořit modul získávající derivace z dat elektronického Slovníku spisovné češtiny. Tato problematika byla vyřešena sestrojením tříúrovňového zpracování vycházejícího z dat slovníku. První úrovní je získání derivací z definic lemmat, druhým krokem je seskupení základních tvarů podle jejich podobností a třetí fází je ohodnocení získaných derivačních dvojic značkou derivační třídy, do které spadají. Zpracováním se podařilo získat více než 4 500 nových slov a ohodnotit nad 20 000 derivačních vazeb. Modul se stal plnohodnotnou součástí Morfologického analyzátoru Výzkumné skupiny znalostních technologií, působící na Fakultě informačních technologií Vysokého učení technického v Brně.



# Abstract

Subject of this thesis is study of word formation in the Czech language. The main aim is to create a module acquiring derivations from data of the electronic Dictionary of the Czech Language. This problematics has been solved by constructing three-level processing based on dictionary data. The first level is to obtain derivations from lemma definitions, the second step is making groups of basic forms according to their similarities, and the third stage is the evaluation of derivation pairs by number tag of derivation class to which they belong. I have managed to get more than 4 500 new words and evaluate over 20 000 derivative couples. The module has become a full-fledged part of the Morphological Analyzer of the Knowledge Technology Research Group, working at the Faculty of Information Technology of the Brno University of Technology.



# Klíčová slova

český jazyk, morfologie, morfém, slovotvorba, derivace, sufixace, vnitřní lingvistika, vnější lingvistika, psycholingvistika, sociolingvistika, Slovník spisovné češtiny, derivační třídy, zahnízďování

# Keywords

czech language, morphology, morpheme, word-formation, derivation, suffixation, internal linguistics, external linguistics, psycholinguistics, sociolinguistics, Dictionary of czech, derivative classes, nesting

# Citace

FALTUSOVÁ, Marie. *Derivační morfologie češtiny na základě rozsáhlých korpusových dat.* Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

# Derivační morfologie češtiny na základě rozsáhlých korpusových dat

## Prohlášení

Prohlašuji, že jsem tuto bakalářskou 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
17. května 2017


## Poděkování

Na tomto místě bych chtěla poděkovat doc. RNDr. Pavlovi Smržovi, Ph.D. za vedení práce, odborné rady z oblasti lingvistiky a za poskytnutí zdrojů dat, se kterými jsem mohla pracovat.

# Obsah

# Kapitola 1

# Úvod

Čeština jako náš mateřský jazyk nás provází od narození. Obecně patří mezi jeden z nejsložitějších jazyků na světě díky své obsáhlé gramatice a rozsáhlé slovní zásobě. Jazyk je živá struktura, která se pod vlivem mnoha faktorů neustále mění. Zkoumání pravidel vzniku nových slov z původních zahrnuje nejen otázky z pole gramatiky, ale také je třeba se ptát, jak vznikla tato struktura jazyka, co nutí člověka vytvářet nové tvary a jak zpracovává jazyk náš mozek, který je vlastně prvotní příčinou vzniku i změny jazyka.

Světoví lingvisté si tyto otázky kladou a zkoumají různé jazyky v kontextu všech jazykovědných disciplín. Nejprobádanějším jazykem z tohoto pohledu je jednoznačně angličtina, kdežto češtině se pozornosti na světové úrovni příliš nedostává. I přes to čeští lingvisté pracují na co nejobsáhlejším zmapování českého jazyka. Dnešní technologický pokrok umožňuje zpracování jazyka pomocí počítačů, což otevírá nové možnosti na poli jazykovědy. Velmi významným příspěvkem k rozsáhlým výzkumům je rozvoj korpusů, které poskytují obsáhlé soubory jazykových dat.

Mým záměrem pro tuto práci bylo prozkoumat jazyk a slovotvorbu z co nejširšího úhlu pohledu, který lingvistika nabízí, jelikož je to jedna z disciplín, kde se velmi silně projevuje vliv člověka. Zajímá mě výzkum pracující s českým jazykem a výsledky či poznání, která může přinést. Český jazyk vnímám jako součást své identity, protože mateřský jazyk si nevybíráme, učíme se jej podle místa, kde se narodíme, a vždy ho budeme mít v sobě uložený nehledě na to, jestli ho v životě budeme nebo nebudeme používat.

Práce je tvořena pro Výzkumnou skupinu znalostních technologií působící na Fakultě informačních technologií Vysokého učení technického v Brně. Jedním z projektů této skupiny je Morfologický analyzátor, který pracuje se slovníky češtiny v elektronické podobě. Nad nimi se provádí různé typy výzkumného zpracování, avšak chybí část, která by se zabývala zpracováním slovotvorby. Proto je hlavním cílem práce rozšířit Morfologický analyzátor o část, která by se tímto tématem zabývala. Kroky, které k tomuto cíli vedly, jsou studium problematiky, využití potenciálu dat, která byla pro tuto práci k dispozici, a zejména integrace nového rozšíření do původního zpracování tak, aby bylo nejen přenositelné mezi různými slovníky, ale zároveň rozšiřitelné o další části budoucího výzkumu. Pro použití co nejaktuálnějších dat a zkoumání živého jazyka dnešní doby je použitá slovní zásoba ověřována českým korpusem vytvořeným firmou Seznam.cz.

V první kapitole je stručný náhled do disciplín lingvistiky, které se slovotvorby dotýkají. Různé úhly pohledu vnitřní i vnější lingvistiky vysvětlují, proč je zkoumání pravidel slovotvorby pro češtinu netriviální. V další kapitole je popsána struktura použitých slovníkových dat a výsledky jejich ověřování korpusem. Čtvrtá kapitola se zabývá dosavadním způsobem zpracování použitého slovníku, popisem jednotlivých fází tohoto procesu a samotným

návrhem jeho rozšíření. Následuje popis způsobu implementace popsaného návrhu a vysvětlení řešení vzniklých složitostí. Celá práce končí testováním vzniklých souborů, uvedením zajímavých statistik a vyhodnocením časové a paměťové náročnosti vůči ostatním skriptům. V přílohách jsou uvedeny vybrané zajímavé grafy zpracování, porovnávací statistiky a ilustrační výpisy z vybraných souborů, důležitých pro tuto práci.

# Kapitola 2

# Derivační morfologie v českém jazyce

Člověk, jakožto sociální bytost, si již od narození osvojuje mateřský jazyk, který mu celý život slouží ke komunikaci s ostatními. Používá jej k popisu vjemů, které jeho mozek mentálně zpracovává. Tyto subjektivní vjemy potom vybírá ze svého vědomí a pomocí slov svou představu popisuje. Slova se uspořádávají do výpovědí a výpovědi tímto způsobem verbalizují lidské mentální operace [1].

Ve flektivních jazycích, kam patří i čeština, se slovo realizuje ve větě prostřednictvím jednoho ze svých tvarů. Jazyk je pozměňováním slov využíván tak, aby se vyjádřil obecný význam nebo vztahy mezi jednotlivými slovy. V těchto jazycích je dobře vyvinutá diferenciace mezi morfém:slovo a slovo:věta. Slovo je diferencovaná gramatická jednotka, u které se rozlišuje případná předpona, kořen a koncovka, která v tomto typu jazyka plní sémantickou i syntaktickou funkci [6].

Všechny aspekty jazyka formálně popisuje **lingvistika**, která se obecně dělí na vnitřní a vnější. Vnitřní lingvistika zkoumá jazyk z hlediska jeho obecných principů, spadá sem morfologie, gramatika a fonetika. Vnější lingvistika se zabývá různými vlivy působícími na jazyk, jako je společnost, která jej používá, nebo mentální zpracování jazyka mozkem člověka. Všechny tyto složky působí na formování, neustálou proměnu a obohacování přirozeného jazyka, který každodenně používáme. Pro snadnou orientaci v následujících kapitolách jsem vytvořila graf dále zmiňovaných disciplín, uspořádaných podle struktury lingvistiky viz Příloha A [16].

## 2.1 Morfologie a morfematika

Morfologie, nebo-li tvarosloví, je spolu se syntaxí jednou z částí gramatiky. Zabývá se vztahy mezi jazykovými znaky či formami, pravidly a způsoby spojování těchto jednotek a vytvářením jednotek vyšších. Dělí se na dvě hlavní části – slovotvorbu (tvoření nových slov) a flektivní morfologii (skloňování a časování slov) [14]. Předmětem této práce je zkoumání slovotvorby a vlivů, které na ni působí, proto se dále v textu zaměřuji zejména na tuto část morfologie.

Díky jasné dělitelnosti slova v češtině hrají pojmy **morfém a morf** důležitou roli. Morfém je nejmenší významonosná jednotka v jazyce. Má stavební funkci, což znamená, že tvoří vyšší významové jednotky – slova a jejich tvary. Morfém je abstraktní jednotka, definovaná na základě významu. Je to množina morfů se stejným lexikálním, gramatickým

i funkčním významem, které jsou si zároveň velmi nápadně podobné svou formální stavbou ze zvukového i grafického hlediska. Morfém je pak ve slovech realizován prostřednictvím jednoho ze svých morfů. Morf je tedy nejmenší nedělitelná jednotka realizovaná v textu, kterou lze vyčlenit na základě zopakování sama sebe minimálně ve dvou slovech [13]. Morfy tvoří základní morfologickou jednotku – slovní tvar, který je syntagma morfů realizované v komunikaci [1]. Morfémy českého jazyka jsou mnohofunkční, mohou zároveň podávat více informací o daném slově (např. o pádu, číslu, rodu nebo životnosti). V takovém případě se nazývají alomorfy [6].

Analýza slova na morfémy, stavební prvky slov jako jsou kořeny, přípony a předpony, se nazývá **morfematika**. Slovotvorba a morfematika jsou disciplíny, které jsou navzájem úzce spjaté, jelikož se obě zabývají zkoumáním morfémů. I když se slovotvorba zabývá především morfémy slovotvornými či kmenotvornými, tak stále dochází k detailnímu zkoumání na morfémové úrovni a proces tvoření slov je tedy vyloženě morfologický. Z tohoto důvodu se slovotvorba také nazývá derivační morfologií [13].

Při skládání morfémů (flektivních i derivačních) dochází někdy k obměně fonémové skladby celého tvarotvorného základu – morfonologické alternaci. Skupina fonémů nebo fonémových skupin, které se alternace v rámci jednoho alomorfu účastní, se nazývají morfonémy (např. v kořeni *ruka* se objevují hned tři alternace: *ru**k**a, ru**č**ní, ru**c**e*).

Soubor tvarů každého slova označujeme jako **paradigma** daného slova. To je ovlivněno ohebností/neohebností nebo případnou nesklonností slova, homomorfií (jeden morfém zastává více funkcí) anebo polymorfií (více morfémů má stejný význam – vznik dublet či triplet slov) [1].

## 2.2 Derivace jako slovotvorný mechanizmus

Slovotvorba zkoumá tvoření nových slov ze slov již existujících. Pro vytvoření nového slova mohou být použity tři způsoby: odvozování, skládání (kompozice) nebo okrajové zkracování [14]. Odvozování, neboli **derivace**, je považována za nejpoužívanější slovotvorný mechanizmus v češtině. Derivace jako taková se dělí na prefixální (odvozování předponami), sufixální (odvozování příponami) a konverzní (skládání s jinými základy), případně jejich kombinace. Pro slovotvorbu v češtině má ale zásadní význam především sufixace, která se mimo jiné účastní odvozování jmenných slovních druhů.

**Sufix** je část slova, která je umístěna přímo za kořen slova nebo za jiný sufix determinující kořen. Nachází se za odvozovacím kmenem a je součástí kmene slova odvozeného. Při derivaci, přidáním sufixu za kmen, vzniká buď nové slovo anebo kmen. Podle účinku sufixu na odvozené slovo rozlišujeme[17] sufix kmenotvorný (tvořící nový kmen odvozeného slova, např. *uč-i-tel*), slovotvorný (tvoří neutrální tvar slov, např. *uči-tel*), tvarotvorný (tvoří tvar slova, např. *učitel-e*), charakteristiku (vytváří slovo jiného slovního druhu, např. *blah-ý -> blah-o*) a konektém (sufix umístěný mezi dvěma kořeny u slov vzniklých kompozicí, např. *les-o-step*).

**Koncovka** je část slovního tvaru, která určuje osobu, pád, číslo, popř. také rod daného slova a někdy může určovat i jeho slovní druh. Koncovka je považována za součást sufixu, u některých slov může být sufix pouze koncovka (např. *zdrav-í*) a někdy může být i nulový.

Proces slovotvorby vždy obsahuje dvě slova – vytvořené slovo a slovo základové, původní. Pokud tato dvě slova porovnáme, dostaneme **slovotvorný základ**. Může jím být celé slovo základové, část tohoto slova nebo nejčastěji kmen základového slova, což je typické právě pro sufixaci. Pokud je tento kmen komplexní – obsahuje kořen, slovotvornou a nebo kmenotvornou příponu (sufix), může být slovotvorným základem samotný kořen. Následující

příklad popisuje takovou situaci: mějme slova *sladkost*, *sladce* a *sladit*. Slovo *sladkost* od slova *sladký* má komplexní kmen – obsahuje kořen *slad-*, sufix *-k-* a koncovku typickou pro podstatná jména vlastnosti *-ost*. *Sladce* se tvoří z vlastního kmene *slad-k-*, ale sloveso *sladit* už je odvozeno od kořene *slad-* a přidává pouze koncovku pro slovesa *-it*.

Přidání slovotvorné přípony za základ slova, a pak až navázání koncovky na konec kmene je při tvorbě nových slov velmi častý postup. Ne vždy tomu tak ale musí být, což je speciální případ derivace, která se nazývá **konverze** neboli koncovkové odvozování. Při konverzi dochází k odvození nového slova pouze za použití koncovky, slovotvorná přípona je vynechána (např. *dobr-o -> dobr-á, kmotr -> kmotr-a* atd.). Někdy se konverze považuje za určitý typ sufixace, při které se použije nulový sufix. Avšak na rozdíl od sufixace, při konverzi v češtině málokdy dochází ke změně slovního druhu. Konverze je tedy převedením základového slova do jiného paradigmatu v rámci stejného slovního druhu, proto k ní dochází zejména u neohebných slovních druhů.

Problematickou částí při slovotvorné analýze je jev, který se nazývá **morfémový uzel**. Je to hláska, která obsáhla dvě hlásky, k jejichž sloučení došlo z důvodu zjednodušení slova. Povaha tohoto zjednodušování je převážně fonetická. Většinou se jedná o hlásky foneticky velmi podobné, ale může jít i o hlásky stejné. Dochází tím k deformaci hranic mezi základem a sufixem foneticky i graficky. Nejčastěji se morfémový uzel nachází u adjektiv končících na příponu *-ský* nebo *-cký*, ale není na ně omezen (např. *čes-ský* se redukuje na *český*, *ob-vléct* se redukuje na *obléct* atd.). U některých případů ale došlo k redukci pouze na fonetické úrovni, nikoliv na grafické (např. *Francouz - francouzský* se písemně neredukuje) [13].

## 2.3 Vnější lingvistika ve vztahu ke slovotvorbě

Z pohledu vnější lingvistiky lze říct, že jazyk ovlivňují dvě složky – vnitřní a vnější. Vnější vliv je působení společnosti, etnika, ve kterém žijeme a vnitřní vliv vytváří náš mozek, způsob jakým myslí, používá a obměňuje přirozený jazyk. Všechny tyto faktory výsledně působí na formování jazyka, vyvíjení nových tvarů slov a na syntax jazyka [16].

### Vnější vlivy

Jazyk je ve společnosti centrální, společnost bez něj nemůže existovat stejně tak, jako by jazyk zanikl bez společnosti. Při vývoji každého člověka je potřebný k socializaci a integraci do společnosti. Je nástrojem pro předávání vědomostí, zkušeností, zákonů i domněnek. Jazyk je poznávacím znakem každého etnika, utvářený a formovaný společenským konsenzem, zvyklostmi a tradicemi. Působí na něj pravidlo tzv. většinové volby, což znamená, že lidé radši přejímají jazyk zažitý, vzniklý v minulosti, a tím se většinou eliminují autoritativní vlivy jednotlivce.

Vztahy mezi společností a jazykem se zabývá **sociolingvistika**. Zkoumá jazyková společenství v rámci národů či skupin, variety jazyka (spisovný jazyk, dialekty, interdialekty, žargon atd.), sociologické faktory povolání nebo postavení ve společnosti a vlastní komunikativní situaci, za jakých podmínek a účelem lidé komunikují. Obecným vlivem situace a povahy obou účastníků na vytváření variací jazyka se zabývá **stylistika**. Umění změny stylu komunikace je součástí našich společenských dovedností. Používáme při tom různé úrovně vyjádření na škále od automatických, zažitých výrazů až po aktuální, inovativní způsob vyjádření. Často hraje roli také exaktnost a ekonomie výrazu, které pak ovlivní naši volbu na této škále. Další faktor, který působí na tvarování jazyka je interakční sociolingvistika, tedy působení našich sociálních rolí na použití vhodných slov. Zejména v evropských

jazycích je pak rozšířená honorifika – vývoj zdvořilostních frází používaných k různým způsobům vyjadřování zdvořilosti.

Do jazyka společnosti také zasahuje tzv. jazyková politika, což je státní propagace určité formy jazyka, která se historicky mění. V dnešní době jsou oblasti, kde státní hranice téměř zmizely a naopak existují místa, kde jsou jasně zřetelné. Docházelo i k posunu státních hranic nebo k jejich úplnému přeskupení, proto je jedním z předmětů zkoumání lingvistů vztah jazyka a geografie. Z hlediska globalizace také na jazyk působí bilingvizmus či multilingvizmus, tedy současné používání dvou nebo více jazyků a jejich vzájemné míšení [16].

Díky působení společenských vlivů se jazyk specifickým způsobem **historicky** vyvíjel. Mluvnická stavba jazyka se vyvíjela po staletí a stejným tempem se i mění. Naproti tomu slovní zásoba přirozeného jazyka je velmi citlivá na změny ve společnosti a neustále se vyvíjí v závislosti na vývoji společnosti. Proto je třeba při zkoumání jazyka zapojit i historickou složku. Jazyky se vyvíjely od kmenových až po národní, zásadní vliv na prvotní jazyky měl také vývoj písemnictví a později vynález knihtisku.

Doba vývoje českého jazyka jako jednoho ze slovanských jazyků se odhaduje na něco přes tisíc let, do doby Velké Moravy. V 9. století naše území jazykově ovlivnil příchod věrozvěstů Cyrila a Metoděje, kteří zavedli kulturní a bohoslužební jazyk staroslověnštinu. V ní se ale postupně začaly silně projevovat vlivy češtiny. Největší vývoj slovní zásoby češtiny, zvukové a lexikální složky je datován mezi 12. až 16. století, což pravděpodobně způsobil fakt, že se čeština stala jazykem feudálním a postupně i literárním. Čeština prošla neobvykle intenzivní fází hláskového vývoje. Vývoj básnictví poté rozšířil či zpřesnil významy jednotlivých slov a s hospodářským rozvojem a prosperitou se čeština dále silně rozvíjela (z původní staroslověnštiny zůstalo pouze náboženské názvosloví). Husitské hnutí češtinu rozšířilo mezi všechny společenské vrstvy. Od 14. století již byla čeština samostatným jazykem, kulturně a literárně jednotným, odlišující se pouze v nářečích. Jak vyplývá z dobové literatury, v této době byla čeština vyvinutější než němčina. V období humanizmu došlo k obohacení češtiny o latinské, románské a německé výrazy. Mluvnický základ češtiny se ale od svých počátků téměř nezměnil.

Tento slibný vývoj češtiny byl násilně ukončen po bitvě na Bílé hoře. Dochází k omezení češtiny na všech úrovních, zůstává živá pouze na venkově. Jazyk se obnoví až s národním obrozením a čeština se stává jazykem národním. S vývojem kapitalizmu v 19. století se zjednodušují některé příliš složité vazby z dob humanizmu a vytrácí se zastaralé výrazy. Ani světovými válkami, ani nástupem komunizmu už čeština nebyla ovlivněna, jelikož jako jazyk už byla zcela utvořena a dokončena. Čeština se stala jazykem celé národní společnosti, a i když se její slovní zásoba vlivem komunizmu pozměnila, její pozice jako jazyku je pevná [5]. Nová slova do češtiny přibývají i dnes, v době extrémního technologického rozvoje se kterým vznikají nové výrazy, a také vlivem sílící globalizace a promícháváním kultur se jazyk obohacuje a v budoucnosti stále obohacovat bude.

Nedílnou součástí společnosti je i **literatura**. Ta je specifická tím, že je lingvisticky neurčitá a neexplicitní. Dává prostor k uměleckému vyjádření, čím že zvyšuje míra interpretativnosti jednotlivých slov nebo obratů a metaforičnosti. Studium metafor uměleckých textů je stále v začátcích kvůli problematické automatické identifikaci v korpusech. U umělecké literatury se klade důraz na estetickou kvalitu textu a na rytmus, dochází ke zkracování nebo obměňování slov. Zejména poznamenaná jsou zájmena, vyjadřování osoby, pasiva nebo také způsob časování celého textu (např. minulý čas a narativní). Otevřenými otázkami také zůstává, zdali je umělecký text založený na jiných kognitivních procesech než běžný text nebo jestli jde o zcela jiný typ verbálního chování [16].

### Vnitřní vlivy

Psycholingvistika je lingvistický obor zabývající se spojitostí mezi vědomím, úžeji myšlením, a jazykem. Zkoumání z pohledu psycholingvistiky se liší podle toho, zdali zkoumáme jazyk v jeho plně rozvinuté formě, s vybudovaným, pevným systémem (u lidí aktivně používajících tento jazyk) nebo zkoumáme zpracování jazyka naším mozkem ve fázi vývojové (nabývání mateřského jazyka u dětí nebo studium cizího jazyka). Z hlediska proměnlivosti a vývoje jazyka je toto rozdělení ale relativní, neboť v obou stádiích dochází k neustálé obměně, jenom v jiné intenzitě.

Systém jazyka reprezentovaný v naší mysli má vybudované a funkční tyto části [16]:

- asociativně vzájemně provázaná jazyková paradigmata
- modely kombinací jednotlivých jazykových jednotek
- pravidla pro realizaci modelů
- síť vztahů jednotek (syntax)
- prostor pro kreativitu, tvořivost

Ať už jde o mateřský nebo cizí jazyk, proces zpracování jazyka má své části – produkci a percepci. K **produkci** jazyka člověk používá dva hlavní způsoby: psaný text a mluvený text. Psaný text je určen k přenosu prostorovému nebo časovému. Může být zdrojem pro zkoumání jazyka z hlediska historie a vývoje. Mluvený text je určený především k vyjádření svých myšlenek. Produkce mluvy je velice rychlá a přesná a zapojuje velké množství slov našeho mentálního slovníku. Pokaždé když volíme slova, náš mozek nabízí velký objem kandidátů, ze kterých poté jednoho vybere. Důkazem o této funkci mozku jsou přeřeknutí, která nastávají ve chvíli, kdy mozek nevybere z nabízených variant dostatečně rychle. Průměrný mluvčí vysloví za minutu asi 120 slov, přičemž pro každé z nich náš mozek nabízí další možné varianty, ale sémantickou chybu uděláme přibližně jednou za 1 milion slov.

Proces produkce začíná vyhledáním vhodných pojmenování pro danou myšlenku (konceptualizace), většinou ve tvaru lemmatu, poté se aplikuje flexe – uvedení slova do správného tvaru a zasazení slova do struktury (syntaktický faktor). Proces tedy zahrnuje přechod od základního tvaru ke gramaticky správné podobě a od dílčího významu k celkovému smyslu ve sdělení. Pojmenovávání obsahuje dvě psychické činnosti: klasifikaci, zařazování něčeho nového do jedné ze známých tříd a kvalifikaci, ohodnocení subjektivním pohledem na věc. Obě se poté promítají do projevu a do systémových distinkcí jazyka, což lze pozorovat zejména ve vztahu podstatné a přídavné jméno, vyjádření vztahu k pojmenovávané věci nebo třeba umístění v celkovém členění věty/projevu. Celkově proces konceptualizace je založen na pojmu, kterým se rozumí celkový úhrn jazykových znaků (morfém, slovo, spojení slov, věta). Pojmy jsou uložené v naší paměti, slouží jako podklad k vymezení dalších slov a spojení a ve své podstatě jsou velmi proměnlivé.

Proces **percepce** začíná samotným zaznamenáním sluchovými nebo zrakovými orgány, poté dochází k rozpoznávání slov a nakonec se provádí celková syntaktická analýza struktury sdělení. Studium produkce a percepce jazyka ukazuje, že rozpoznávání jednotlivých slov je založeno na modelech jazyka. Jde o porovnávání a odhad podobností mezi signály a jazykovou reprezentací. Člověk na přijímající straně komunikace po zaznamenání začátku slova začne automaticky prohledávat svůj mentální lexikon slov a vybere z něj všechna, která stejně začínají. Správné slovo je nakonec vybráno na základě korelace s významem

a s dalšími slovy, kdy na konci věty zůstanou mentálně vybraná ta slova, která zapadají do kontextu [16].

Častým předmětem psycholingvistického zkoumání je **morfologická dekompozice**, která se zabývá mentální reprezentací slova při vybavování. Na základě četných pokusů se obecně lingvisté přiklání k možnosti, že v mentálním lexikonu jsou zaznamenány kmeny slov a v případě potřeby je potom aplikováno morfologické pravidlo pro vytvoření požadovaného tvaru. Znamená to tedy, že slova se rozkládají na kmeny a afixy, které se samostatně vyhledávají, a poté spojují podle morfologických pravidel. Podle výzkumu[4] se liší zpracování předpon a přípon, avšak detailnímu zkoumání se podrobily pouze derivativní sufixy. Tato funkce mozku urychluje zpracování slov se stejnými příponami. Po dekompozici slova dochází k rozpoznání funkce přípony a mentálně je vyjasněn celkový význam slova [11].

Čeština obecně nepatří mezi psycholingvisticky prozkoumané jazyky, ač nabízí zajímavá témata pro zkoumání jakožto vysoce flektivní jazyk. U češtiny se nedá předpokládat tak markantní rozdíl mezi flektivní a derivační morfologií jako u jiných jazyků díky vysoké homonymii a synonymii mezi sufixy. Některé přípony jsou kvůli své nejednoznačnosti bez kontextu nepredikovatelné [2].

Závěrem lze tedy říci, že během učení a používání jazyka si člověk buduje určitý jazykový systém ve své paměti. Tento systém obsahuje morfémy, slova, slovní spojení nebo dokonce celé věty. Je nepostradatelný, ať už chceme jazykem něco vyjádřit a hledáme vhodná slova pro své myšlenky nebo rozpoznáváme, co nám sděluje někdo jiný. Vždy vycházíme z naučených základních slov nebo jejich částí, rozkládáme a skládáme je na kmen a morfémy podle morfologických pravidel, a poté určujeme či odvozujeme význam slova a nakonec celého sdělení. Ač čeština není příliš psycholingvisticky probádaným jazykem, stále se na ni aplikují výše zmíněné základní postupy naší mysli. Další výjimky, rozlišení mentální reprezentace předpon a přípon nebo derivace a flexe jsou volným polem pro budoucí výzkumy, aby potvrdily nebo vyvrátily některé teorie psycholingvistů, které bez zapojení slovanských a vysoce flektivních jazyků budou mít pouze dílčí povahu.

## 2.4 Existující nástroje pro morfologii češtiny

Tato kapitola se věnuje stručnému shrnutí vybraných nástrojů pro flektivní a derivační morfologii v češtině. Popisuje metody, jakými byla zpracována vstupní data a jejich přístup k morfologii. Zmíněny jsou také související nástroje a výzkumy, které je ovlivnily nebo z nich části těchto nástrojů přímo vycházejí.

### MorphoDiTa (Morphological Dictionary and Tagger)

MorphoDiTa je nástroj provádějící morfologickou analýzu s lemmatizací, značkování slovních druhů a rozpoznávání pojmenovaných entit v češtině. Je natrénován pro češtinu, ale autoři jej uvádějí jako rozšiřitelný i pro jakýkoliv jiný flexivní jazyk.

Čeština obsahuje velké množství koncovek, které se připojují ke kmeni slova a formují jej. Z tohoto důvodu je morfologický slovník analyzátoru navržen zejména pro práci s velkým objemem sufixů a rozpoznávání základních lemmat slov. Na základě těchto dat poté analyzátor pozoruje pravidelně se objevující vzorce kmen-sufix a automaticky je zařazuje do morfologických šablon.

MorphoDiTa zpracovává lemma jedno za druhým a rozřazuje je do stromové struktury dat trie, kde každý uzel je jeden znak slova. Vzniknou tak potomci, kteří sdílejí stejnou předponu, čímž lze odhadnout společného předka (předpona nebo kmen slova). Pokud je

určena společná předpona, všechny koncovky daného podstromu včetně jejich značek poté utvářejí šablonu. Každý strom tedy může obsahovat 1 a více šablon.

Morfologický slovník je uložen v binární podobě. Při dotazu vyhledá všechny možné dvojice lemma-značka, kde lemma odpovídá předponě (začátku) slova a značka jeho koncovce, a MorphoDiTa poté vytvoří správný tvar dotazovaného slova. Není tedy využita žádná lingvistická znalost, ale tento způsob zpracování umožňuje rychlou a efektivní analýzu, což je využíváno během značkování.

Značkovač slovních druhů tohoto nástroje přímo vychází z výzkumu nazvaného Morče, jehož výsledkem byly dva značkovače – Morče a Featurama. Pro každý tvar v textu morfologický slovník vyhledá skupinu možných kandidátů ve tvaru lemma-značka. Tyto dvojice jsou potom disambiguovány značkovačem, který je implementován jako algoritmus učící se s učitelem, perceptronem. Klasifikátory byly převzaty z původního značkovače Morče.

Poslední částí nástroje MorphoDiTa je NER – rozpoznávač pojmenovaných entit v češtině. Je implementací výše zmíněného výzkumu Morče a zakládá se na Markovových Modelech maximální entropie. Tento model předpovídá pro každé slovo ve větě distribuci jeho tříd a pozici. Poté se určí optimální kombinace tříd a délek pojmenovaných entit. Následná klasifikace se zakládá na morfologické analýze, dvoustupňové predikci a zařazování slov do tříd. Jako trénovací data byl použit CNEC (Czech Named Entity Corpus).

Autoři plánují budoucí rozšíření tohoto nástroje, zejména chtějí vytvořit guesser, který by dokázal analyzovat validní tvary slov, které nikdy před tím neviděl. Dále také plánují rozšířit nástroj MorphoDiTa na další jazyky a do zahraničí [12].

### MorfFlex CZ

MorfFlex CZ je morfologický seznam trojic, který je generován ze stromové struktury morfologického slovníku, původně vytvořeného Janem Hajičem pro nástroj MorphoDiTa (viz předchozí podkapitola). Trojice, které jsou obsažené v MorfFlex CZ jsou ve tvaru lemma-značka-tvar. V současné verzi obsahuje seznam více než 987 tisíc unikátních lemmat. Každý slovní tvar je označkován podle původního značkovacího schématu, použitého v MorphoDiTa.

Zjednodušení datové struktury na seznam vychází z toho, že každé z původních paradigmat je vlastně údaj o koncovce a značka. Lemma se pak skládá z kořene a paradigmatu. Konkatenací kořene a koncovky slova vzniká tvar slova a značka.

Díky tomuto zobrazení dat mohou být odlišena derivační paradigmata a lemmata, která pod takové paradigma spadají, mohou být použita pro odvození dalších slov. Veškeré derivační informace, včetně referencí na původní lemma, jsou uloženy v části sufixu – tedy příponě, která vytvořila nový tvar z původního [19].

### DeriNet

DeriNet je síť dat a údajů o derivační morfologii. Vztahy mezi slovem původním a odvozeným jsou zobrazeny v orientovaném grafu, jehož uzly reprezentují lemmata a hrany derivační kroky. DeriNet jako takový má již 5 verzí, během kterých došlo postupně k rozšíření sítě z podstatných jmen na ostatní slovní druhy a k optimalizaci. Nakonec byl spojen s nástrojem MorfFlex CZ, čímž došlo k masivnímu obohacení sítě (zhruba o 300% původního obsahu DeriNetu) o lemmata, která v ní do té doby chyběla. Rozvojem této sítě byla zaznamenána nová derivační pravidla.

Mezi podobné nástroje určené pro derivaci patří např. Deriv[8], což je nástroj pro automatické vyhledávání slovotvorných vztahů, založený na pravidlech aplikovaných regulárními

výrazy [19]. Za jeho nástupce lze považovat nástroj Derivancze, který je webovou aplikací[1]. Hledá lemma zadaného slova a na základě analýzy poté zobrazí jeho známé derivace a značky udávající další gramatické detaily. Názorně zobrazuje derivační vztahy mezi českými slovy a výsledky porovnává se sítí DeriNet [9].

### ajka

Ajka je český morfologický analyzátor zaměřující se na efektivní zpracování a uložení velkého objemu dat. Rozlišuje mezi flektivní a derivační morfologií a ke každé disciplíně přistupuje odlišně. Jako hlavní složku flektivní morfologie v češtině udává definici vzorů. Proto byl autory navržen rozsáhlý systém těchto vzorů a na základě pozorování byl také vytvořen systém setů koncovek.

Oproti tomu pro derivační morfologii byl sestrojen hierarchický systém morfologických paradigmat pro zachycení všech úrovní české derivační morfologie. Tyto vzory se poté konstruují vzájemným propojováním lemmat speciálním typem hrany. Autoři považují tento proces za n-ární relaci, která může být rozdělena na n - 1 binárních relací. Tento přístup umožňuje větší flexibilitu a generalizaci derivací a jejich vztahů.

Vzhledem k celkové paměťové náročnosti analyzátoru z důvodu uchovávání lexikálních dat, autoři také řešili otázku efektivního uložení a rychlého procházení datového prostoru. Řešením se stalo uspořádání dat do datové struktury trie ve formě konečného automatu. Takto uloženy byly dva základní soubory analyzátoru – definice setů koncovek a vzorů pro morfologii a binární obraz strojového slovníku češtiny obsahující kmeny a pomocná data. Pro čtení těchto souborů byl vytvořen program `abin`. Po načtení dat do paměti dochází k běhu algoritmu pro morfologickou analýzu na základě dělení a určí se jednotlivé gramatické kategorie. Analyzátor má také funkci pro zobrazení lemmat a jejich derivací.

Uživatel může mít více variant binárních souborů a počet rozpoznaných slov pak záleží na rozsahu a bohatosti použitého slovníku. Z původní databáze 223 600 kmenů dokáže analyzátor vytvořit 5 678 122 validních tvarů českých slov. Rychlost analyzátoru je 20 000 slov za sekundu na počítači s procesorem Pentium III s frekvencí 800MHz [10].

[1] Aplikace dostupná na: https://nlp.fi.muni.cz/projects/derivancze/index.cgi

# Kapitola 3

# Použitá data pro derivační morfologii

Tato práce je vytvářena pro Výzkumnou skupinu znalostních technologií (dále jen KNOT) působící na Ústavu počítačové grafiky a multimédií, Fakultě informačních technologií Vysokého učení technického v Brně. Hlavním cílem práce je rozšířit Morfologický analyzátor výzkumné skupiny KNOT o část, která by se zabývala derivační morfologií. Celý projekt je založen na slovníkových datech v elektronické podobě.

Derivační morfologie je disciplína, která je určována především slovní zásobou a odráží se v ní i její dynamická proměnlivost v čase tím, jak vznikají, mění se a případně i zanikají slova a jejich tvary. Aby byla práce co nejaktuálnější, je třeba získat reprezentativní data současné češtiny. Pro tento účel posloužil český korpus, poskytnutý v dubnu 2017 výzkumné skupině KNOT firmou Seznam.cz.

Korpusy jsou vnitřně strukturované, unifikované a ucelené soubory jazykových dat v elektronické podobě. Důvodů, proč mají pevnou pozici ve světových lingvistických výzkumech, je mnoho a patří mezi ně zejména velký rozsah dat, aktuálnost a přirozenost získaných slov (resp. textů) a v neposlední řadě také jejich elektronická forma umožňující zpracování počítačem [15].

Korpus jako takový má i své meze. Jednou z jeho vlastností je, že nepokryje všechna slova v jazyce, pouze frekventovaná, a slova ne tolik častá na okraji jazyka mohou být vynechána. S tím také souvisí, že korpus není schopen postihnout jevy potenciální. To by mohlo v některých typech výzkumů být výhodou, nikoliv však pro derivační morfologii. Je sice důležité pokrýt aktuální slova, avšak je třeba zahrnout i slova neobvyklá, a to vše včetně různých tvarů, kterých mohou nabývat. Korpusy jsou obrazem především situačně standardních textů, příliš se v nich nevyskytují specifická slova z momentů jako např. hádka, intimní konverzace anebo výraz překvapení [3].

Pro účely této práce jsem tedy zvolila přístup korpusové lingvistiky označovaný pojmem „corpus-based“, což je přístup, kdy je výzkum na korpusu založený, resp. korpusem ověřovaný [3]. Slova vyskytující se v korpusu firmy Seznam.cz jsou zpracována do frekvenčního slovníku, který je rozdělen na dvě části – na slova spisovné češtiny (zastoupení všech slovních druhů, různé tvary slov včetně velkých písmen) a na seznam slov nespisovných, názvů společností a zkratek. Patří sem výrazy jako např. „jj“, „com“, „Wikipedia“ nebo „DVD“, která pro derivační morfologii českého jazyka nemají žádnou výpovědní hodnotu, proto budu dále pracovat pouze s první částí, obsahující validní česká slova. Ta jsou zároveň

reprezentována ve Slovníku spisovné češtiny, který je součástí Morfologického analyzátoru skupiny KNOT.

Slovník spisovné češtiny je moderní výkladový slovník českého jazyka vytvořený Ústavem pro jazyk český Akademie věd České republiky. Jeho elektronickou podobu vydalo nakladatelství LEDA v roce 2005. Obsahuje hesla z okruhů spisovné češtiny, obecné češtiny, některé slangové výrazy, odborné názvy a objevují se v něm i neologizmy anebo zeměpisné názvy a národnosti [7].

Každé heslo obsahuje nejdříve část se značkami, informujícími o jeho gramatických detailech jako je: slovní druh, podoba slova v nominativu singuláru, příp. informace o jeho nesklonnosti, dále se mohou nacházet koncovky pro skloňování v některých dalších pádech a poté následuje výklad hesla. Tato definice je ve Slovníku spisovné češtiny obohacena o dublety, varianty slov, synonyma, antonyma nebo přímo v definici jsou obsažena slova odvozená. Právě toto rozšíření slovníku z něj činí cenný materiál pro zkoumání derivační morfologie v českém jazyce. Ukázku dat ze slovníku si lze prohlédnout v Příloze D.1.

Lemmata ze Slovníku spisovné češtiny jsem ověřila použitím porovnávacího skriptu `CorporaComp.py`, který srovnal základní tvary slov obsažených ve slovníku se spisovnou částí frekvenčního slovníku korpusu. Aby bylo možné tato data srovnávat, musí být v takovém složení, aby byla porovnatelná. Frekvenční slovník korpusu obsahuje velká písmena i u slov, kde se v základním tvaru nepíší (např. u slov, která byla v textu na začátku věty). Z toho důvodu jsem celé srovnání prováděla nezávisle na velikosti písmen. Výsledky tohoto srovnání (detail viz tabulka 3.1) přinesly následující poznatky: můžeme říct, že 95% lemmat ve Slovníku spisovné češtiny patří do současného, živého jazyka. Zbývajících 1 890 slov lze označit za slova ne tolik frekventovaná, která dotváří celkový obraz jazyka a obohacují tuto práci o další data.

| **Měřený aspekt** | **Výsledek** |
|---|---:|
| Počet slov ve frekvenčním slovníku korpusu | 54 904 540 slov |
| Počet slov ve slovníku bez derivací | 50 848 slov |
| Počet slov ve slovníku se získanými derivacemi | 53 025 slov |
| Počet slov korpusu pokrytých slovníkem | 230 152 slov |
| Počet slov pouze ve slovníku | 1 890 slov |
| Slova slovníku pokrytá korpusem | 95,103 % |

Tabulka 3.1: Výsledky srovnání slovníku s korpusem

Korpus nám tedy poskytl důkaz, že Slovník spisovné češtiny je stále aktuální a použitelný pro zkoumání derivační morfologie spisovné slovní zásoby současného jazyka. Zároveň srovnání dokázalo, že slovník obsahuje slova vyskytující se v korpusech a následně pak může pokrýt derivace většiny z nich.

# Kapitola 4

# Současné zpracování Slovníku spisovné češtiny

Slovník spisovné češtiny (dále jen SSČ) je jedním z deseti slovníků dostupných v rámci Morfologického analyzátoru. Každý ze slovníků je unikátní, obsahuje jiné informace a jinou strukturu dat, což znamená, že slovníky zpracovávají v první úrovni různé skripty. Slovník SSČ prochází několika fázemi zpracování pro různé výzkumné účely. V této kapitole bych proces zpracování slovníku ráda popsala. Jelikož se do něj zapojuje několik skriptů a vzniká velké množství různých souborů, vytvořila jsem pro lepší přehled diagram celého procesu (viz Příloha B).

Na zpracování slovníků se podíleli různí lidé a často nebyla pečlivě vedena dokumentace k těmto skriptům, což spolu s ostatními chybami způsobilo, že některé soubory nebyly zcela v pořádku. Abych mohla svou práci zakládat na co nejlepších datech, provedla jsem před zahájením samotné bakalářské práce opravy a přidala některé funkce do již vytvořeného kódu.

Původní data slovníku jsou uložena ve formátu *xml* v souboru nazvaném `ssc.xml.lines`. První zpracování těchto dat provádí skript `xml2lntrf.py`, jehož autorem je Štěpán Rydlo. Nejprve se provede získání slovníkových dat do textové podoby z formátu *xml*. V této fázi jsem přidala výpis textové podoby zdrojových dat bez tagů, čímž vznikl soubor `ssc.def`, obsahující vždy na prvním řádku lemma a na řádku následujícím jeho slovníkovou definici. Data v tomto formátu poté skript prochází řádek po řádku a rozpoznává slovní druh, informace o skloňování a čísle z definice uvedené slovníkem. Pokud z nějakého důvodu nebyl slovní druh rozpoznán, slovo skončí v chybovém souboru `ssc.undefined`.

V této fázi vznikají i další soubory obsahující získané předpony, přípony nebo víceslovné výrazy ze slovníku (konkrétně viz Příloha B). Ze slov, která prošla předchozím tříděním, vzniká nejdůležitější soubor této fáze zpracování – soubor typu *lntrf* (ilustrační výpis ze souboru viz Příloha D.2). Data v souboru `ssc.lntrf` jsou ve tvaru *lemma:note:tag*, kde *lemma* je slovníkové heslo, *note* je případná poznámka o zvratném zájmenu u sloves nebo o stylu slova a *tag* je značka, která nese informace o slovním druhu (k1,k2,k3...k0, příp. kA pro zkratky), rodu, čísle, pádu a o koncovce tvaru, který v tomto pádu vzniká. Detailní popis všech variant značek v souboru *lntrf* si lze prohlédnout v oficiální dokumentaci k projektu[1].

Soubor `ssc.lntrf` je již souborem, který má unifikovanou podobu, ať pochází z jakéhokoliv slovníku Morfologického analyzátoru, a proto je skript `lntrf2lpn.py`, který jej

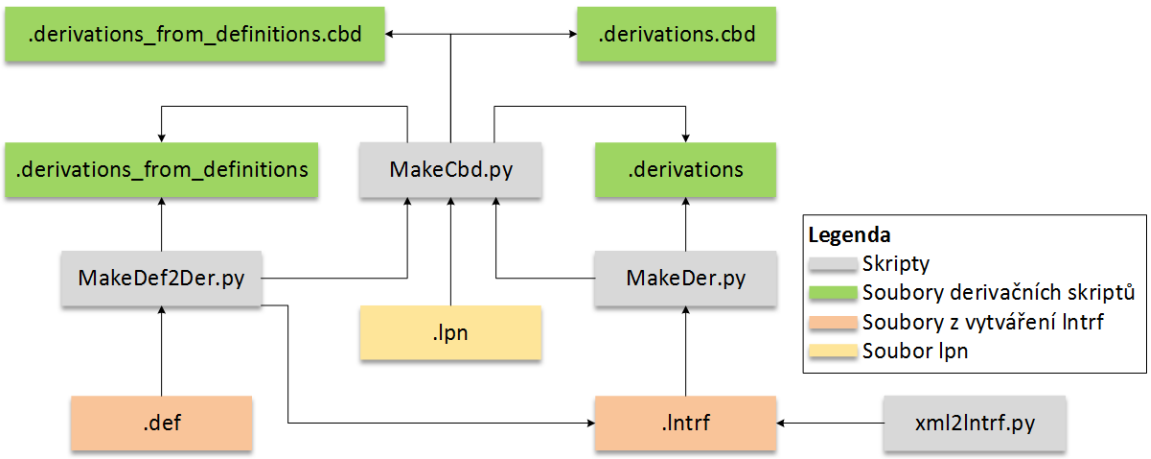
[1] http://knot.fit.vutbr.cz/wiki/index.php/Morfologick%C3%BD_slovn%C3%ADk_a_morfologick%C3%BD_analyz%C3%A1tor_pro_%C4%8De%C5%A1tinu

dále zpracovává, společný pro všechny slovníky. Jeho úkolem je nacházet pro každé slovo v souboru *lntrf* vzor ze seznamu paradigmat `czech.paradigms` a výsledkem je soubor `ssc.lpn` (ilustrační výpis ze souboru viz Příloha D.3). Formát dat v souborech typu *lpn* je *lemma#paradigm#note*, kde *lemma* je slovníkové heslo zpracované v souboru *lntrf*, *paradigm* je nalezený vzor a *note* je případná poznámka. Slova, která se nepodaří přiřadit k žádnému ze vzorů, jsou vypsána včetně chybové hlášky o vzniklém problému do souboru `ssc.guess_err`. Dále se provádí už jenom různá porovnávání ostatních souborů se soubory *lpn* nebo případě s `czech.paradigms` pro získání statistických údajů o slovech v rámci celého Morfologického analyzátoru.

Oblast derivační morfologie nad Morfologickým analyzátorem zpracovával také Ing. Stanislav Černý v rámci své diplomové práce nazvané Analýza slovotvorných vazeb v češtině [18]. V rámci této práce navrhl a vytvořil systém ohodnocující derivační vazby čtyřmístnou značkou na základě odvozujících pravidel. Výsledky jeho práce si lze prohlédnout v souborech typu *tbdp* v adresáři Morfologického analyzátoru.

## 4.1 Návrh Derivačního modulu

Na základě současného zpracování Slovníku spisovné češtiny jsem navrhla rozšíření Morfologického analyzátoru o skupinu skriptů, pracovně nazvanou jako „Derivační modul“. Jeho struktura je názorně zobrazena na Obrázku 4.1 a jeho celkové zasazení do struktury zpracování slovníku si lze prohlédnout v Příloze C.




Obrázek 4.1: Struktura Derivačního modulu

Prvním krokem celého Derivačního modulu je zpracování dat ze slovníku derivačním skriptem `MakeDef2Der.py`, který používá data v souboru `ssc.def` a prohledává definice lemmat. Jeho úkolem je nalézt co největší množství dalších lemmat uvedených v definicích autory slovníku. Takto nalezená slova se ve skriptu `xml2lntrf.py` předávají ke zpracování, vytváří se jim značka a začleňují se do souboru *lntrf*. Rozšiřují tak počet zpracovaných slov, předávaných k dalšímu zpracování do souboru typu *lpn*. Z tohoto procesu vzniká také samostatný soubor `ssc.derivations_from_definitions`, ve kterém jsou vypsána základní slova a jejich derivace v grafické podobě, připravené k dalšímu zpracování. Získávání dalších slov přímo ze slovníkových definic je efektivní způsob jak získat různorodou škálu tvarů a typů derivací či dublet.

V druhé fázi zpracovává skript `MakeDer.py` již obohacený soubor `ssc.lntrf` a vzájemně mezi sebou porovnává slova a seskupuje je do skupin tzv. zahnízďováním, což je metoda, kdy je v centru jedno slovo a kolem něj se seskupují další. Vznikají takto hnízda slov, která jsou si vzájemně podobná svým začátkem a poté se odlišují v sufixu, který určuje jejich tvar. Toto řešení má podobný princip přístupu, jakým řeší derivace nástroj MorphoDiTa (viz kapitola 2.4). Výsledkem tohoto procesu je poté soubor `ssc.derivations`, který obsahuje derivace vypsané v grafické podobě *slovo základové -> derivace*. Slovo, které se vypisuje jako základ pro celou skupinu, jsem se rozhodla vybrat tak, aby bylo nejkratší ze všech, jelikož právě toto slovo má nejblíže k největšímu společnému kmeni celého uskupení nebo je s ním shodné.

Další funkcí, kterou tento skript má, je rozkládat slova na kmen, popř. kořen, sufix a koncovku. Touto dekompozicí na morfémy získáváme v prvé řadě statistiku o kmenech a koncovkách. Ta, spolu s derivačními třídami, může v budoucnosti sloužit jako jako jeden z podkladů pro vytváření derivačních vzorů. V druhé řadě díky porovnávání sufixů můžeme odhalovat paronyma, která způsobují chyby ve vytvořených derivačních skupinách.

Paronyma jsou slova graficky a nebo foneticky podobná, avšak mají zcela odlišný význam. Problém způsobují speciální druhy paronym, kdy se slova většinou odlišují až v samotné koncovce, jako např. slova *sklep* a *sklepat* nebo *sterilace* a *sterilizace*. Jediným způsobem, jak od sebe taková slova odlišit, je dekompozicí na morfémy a srovnáním kmenotvorných sufixů. Konkrétně od slova *sterilace* pochází adjektivum *sterilační*, které bude rozděleno na *steril-ač-ní*. Pokud ho budeme porovnávat s adjektivem *sterilizační* od slova *sterilizovat*, rozkladem na *steril-izač-ní* porovnáním zjistíme, že se jedná o odlišná slova. Pokračováním v rozboru derivací slov bude zřejmé, že slova, která spolu souvisí, budou mít minimálně stejně začínající sufixy. Jelikož ale v češtině dochází k morfémovým uzlům a záměnám některých samohlásek, navrhla jsem systém tolerance vůči některým rozdílům. Pokud prozkoumáme další derivaci, slovo *steril-ova-t*, zjistíme, že písmeno *a* se mění na *o*, ale je nezaměnitelné s *i*, které je ve slovu *steril-izova-t*. Další tolerované rozdíly jsou např. mezi *i* a *y*, mezi písmeny bez diakritiky a s ní (*a* a *á*, *e*, *é* a *ě*, *c* a *č* atd.), ale také mezi změnou zápisu z *ď* a *dě* (a ostatní podobné dvojice).

Poslední částí zpracování je průchod dat modulem `MakeCbd.py`. Tento soubor není samostatným skriptem, jedná se o modul jazyka Python, který je volán z ostatních derivačních skriptů a aplikuje na vytvořená data příslušné funkce. Pro návrh řešení tímto způsobem jsem se rozhodla ze dvou důvodů. Prvním důvodem je, že potřebuji, aby funkce tohoto modulu byly aplikovány během zpracování dat derivačními skripty. Druhým důvodem je jeho snadná použitelnost na jakýkoliv slovník Morfologického analyzátoru. Data, která vytváří derivační skripty jsou již unifikovaná, mají jednotnou podobu nezávisle na slovníku, který zpracovávají, a proto není třeba vytvářet celý skript, stačí pouze přenášet funkce modulu `MakeCbd.py` mezi jednotlivými derivačními skripty.

Jeho prvním a základním úkolem je přidat k získaným datům z derivací informaci ve tvaru *class:odtržená_koncovka:nová_koncovka*, kde *class* je číselná značka derivační třídy, *odtržená_koncovka* je koncovka, která musí být odtržena z původního slova pro získání nového tvaru a *nová_koncovka* je koncovka, která po přidání k původnímu kmeni vytvoří nový tvar slova – derivaci. Tento údaj je poté vypsán za každou získanou dvojicí přímo do derivačního souboru `ssc.derivations` nebo `ssc.derivations_from_definitions`.

Značky derivačních tříd vycházejí původně z diplomové práce pana Ing. Stanislava Černého [18], kde jeho systém ohodnocování derivací vytváří čtyřmístné číselné značky. Od roku 2010, kdy byl tento systém vytvořen, prošel zjednodušovacím procesem. Zjednodušený koncept na třímístnou značku vytvořil Martin Dostál a nový návrh zdokumentoval do souboru

`vazby-navrh`. Některé kategorie ale byly zcela nerozlišitelné z pohledu dostupných dat ze zpracování slovníku SSČ a naopak, některé kategorie zcela chyběly (např. odvozování jména obyvatel od názvu státu nebo odvození ženského jména od mužského apod.), jelikož slovník SSČ obsahuje poměrně bohatá data z různých odvětví. Proto jsem upravila původní verzi a vytvořila vlastní, která je popsána v souboru `ssc.deriv_tags.txt`. Některé kategorie jsem odstranila a potřebné naopak přidala. Výsledný systém trojmístných značek je poté použit v modulu `MakeCbd.py`.

Tvoření souborů `ssc.derivations_from_definititions.cbd` a `ssc.derivations.cbd` je druhým úkolem tohoto skriptu. Formát zápisu dat vychází ze souboru *lpn* a je ve tvaru *class:base#paradigm:derivation#paradigm*, kde *class* je číselná značka třídy derivace, *base* je slovo základové s jeho vzorem podle souboru *lpn* a *derivation* je slovo odvozené se vzorem podle souboru *lpn*. Každý soubor obsahuje derivační dvojice z předchozího zpracování derivačním skriptem. Takto dojde k propojení celého modulu se zbytkem zpracování slovníku a k vytvoření různorodých dat s možnostmi vytváření nových zpracování a zkoumání v budoucnosti.

# Kapitola 5

# Implementace Derivačního modulu

Pro implementaci byl použit jazyk Python verze 3. Hlavním důvodem výběru tohoto jazyka je jeho použití ve všech ostatních skriptech Morfologického analyzátoru. Mezi další výhody patří podpora zpracování textu a přehlednost kódu díky použití vestavěných funkcí, které tento jazyk nabízí. Nevýhodou jazyka Python je, že patří mezi časově náročnější. Tento aspekt je však relevantní vzhledem ke zpracování celého Morfologického analyzátoru na školním serveru Minerva 1, který obsahuje dvakrát procesor Intel Xeon E5-2640 s frekvencí 2,50 GHz, 128 GB paměti RAM a 75 TB prostoru na pevných discích.

Jak již bylo popsáno v návrhu, Derivační modul se dělí do tří celků poskytujících různé typy zpracování derivací nad daty Slovníku spisovné češtiny. V následujících podkapitolách bych tyto tři části ráda popsala z hlediska jejich implementace a řešení problematiky zpracování dat ze slovníku.

Ve skriptech `MakeDef2Der.py` a `MakeDer.py` bylo třeba implementovat správné řazení nalezených derivací podle české abecedy. Pro tento účel jsem použila modul jazyka Python `locale`, kde jsem nastavila jako použitý jazyk češtinu. Jelikož seznamy a slovníky v jazyce Python nedodržují pořadí, v jakém byly načteny, bylo třeba pro data, která měla být seřazena, nastavit speciální typ datové struktury – slovník s řazením (v jazyce Python označovaný jménem *OrderedDict*). Samotné řazení bylo poté implementováno pomocí funkce `sorted`, v jejímž parametru byla použita funkce *lambda* pro seřazení klíčů slovníku podle nastavené české lokalizace.

## 5.1 Derivace z definic lemmat

Nejdříve prochází zpracováním soubor `ssc.def`, ze kterého se získávají lemmata a jejich definice. Ten je načten do datové struktury slovníku, jelikož v něm jsou data přehledně uložena podle lemmat a jejich definic. Takto načtená slovníková data jsou poté předávána funkci `find_lemmas`.

Tato funkce tvoří samotné jádro celého skriptu, neboť jejím hlavním úkolem je vyhledávat skrytá lemmata v definicích základních slovníkových hesel. Definice ve slovníku SSČ jsou velice různorodé, obsahují velké množství různých značek ale i různé typy zápisů pro nová lemmata skrytá v definicích (náhled do struktury dat v souboru `ssc.def` viz Příloha D.1). Abych pokryla co největší množství takto skrytých lemmat, vytvořila jsem síto regulárních výrazů, které postupně filtrují definice jednu po druhé a testují případnou shodu s jedním ze vzorů pro lemma v definici. Regulární výrazy musí být velmi konkrétní, aby nedocházelo k výběru slov, které derivacemi ve skutečnosti nejsou. V případě, že je nalezeno nové slovo,

dochází k dělení původní definice na lemma a část definice, která k němu patří a na derivace a jejich části původní definice. Díky tomuto rozdělení lze nově nalezená slova poskytnout skriptu `xml2lntrf.py`, který pak pro ně dokáže sestavit značku na základě definice, která k nim patří, a začlenit je tím do souboru *lntrf*.

Jedním z nejkomplexnějších zpracování prochází získávání názvů národů, národností a z nich vycházejících přídavných jmen. Názorným příkladem je lemma *Afghánistán* a její definice (viz výpis 5.1). Regulární výraz pro národnosti dělí definici podle výskytů údajů o skloňování, čárek a pozice přípony *-ka*, která naznačuje ženský tvar pro danou národnost. Tato lemma v prvé řadě obsahuje v definici svou dubletu, která se rozpoznává podle toho, že stojí mezi začátkem řádku a skloňováním samotného lemmatu. Dále se odděluje mužský tvar pro národnost, pokud je přítomen. V některých případech chybí název státu a místo něj je základním lemmatem právě národnost v mužském tvaru. Také je možný výskyt dublety mužské národnosti, která se musí nejen získat, ale zároveň se k ní musí vytvořit i příslušný ženský tvar. Jako poslední se v definici nachází ženský tvar národnosti, a to buď v podobě přípony *-ka*, anebo v nepravidelných tvarech je vypsán celý. Například lemma *Čech* (viz výpis 5.1) názorně ukazuje jednu z definic, kde chybí název národa a zároveň se zde vyskytuje nepravidelný ženský tvar uvedený v plné podobě.

```
Afghánistán
  , [-ny-] , Afganistan , -u m st.; Afg(h)án|ec , -nce m; -ka , -y ž;
  afg(h)ánský
Čech
  , -a m (1. mn. Češi , Čechové); Češka , -y ž; český
```

Výpis 5.1: Definice národností se zvýrazněnými derivacemi

Ženský tvar je vždy závěrečnou částí, o kterou se regulární výraz zastaví. Přídavná jména na konci nelze rozpoznávat regulárním výrazem, jelikož jejich počet je různý (nemusí zde být žádné) a především neobsahují žádnou gramatickou poznámku, o kterou by se šlo opřít. Abych získala i tato slova, testuji konec definice oddělené pro ženskou národnost na výskyt slov s koncovkami *-ský* a *-cký*, což je možné jen díky pravidelným tvarům tohoto typu adjektiv. Konkrétně definice lemmatu *Afghánistán* je zajímavá také tím, že obsahuje dvě varianty zápisu slov uvedením možného písmenka v závorce. I toto je ve zpracování zohledněno, což znamená, že z této definice bylo získáno celkem 7 derivací.

Další početnou skupinou derivací z definic je získávání derivací za znakem „■“. Zde se mohou nacházet homonyma nebo slova slangová s přeneseným významem. I když mají tyto derivace snadný způsob získávání, je u nich důležitý faktor správně vedených poznámek. Pokud získáváme homonyma, pak se v souboru *lntrf* začnou vyskytovat stejně vypadající slova, ale různých slovních druhů (např. slovo „*co*“ je zde celkem šestkrát). Aby bylo později rozpoznatelné, jakým směrem derivace vznikla, a byla správně určena derivační třída modulem `MakeCbd.py`, je třeba vést systém poznámek, které jsou pak modulu předány jako jeden z parametrů. Poznámka obsahuje dvojici slov *lemma -> derivace* jako klíč k pozdějšímu vyhledávání a v hodnotě je uložena informace o jejich slovních druzích v daném pořadí. Názorným příkladem tohoto typu definic je např. lemma *dozorčí* (viz výpis 5.2).

```
dozorčí
příd. kontrolní, inspekční : dozorčí orgán ■ dozorčí, -ho m (2. mn
. -ích) : voj. dozorčí roty, kuchyně
```

Výpis 5.2: Definice lemmatu dozorčí se zvýrazněnými derivacemi

Různé varianty jednoho slova ve slovníku mohou být zapsány více způsoby. První možností je, že oba tvary mají stejnou koncovku v druhém pádě jednotného čísla i stejný rod a nemění se slovní druh. Pak jsou zapsány přímo za sebou, jako např. u slova *česač* (viz výpis 5.3). Pokud dochází ke změně jednoho z výše zmíněných údajů, slovník je zapisuje rozděleně, každé slovo s jeho gramatickými poznámkami, jako je tomu u slova *dareba* (viz výpis 5.3). Často je situace také zkomplikována přidáním údajů o skloňování i v jiných pádech, než je nominativ singuláru, čím z hlediska analýzy textu přibývají závorky, například u lemmatu *křemenáč* (viz výpis 5.3). Každá z těchto variant musí mít vlastní regulární výraz, jelikož při použití příliš obecných vzorů by docházelo k chybám.

```
česač
  , česáč, −e m kdo češe, sklízí plody : česači ovoce, chmele;
dareba
  , −y m, darebák, −a m (6. mn. −cích ) lotr, darebný člověk, ni
    čema, lump, padouch expr. nezbedné, rozpustilé dítě, uličník;
křemenáč
  , −e m (4. j. i −e), křemeňák, −a m (4. j. i −a, 6. mn. −cích)
    (než./živ.) jedlá hřibovitá houba s oranžovým kloboukem na šti
    hlém třeni ;
```

Výpis 5.3: Definice se zvýrazněnými dubletami

Poslední velkou skupinou derivací získávaných z definic jsou deverbativa. Mohou být z řad nominativ i adjektiv, ale může to být i verbum, jakožto dubleta původního lemmatu. Pro získání všech výrazů se aplikují dva regulární výrazy. První, v hlavním sítu regulárních výrazů, má za úkol rozpoznat, že definice slovesa obsahuje deverbativum, čímž dojde k zachycení této definice a jejímu rozdělení na dvě části – část definice patřící k lemmatu a první deverbativum a zbytek definice. Tato druhá část poté prochází méně složitým regulárním výrazem, jehož úkolem je rozdělit ji na další části podle případných deverbativ. Názornou ukázkou může být definice lemmatu *anektovat* (viz výpis 5.4).

```
anektovat
  dok. i ned. násilně připojit, zabrat, zabírat : anektovat
    pohraničí; anexe, −e ž :anexe cizího území;anexní, anekční příd.
```

Výpis 5.4: Definice národností se zvýrazněnými derivacemi

Po zpracování všech definic slovníku je volána funkce `derivation_print`, která v nejdříve volá modul `MakeCbd.py`. Modulu je pomocí parametrů předán seznam získaných derivací a mimo jiné také seznam vzniklých poznámek pro správné určení derivačních tříd. Modul poté vrací modifikovaný seznam derivací, obohacený o derivační třídu a změnu koncovek. Blíže bude fungování modulu popsáno v kapitole 5.3.

Poté dochází k výpisu získaných derivací do grafické podoby, která názorně zobrazuje lemmata a od nich odvozené derivace. Graf, ve kterém jsou slova vypsána, může mít až dvě úrovně, které vznikají zejména pokud od dublet vznikají další derivace. Výsledným souborem celého procesu je `ssc.derivations_from_definitions` (viz Příloha D.4).

Získané derivace jsou testovány skriptem `CheckLemma.py`, který je popsán v kapitole 6.1. Skript `MakeDef2Der.py` je nepřenositelný mezi slovníky, jelikož pracuje s obsahem konkrétního slovníku a je mu tzv. „šitý na míru“. Každý slovník Morfologického analyzátoru bude pro získání derivací z definic vyžadovat vytvoření svého vlastního skriptu, který postihne všechna specifika daného slovníku. Vzniklý soubor derivací je však již univerzální a dále zpracovatelný společnými skripty pro všechny slovníky.

## 5.2 Derivace podle podobnosti slov

Derivace podle podobnosti slov se také jinak nazývají derivace ze zahnízďování. Tento název vznikl podle způsobu, jakým se slova sdružují do skupin. Celý proces začíná u jednoho slova, kolem kterého se podle podobnosti shlukují další, až vytvoří skupinu vzájemně si podobných a významově souvisejících slov. V případě slovníku SSČ tyto derivace vznikají na základě souboru `ssc.lntrf` (viz Příloha D.2). Získávání lemmat z tohoto typu vstupu bylo zvoleno proto, že obsahuje všechny doposud získané derivace z definic. Zpracování těchto lemmat prochází několika důležitými kroky, které bych v této kapitole ráda popsala.

Prvním krokem je získat základní lemma, kolem kterého se ostatní budou shlukovat. Abych vytvořila základy pro budoucí hnízda, vzala jsem všechna podstatná jména, která soubor `ssc.lntrf` obsahuje. Poté přichází na řadu srovnávání s ostatními lemmaty, která již tvoří základ hnízd. Je zde možnost, že k lemmatu, které je právě zpracováno, již existuje základ hnízda, kam by mělo být přiřazeno. Buď se jej tedy podaří přiřadit, anebo samo začne tvořit základ pro potenciální budoucí skupinku.

Přiřazování lemmat k hnízdům je závislé na vzájemné podobnosti srovnávaných slov. Výpočet podobnosti zpracovává funkce `count_similarity`, jejímž vstupem jsou dvě slova, která mají být srovnána. Funkce počítá procentuální podobnost kratšího slova vůči delšímu, jinými slovy kolik procent delšího slova tvoří společná písmena. Je třeba vzít v úvahu, že slovník obsahuje slova o délce 1 až 24 písmen. Pokud bychom srovnávali např. slova *afekt* a *afektovanost*, je jasné, že v případě kratšího slova je shoda 100 %, jelikož všech 5 znaků je společných. Toto číslo o ničem nevypovídá, avšak pokud je srovnáváme podle delšího slova, zjistíme, že podobnost je ve skutečnosti pouhých 45 %. Funkce nevrací jenom procentuální podobnost, ale také pozici písmene ve kterém se daná slova liší a je volána z různých pozic v kódu díky svému univerzálnímu využití.

Utváření prvotních potenciálních základů pro hnízda je velmi citlivý bod z hlediska rozhodování o přiřazení či nepřiřazení slova k již existujícímu. Cílem je eliminovat co největší počet nesprávně přiřazených slov, jelikož u podstatných jmen existuje velké množství podobných slov, která spolu ale ve skutečnosti nijak nesouvisí. Rovnou jsem vyloučila ta slova, jejichž délka je kratší než 5 písmen. Tato lemmata jsou bez dalších údajů nepřiřaditelná, vždy se shlukují mezi sebou, jelikož se obvykle liší pouze v posledním písmenku. Velké rozdíly mezi délkami slov vyžadují, aby se algoritmus rozhodoval v závislosti na délkách srovnávaných slov. Proto je u přiřazování substantiv nastavena podmínka následujícím způsobem: pokud je alespoň jedno ze srovnávaných slov delší než 7 písmen, stačí, aby jejich podobnost byla 63 %. Pokud jsou však obě kratší než 7 písmen, jejich podobnost musí být více než 81 %. Tato čísla byla vybrána cíleně, na základě testování různých hodnot. Podobnost 81 % také zcela záměrně vylučuje možnost, že by k sobě byla přiřazena dvě slova o délce 5 písmen, která by se lišila pouze v poslední hlásce. Číslo 63 % pro delší slova zase eliminuje přiřazení slov s 5ti písmeny ke slovům delším než 7 písmen. Obě nastavené hranice pokrývají případy s nejčastějším chybným přiřazením z důvodu velkého nebo naopak příliš malého rozdílu mezi slovy.

Ve chvíli, kdy máme vytvořené základy pro vytváření skupin, můžeme zkusit přiřadit další část slov. V druhém kroku se připojují přídavná jména. Z hlediska podobností má tento slovní druh skvělé vlastnosti, jelikož má většinou stejný začátek jako základní slovo a typickou koncovku. Právě koncovka prodlužuje slovo o nejméně 3 znaky, proto je hranice podobnosti u slov delších než 7 písmen posunuta níž na 57 %.

Podobně dobré vlastnosti jako přídavná jména mají i slovesa. Ta mají hranici pro delší slova sníženou až na 55 % a to z důvodu častého výskytu derivací, kde je shoda

na 5 písmen, ale delší slovo má délku 9 písmen. Konkrétním příkladem může být dvojice *abdikace -> abdikovat*.

Poslední fází utváření derivační skupiny je připojení slov ostatních slovních druhů. Neboť jde o velmi rozmanitou směs slov, jsou opět podmínky přiřazení nastaveny přísněji, jako u podstatných jmen. Pro delší slova je hranice nastavena na 60 %, u srovnávání slov kratších než 7 znaků je vyžadována podobnost 81 %.

Pokud se nějaké slovo během tohoto procesu nepodaří přiřadit, pak končí v seznamu nazvaném `word_bin`. Uchovávání nepřiřazených slov umožňuje jejich zpracování a pokus o opětovné přiřazení později.

Na závěr tohoto kroku se ještě provede třídění získaných skupin tak, že všechna podstatná jména, která zbyla sama a nebyla k nim přiřazena žádná další slova, jsou přesunuta do seznamu `singleton_bin`, kde čekají na další pokus o přiřazení. Tento krok rapidně sníží počet položek ve slovníku, ve kterém jsou nalezené definice přechovávány, a zrychlí to jeho procházení.

Na základě této dekompozice se skript snaží vypořádat s co největším počtem chybně zařazených slov (problematika paronym popsaná v kapitole 4.1). Nejlepším ukazatelem, že slovo patří do jiné skupiny, je odlišnost v sufixové části. Pokud při porovnávání sufixů nedojde ke shodě a odlišné písmeno není ani v tolerovaných výjimkách, pak je slovo označeno za chybné, je přesunuto do seznamu `word_bin`, a jeho sufix a koncovka jsou vymazány z údajů skupiny. Pokud po tomto procesu ve skupině zůstalo více než jedno slovo, je skupina zařazena do finálního seznamu derivačních skupin.

Ten je předáván modulu `MakeCbd.py` pro zpracování a určení derivačních tříd. Výstupem je poté soubor `ssc.derivations` (viz Příloha D.5), který obsahuje graficky vypsanou skupinu derivací, údaje o derivační třídě, změnu koncovky pro každou dvojici a pod skupinou je vypsána statistika z dekompozice slov skupiny. Zde je uveden největší společný kmen, sufixy a koncovky.

Výpis informací o dekompozici je v souboru především pro účely bakalářské práce, pokud v budoucnosti bude docházet k dalšímu zpracování tohoto souboru, budou se dekompoziční informace předávat jako jeden z parametrů a zůstanou nevypsány. Vzniklý soubor derivací je univerzální, všechny soubory tohoto typu budou mít stejnou podobu nezávisle na slovníku, ze kterého byly získány. Znamená to také, že vypsaná data jsou dále zpracovatelná společnými skripty pro všechny slovníky.

## 5.3 Vytváření derivačních tříd

Tento modul je umístěn v samostatné složce mimo ostatní slovníky, jelikož jeho účelem je být volán z jakéhokoliv derivačního skriptu libovolného slovníku Morfologického analyzátoru. S těmi je propojen pomocí zadání cesty k jeho složce a příkazu `import MakeCbd`.

Práce tohoto modulu začíná ve chvíli, kdy si jej zavolá jeden z výše popsaných derivačních skriptů. Konkrétně se volá funkce `derivation_class`, které skript pomocí parametrů předá informaci o slovníku, se kterým pracuje, seznam nalezených derivací, slovník lemmat a definic ze souboru typu *def* a případné poznámky k některým dvojicím. Tato funkce pak řídí zpracování dat a volá další interní funkce modulu.

První volanou a hlavní funkcí je `create_class`, která dostává postupně získané a ohodnocené derivační dvojice a určuje třídu derivace, do které dvojice spadá. Toto rozhodnutí se provádí na základě velkých a malých písmen ve slově, koncovky, slovního druhu podle souboru *lntrf*, některých údajů v souboru *def* nebo případných poznámek poskytnutých

derivačním skriptem. Struktura implementace rozhodovacího algoritmu odpovídá rozdělení derivačních tříd na kategorie následujícím způsobem.

Derivační třídy jsou rozděleny na 7 kategorií podle slovního druhu získaného slova odvozeného na podstatná jména, přídavná jména, zájmena, číslovky, slovesa, příslovce a ostatní. Podstatná jména jsou nejobsáhlejší skupinou, jejich značka začíná číslem 1 a dělí se na další podkategorie: osoba, vlastnost, místo, atd. až po poslední podkategorii s číslem 9 – ostatní. Sem spadají derivace jmen, dublety neživotných slov a odvození podstatného jména od příslovce, předložky, spojky, částice nebo citoslovce. Přídavná jména, zájmena, číslovky, slovesa ani příslovce se na podkategorie již nedělí, jejich značku udává pouze to, od jakého slova, resp. slovního druhu, byly odvozeny. Poslední kategorie se značkou začínající na číslo 7 se nazývá Ostatní. Pokrývá vzájemná odvození mezi dalšími slovními druhy od příslovcí až po citoslovce, kterých je velké množství zejména díky derivacím, získaným z definic slovníku. Přesné rozdělení kategorií si lze prohlédnout v souboru `ssc.deriv_tags.txt`.

Ve chvíli, kdy je určena značka derivační třídy, funkce `get_change` určí změnu koncovky. Tento údaj je připojen za získanou číselnou značku a připsán k dané dvojici slov. Kompletní seznam derivací s přidanými informacemi je pak vrácen volajícímu skriptu, který jej vypíše do svého souboru s derivacemi.

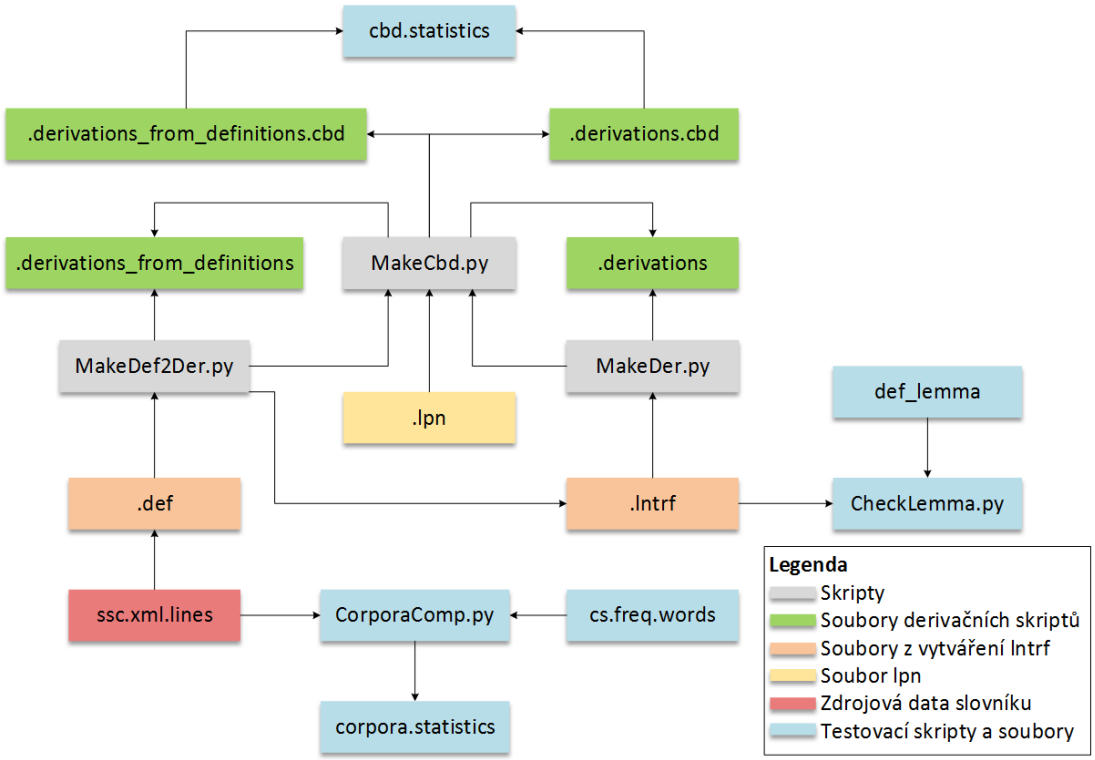
Údaje o značce, původním slovu a jeho derivaci jsou zároveň ukládány do seznamu pro vypsání souboru *cbd*. Ten se sestavuje až po získání značek pro všechny dvojice. Tuto část modulu spravuje funkce `print_cbd`. Používá data ze souboru *lntrf*, jelikož ten obsahuje stylistické poznámky ke každému slovu. Tyto poznámky se promítají i do souboru *lpn* a v případě, kdy se jedno slovo vyskytuje v *lpn* vícekrát, stylistická poznámka je může pomoci rozlišit.

Pokud jsou vzory a poznámky úspěšně nalezeny, jsou jednotlivé dvojice vypsány s jejich derivační třídou a flektivními vzory ze souboru *lpn*. Data jsou seřazena vzestupně podle číselné značky, jelikož na tomto zobrazení je názorně vidět rozložení derivací mezi jednotlivými kategoriemi. Jména výsledného souboru *cbd* se odvíjí od skriptu, ze kterého byla data získána. Při současném zpracování je to buď `ssc.derivations_from_definitions.cbd` nebo `ssc.derivations.cbd` (viz Příloha D.6).

# Kapitola 6

# Vyhodnocení Derivačního modulu

V této kapitole jsou uvedeny způsoby, jakými byly výsledky a algoritmy testovány. Také jsou zde zmíněna úskalí a problematiky některých derivačních skriptů a navržení jejich případných řešení, která přesahují téma této práce. Zároveň jsem v kapitole 6.2 navrhla možné budoucí rozšíření a téma pro další výzkum, pro který tato práce vytvořila podklad. V závěru kapitoly je pak vyhodnocena rychlost derivačních skriptů vůči ostatním částem zpracování slovníku SSČ. Pro přehlednost jsem vytvořila názorný diagram Derivačního modulu (viz obrázek 6.1) s testovacími skripty, soubory a získanými statistikami.




Obrázek 6.1: Struktura Derivačního modulu s testovacími skripty a získanými statistikami

## 6.1 Testování vzniklých derivací z definic

Pro sestavení a odladění regulárních výrazů jsem použila webovou aplikaci[1] Regular expressions 101, která poskytuje jednoduché webové rozhraní pro testování regulárních výrazů v různých programovacích jazycích na uživatelem poskytnutých datech. Umožňuje také regulární výraz i s příslušnými daty uložit a kdykoliv se k němu vrátit pomocí *URL*. Této funkce jsem využila tak, že u každého regulárního výrazu v sítu skriptu `MakeDef2Der.py` je v komentáři uveden daný odkaz a kdokoliv si může zobrazit výraz i s testovacími daty online a provádět nové testování nebo odladění chyb. Všechny následující změny jsou ukládány jako verze, které aplikace ukládá pro každý uložený regulární výraz.

Pro testování výsledků skriptu `MakeDef2Der.py` jsem vytvořila skript `CheckLemma.py`. Ten načte lemmata z testovacího souboru, kontroluje, jestli jsou obsažena a správně zpracována v souboru `ssc.lntrf` a poté vypíše na výstup počet a seznam lemmat, která stále chybí. Testovací soubor jsem sestavila ručně procházením zdrojových dat slovníku, přičemž jsem vypisovala různé typy skrytých derivací v definicích tak, abych pokryla všechny možné způsoby zápisu a výskytu nových slov. Vznikl tak soubor `def_lemma` obsahující 1322 různých slov, která by se měly v souboru *lntrf* objevit. Porovnávání s tímto souborem jsem zvolila záměrně proto, abych měla jistotu, že nové derivace se objeví nejen v souboru `ssc.derivations_from_definitions`, ale budou také úspěšně zahrnuty do dalšího zpracování dat slovníku, které vychází právě ze souboru `ssc.lntrf`. Neméně důležitý je také fakt, že ze souboru *lntrf* vychází další derivační skript `MakeDer.py`, což znamená, že nové derivace budou s určitostí zahrnuté i v derivacích, které vytváří.

## 6.2 Vyhodnocení derivací podle podobností

Na vyhodnocení správnosti tohoto typu derivací nelze vyrobit automatický skript. Kontrola získaných dat byla prováděna ručně, procházením souboru, stejně jako kontrola číselných značek derivačních tříd. Pro zajímavost byla vytvořena statistika `cbd.statistisc`, kde je vypsáno, kolik derivačních dvojic bylo ke každé třídě přiděleno. Výpis si lze prohlédnout v přiložených souborech na CD.

Problém derivací získávaných zahnízďováním jsou paronyma. Velké množství jich bylo odstraněno při použití údajů o sufixech a koncovkách při dekompozici slov. Účinnou obranou proti tomuto typu chyby je také správné nastavení hranic pro přiřazování slov ke skupinám. Jsou ale taková, která touto metodou nelze odstranit. Např. dvojice *Atlas -> atlasový* nebo *následek -> následník* od sebe nelze rozdělit na základě dekompozice. Rozdíl mezi těmito slovy je na sémantické úrovni. Na základě analýzy definice slov ve slovníku by se muselo rozhodnout, zdali kromě společného kmene mají slova i související význam. Tímto krokem by se do skriptu přidala další rovina analýzy. V této práci jsem zůstala pouze na úrovni morfologické, ale vnímám to jako další možnosti budoucího vývoje Morfologického analyzátoru.

Další možností pro rozšíření tohoto skriptu je vývoj algoritmu pro znovu-přiřazování slov, která se nepodařilo přidat k žádné ze skupin. Seznam těchto slov je uchováván v proměnné `word_bin`. Na základě údajů z dekompozice by se mohlo porovnávat, jestli dané slovo po dekompozici obsahuje podobnou strukturu a podle toho jej zařadit do dané skupiny. Podobně problematickou skupinou jsou slova o délce kratší než 5 písmen, což je problematika zejména podstatných jmen. U takto krátkých slov lze předpokládat spíše shodu s kořenem

[1] Aplikace je dopstupná na: https://regex101.com/

jedné z utvořených skupin, než porovnávání sufixů a přítomnost koncovek. Podstatná jména na rozdíl od přídavných jmen, sloves nebo příslovcí nemají tak jasně vyhraněné koncovky.

## 6.3 Rychlost a paměťová náročnost derivačních skriptů

Z hlediska rychlosti skripty Morfologického analyzátoru nejvíce ovlivňují pasáže, kdy srovnávají nebo zpracovávají obsáhlé soubory a iterace cyklů. Tabulka 6.1 ukazuje skripty zpracovávající slovník SSČ, jejich dobu běhu a soubory, se kterými pracují (v závorce jsou uvedeny velikosti těchto souborů). V případě derivačních skriptů je naměřený čas včetně modulu `MakeCbd.py`. Nejdelší dobu běhu má stále jednoznačně skript `lntrf2lpn.py`, který vyhledává vzory, a poté skript `lpn_split.py` porovnávající tři obsáhlé soubory. Třetí nejdéle běžící skript je `MakeDer.py`, jehož dobu běhu ovlivňuje porovnávání a rozřazování slov do derivačních skupin.

| **Skript** | **Doba běhu** | **Soubory** |
|---|---:|---|
| xml2lntrf.py | 1 m 34 s | ssc.xml.lines (152,9 MB) |
| lntrf2lpn.py | 517 m 24 s | czech.paradigms (2,7 MB) a ssc.lntrf (1,5 MB) |
| lpn_split.py | 115 m 19 s | ssc.singleword.lntrf (1,5 MB) a ssc.lpn (1,1 MB) a czech.lpn (8,5 MB) |
| MakeDef2Der.py | 0 m 17 s | ssc.def (6 MB) |
| MakeDer.py | 14 m 34 s | ssc.def (6 MB) a ssc.lntrf (1,5 MB) |

Tabulka 6.1: Srovnání rychlosti zpracování jednotlivých skriptů

Pro zjištění údajů o použité paměti byl použit nástroj `memory_profiler`. Byla změřena použitá paměť pro nejdůležitější funkce derivačních skriptů a celková použitá paměť. Srovnání paměťové náročnosti obou derivačních skriptů a jejich funkcí si lze prohlédnout v tabulce 6.2.

| **Skript** | **Funkce** | **Využitá paměť** |
|---|---:|---|
| MakeDef2Der.py | | |
| | main | 434,9 MiB |
| | get_args | 81,4 MiB |
| | find_lemmas | 8 736,0 MiB |
| | derivation_print | 780,5 MiB |
| | Celkem | 10 059,8 MiB |
| MakeDer.py | | |
| | main | 642,2 MiB |
| | get_args | 78,8 MiB |
| | count_similarities | 670,6 MiB |
| | get_derivation | 3 965,0 MiB |
| | decomposition | 2 576,8 MiB |
| | print_derivations | 812,4 MiB |
| | Celkem | 8 745,8 MiB |

Tabulka 6.2: Srovnání využití paměti jednotlivými skripty

# Kapitola 7

# Závěr

Cílem této práce bylo rozšířit Morfologický analyzátor Výzkumné skupiny znalostních technologií o část, která by zpracovávala derivační morfologii. Tento záměr byl splněn implementací Derivačního modulu.

Nastudovala jsem problematiku z hlediska lingvistických disciplín a popsala již existující nástroje pro derivační morfologii. Práce se zakládá na Slovníku spisovné češtiny, který byl porovnán s českým korpusem firmy Seznam.cz. Toto srovnání ukázalo, že jsou pokryty derivace velké části slov, která se vyskytují v korpusu. Na základě současného zpracování slovníku jsem navrhla a implementovala rozšíření Morfologického analyzátoru, které je integrované do jeho struktury. Získává velké množství dat, mezi která patří derivace z definic, derivace podle podobností slov, dekompozice slov na slovotvorné či kmenotvorné morfémy a ohodnocené derivační dvojice číselnou značkou jejich derivační třídy, čímž byla vytvořena podpora pro generování slovotvorných vazeb. Na závěr byla vyhodnocena rychlost a paměťové nároky všech tří částí implementovaného Derivačního modulu a vytvořen plakát prezentující cíle a výsledky práce. Tímto byly splněny všechny body formálního zadání.

Pomocí navrženého modulu se podařilo získat více než 4 500 nových slov z derivací skrytých v definicích. Tato slova obohatila nejen další zpracování slovníku, ale i slovní zásobu Morfologického analyzátoru. Z celkového počtu 53 000 slov bylo v další fázi utvořeno téměř 8 800 skupin podle podobnosti jednotlivých slov, nad kterými byla provedena dekompozice. Celkem tedy bylo ohodnoceno a zařazeno do derivační třídy více než 20 000 derivačních dvojic.

Největší potenciál získaných dat vidím v jejich budoucím využití. Všechny dvojice, číselné značky i dekompozice tvoří podklad pro možné budoucí výzkumy. Jako jedno ze zajímavých témat vnímám vytvoření derivačních vzorů. Morfologický analyzátor již obsahuje systém flektivních vzorů, který se zakládá na údajích ze souborů *lntrf* a určuje vzory pro skloňování zpracovaných slov, ze kterých pak lze odvodit skloňování i v pádech, které slovník neuvádí. Pokud bychom na základě dekompozice a značek mohli vytvořit systém derivačních vzorů, mohli bychom odhadovat derivace, které zatím neznáme.

Zajímavým výzkumným podnětem může být také studium potenciálních jevů, resp. tvarů v derivační morfologii a vytváření zcela nových slov. Tímto způsobem by se také snadno začleňovaly do Morfologického analyzátoru neologizmy nebo slova počeštěná, původem z jiných jazyků. Tyto typy slov se objevují především díky globalizačnímu trendu současné doby, jelikož slovní zásoba jazyka je na změny ve společnosti velmi citlivá.

# Literatura


[1] Adam, R.: *Morfologie Příručka k povinnému předmětu bakalářského studia oboru ČJL*. Nakladatelství Karolinum, 2015, ISBN 978-80-246-2800-4, dostupné také z: http://www.cupress.cuni.cz/ink2_stat/dload.jsp?prezMat=53806.

[2] Bordagová, D.: Psycholingvistika a čeština: některá slibná témata. *Naše řeč*, ročník 92, 2009: s. 213–217, ISSN 0037-7031. Dostupné také z: http://sas.ujc.cas.cz/archiv.php?lang=en&art=4078.

[3] Cvrček, V.; Kováříková, D.: Možnosti a meze korpusové lingvistiky. *Naše řeč*, ročník 94, 2011: s. 113–133, ISSN 0027-8203. Dostupné také z: http://nase-rec.ujc.cas.cz/archiv.php?lang=en&art=8191.

[4] Giraudo, H.; Gaigner, J.: On the role of derivational affixes in recognizing complex words: Evidence from masked priming. In *Morphological structure in language processing*, ročník 151, editace R. Baayen; R. Schreuder, Mouton de Gruyter, 2003, str. 209–232.

[5] 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.

[6] Jaroslav Popela, v. B. V. a. V. B.: *Skaličkova jazyková typologie*. [Online; navštíveno 30.04.2017].
URL http://www.ujc.cas.cz/miranda2/export/sitesavcr/ujc/sys/galerie-download/oddeleni-etymologicke/popela_skalickova-jazykova-typologie.pdf

[7] LEDA: *Slovník spisovné češtiny - elektronická verze pro PC*. [Online; navštíveno 09.05.2017].
URL https://www.leda.cz/Titul-detailni-info.php?i=87

[8] Osolsobě, K.: Deriv - nástroj pro automatické vyhledávání slovotvorných vztahů (Deriv - the Word Derivation Tool). In *Přednášky a besedy ze XLII. běhu LŠSS*, Brno: Masarykova Univerzita, první vydání, 2009, s. 132–137.

[9] Pala, K.; Šmerk, P.: *Derivancze - Derivational Analyzer of Czech*. [Online; navštíveno 07.05.2017].
URL
https://link.springer.com/chapter/10.1007%2F978-3-319-24033-6_58#page-1

[10] Sedláček, R.; Smrž, P.: *A New Czech Morphological Analyser ajka*. [Online; navštíveno 07.05.2017].

URL https://nlp.fi.muni.cz/publications/tsd2001_rsedlac_smrz/tsd2001_rsedlac_smrz.pdf

[11] Smolík, F.: Psycholingvistika a čeština: některá slibná témata. *Naše řeč*, ročník 92, 2009: s. 240–251, ISSN 0027-8203. Dostupné také z: http://nase-rec.ujc.cas.cz/archiv.php?art=8063.

[12] Straková, J.; Straka, M.; Hajič, J.: *Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition*. [Online; navštíveno 07.05.2017].
URL http://www.aclweb.org/anthology/P/P14/P14-5003.pdf

[13] Varadová, E.: *Pojetí přípony a koncovky v současné české lingvistice*. Bakalářská práce, Západočeská univerzita v Plzni, Fakulta pedagogická, Katedra českého jazyka a literatury, Plzeň, 2012, Vedoucí práce PaedDr. Helena Chýlová. Dostupné také z: https://otik.uk.zcu.cz/bitstream/11025/4490/1/E_Varadova_2012_final.pdf.

[14] Václav Cvrček, a. k. a.: *Mluvnice současné češtiny 1 Jak se píše a jak se mluví*. Nakladatelství Karolinum, 2015, ISBN 978-80-246-2834-9.

[15] Čermák, F.: Jazykový korpus: Prostředek a zdroj poznání. *Slovo a slovesnost*, ročník 56, 1995: s. 119–140, ISSN 0037-7031. Dostupné také z: http://sas.ujc.cas.cz/archiv.php?art=3627.

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

[17] Čermák, F.: *Morfématika a slovotvorba češtiny*. Nakladatelství Lidové noviny, 2012, ISBN 978-80-742-2146-0.

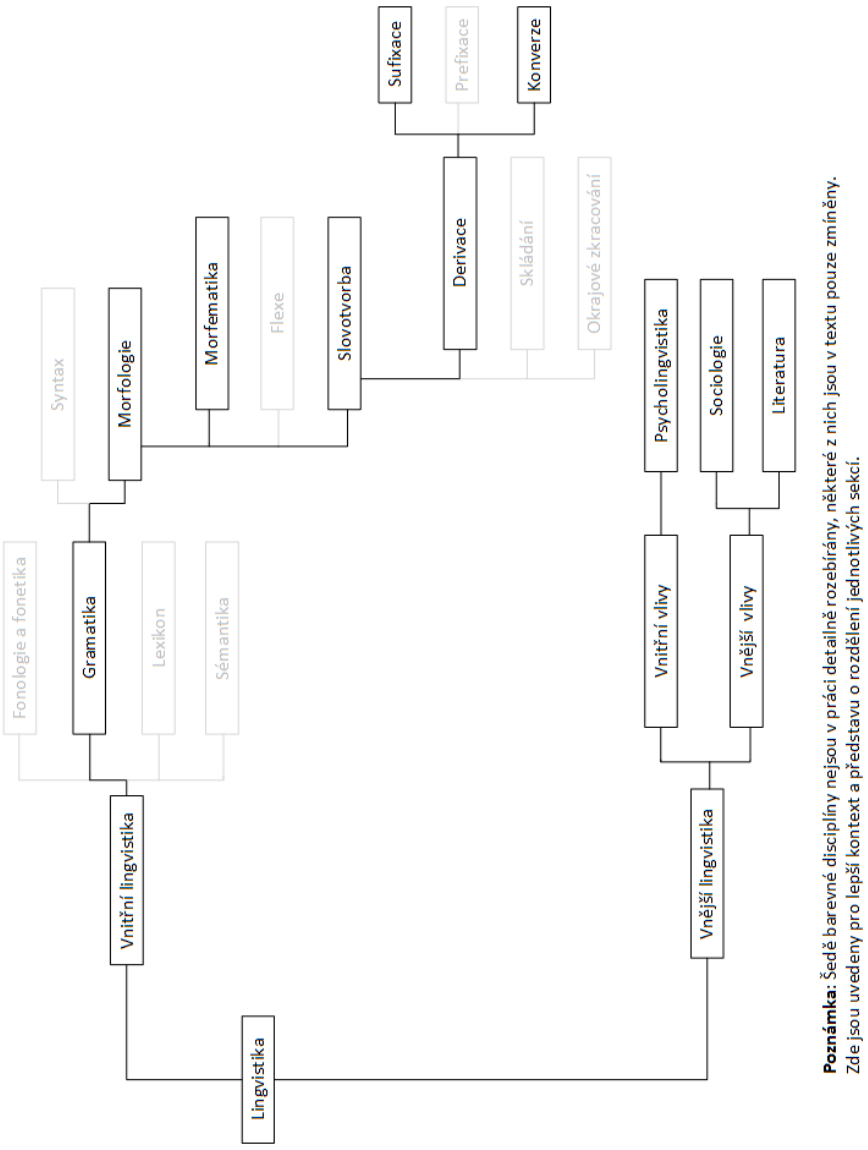
[18] Černý, S.: *Analýza slovotvorných vazeb v češtině*. Diplomová práce, Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, 2010, Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

[19] Žabokrtský, Z.; Ševčíková, M.; Straka, M.; aj.: *Merging Data Resources for Inflectional and Derivational Morphology in Czech*. [Online; navštíveno 07.05.2017].
URL http://ufal.mff.cuni.cz/~straka/papers/2016-lrec_derinet.pdf

# Přílohy

# Příloha A

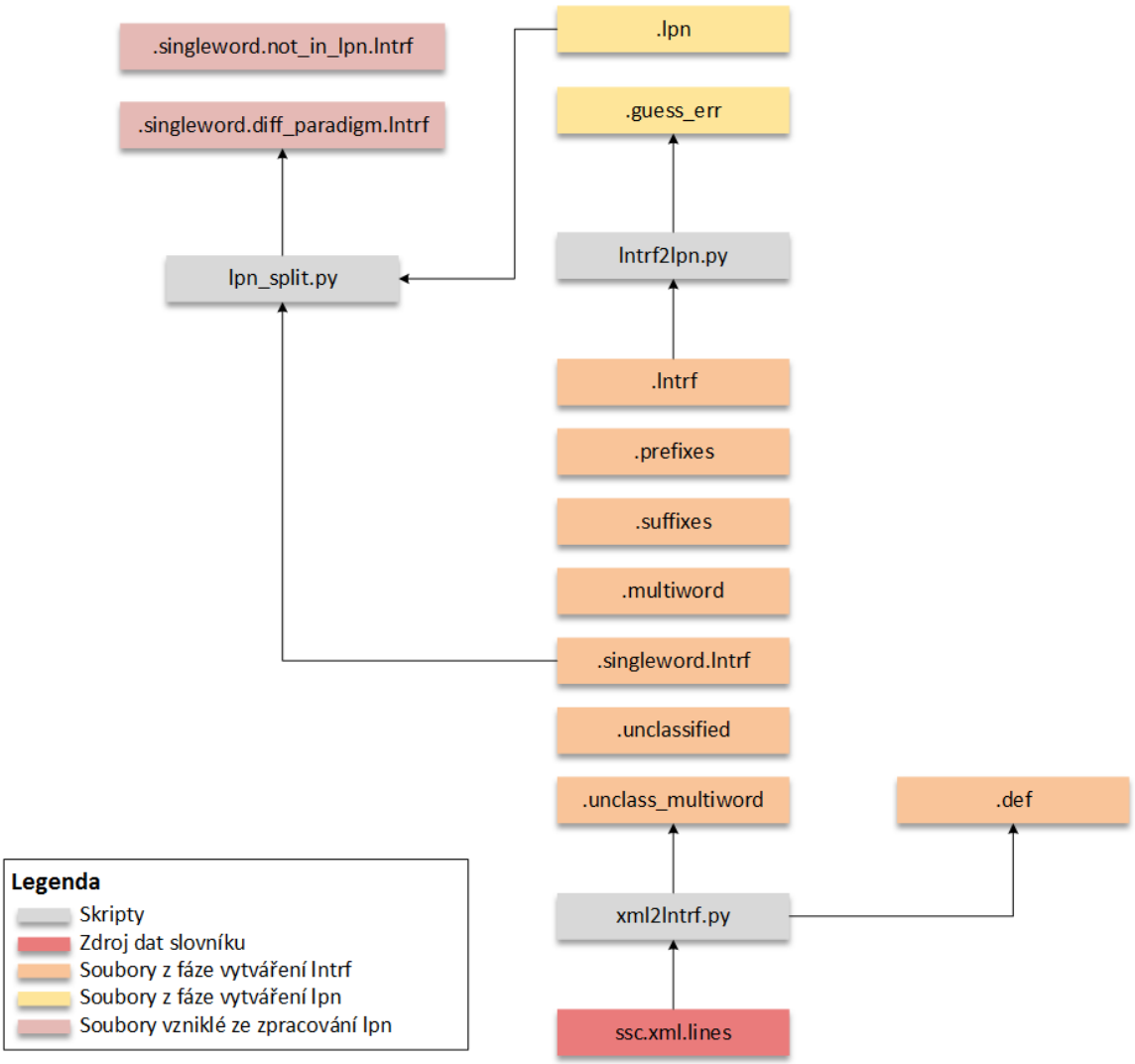
# Lingvistické disciplíny v textu




Obrázek A.1: Graf znázorňuje zařazení použitých pojmů lingvistiky

# Příloha B

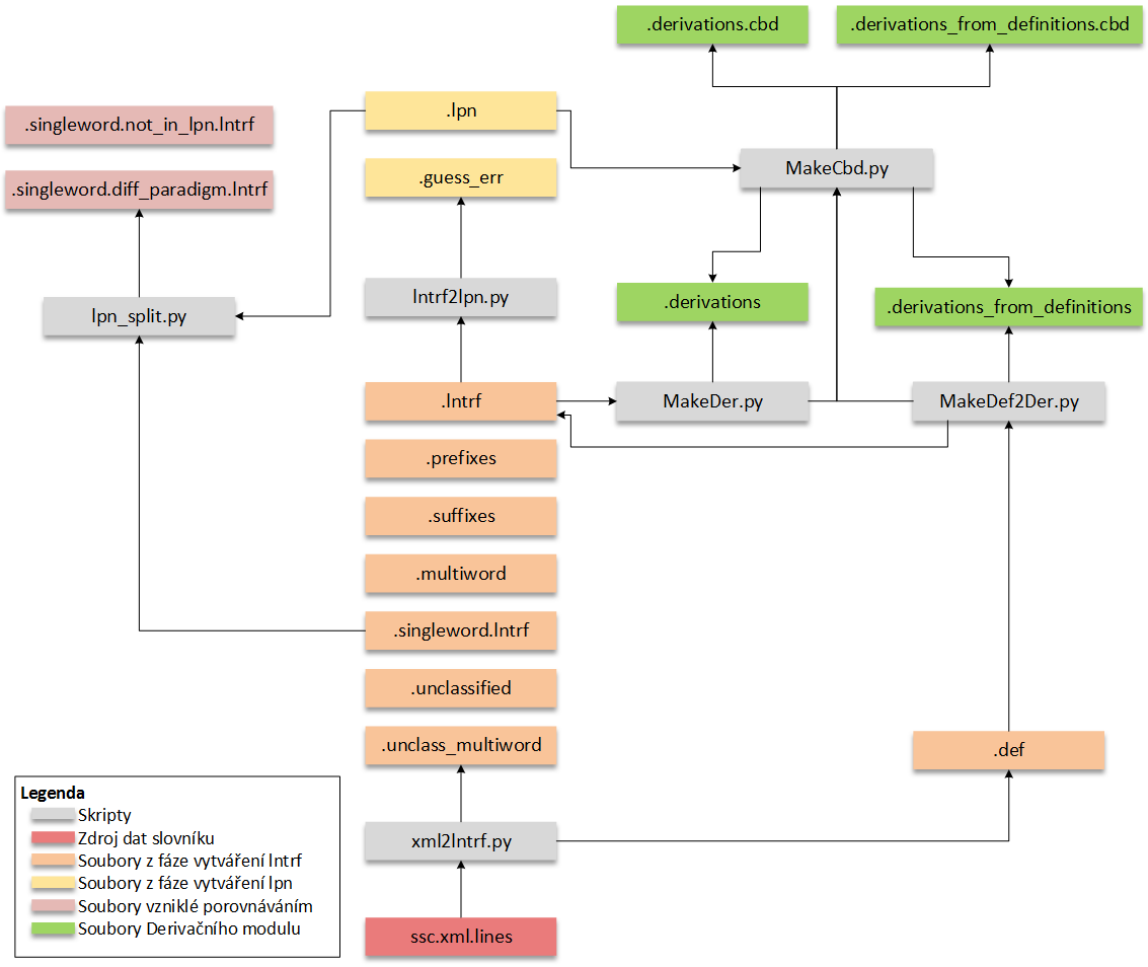
# Graf současného zpracování slovníku SSČ




Obrázek B.1: Graf skriptů zpracovávajících data slovníku SSČ a vznikajících typů souborů.

# Příloha C

# Graf zpracování slovníku SSČ s Derivačním modulem




Obrázek C.1: Graf zpracování slovníku SSČ včetně Derivačního modulu.

# Příloha D

# Ilustrační výpisy z nejdůležitějších souborů práce

Výpis D.1: Ukázka výpisu ze souboru ssc.def se zvýrazněnými lemmaty.

**dobré**
, -ho s (× zlé ) dobro 1 , prospěch 1 : obrat k dobrému po dobrém, v dobrém ve shodě, s vlídností, s přívětivostí : vyřídil to po dobrém ; v dobrém se rozešli ♦ dát něco k dobrému pohostit, pobavit něčím ; když to nejde po dobrém, půjde to po zlém (pořek.) ;

**dobro**
, -a s kladná morální hodnota , prospěch 1 , blaho (× zlo ): konat dobro dobré skutky ; dělám to pro tvé dobro mít k dobru mít (jako ) pohledávku: připsat k dobru v prospěch (účtu ) (× k tíži (v. tíže 3) ) (i přen.) ■ dobrá , dobře část. přitak. hovor. ano 1, 2 : dobře, nechám si to ; - no dobře

**dobudovat**
dok. dokončit stavbu , dostavět : dobudovat přehradu dokončit výstavbu , vybudovat : dobudovat základy demokracie

**dobýt**
dok. (1. j. -budu , hovor. -bydu , rozk. -buď , čin . -byl , trp. -byt ) zmocnit se 1 (bojem) : dobýt města, město (velkým úsilím) dosáhnout 5 : dobýt svobody vynutit , vymoci 1 : dobýt z někoho přiznání ;

**dobývat**
ned. k 1-3 získávat ( ze země ): dobývat uhlí dolovat ; dobývat brambory kopat ; dobývat se ned. násilím pronikat : dobývat se do bytu

Výpis D.2: Ukázka výpisu ze souboru ssc.lntrf.

```
abdikace     k1gFnSc2[^:]*::
abdikovat    k5eAaBmF::
Abdon    k1g[MI]nSc2[^:]*::a
abeceda    k1gFnSc2[^:]*:a:y
abecední     k2eAg.nSc1d1::
Abel     k1g[MI]nSc2[^:]*::a
Ábel     k1g[MI]nSc2[^:]*::a
abertamský     k2eAg[MI]nSc1d1::
Abertamy     k1g[MI]nPc2[^:]*:y:
Abidžan    k1g[MI]nSc2[^:]*::u
abidžanský     k2eAg[MI]nSc1d1::
Abigail    k1gFnSc2[^:]*::y k1gFnSc2[^:]*::
abiturientka     k1gFnSc2[^:]*:a:y
abiturient     k1g[MI]nSc2[^:]*::a
abiturientský    k2eAg[MI]nSc1d1::
abnormalita    k1gFnSc2[^:]*:a:y
abnormálně     k6eAd1::
abnormální     k2eAg.nSc1d1::
abnormálnost     k1gFnSc2[^:]*::i
```

Výpis D.3: Ukázka výpisu ze souboru ssc.lpn.

```
scenárista#husita#
scenáristický#africký#
scenáristka#aktovka#hP
scénář#stroj#
scenerie#abatyše#hT
scénický#otrocký#
scénka#aktovka#hT
scénografický#africký#
scénografie#abatyše#hT
scestí#stavení#
scestný#kupovaný#
scestovat#politovat#
scípat#dřímat#
scvrkat#klusat#rE
scvrklý#čilý#
sčesat#dokysat#
sčesávat#kousat#
sčetlý#čilý#
sčítanec#dělenec#
SČ#atd.#
SDA#atd.#
sdělení#stavení#
sdělit#dobavit#
sdělnost#kost#hT
sdělný#kupovaný#
```

Výpis D.4: Ukázka výpisu ze souboru ssc.derivations_from_definitions.

```
falzifikát -> falzum  1.9.7:ifikát:um
fanatizmus -> fanatismus  1.8.0:zmus:smus
fašizmus -> fašismus  1.8.0:zmus:smus
fatalizmus -> fatalismus  1.8.0:zmus:smus
fazole -> fazol 1.9.8:e:
Felicita -> Felicitas 1.9.5::s
férie -> féerie 1.1.3:rie:erie
fetišizmus -> fetišismus  1.8.0:zmus:smus
feudalizmus -> feudalismus  1.8.0:zmus:smus
Fidži -> Fidžijec 1.0.3::jec
      -> Fidžijka 1.0.4::jka
      -> fidžijský  2.0.11:Fidži:fidžijský
```

Výpis D.5: Ukázka výpisu ze souboru ssc.derivations.

```
nepořádek -> nepořádnost  1.1.1:ek:nost
          -> nepořádný  2.0.5:ek:ný
          -> nepořádně  6.0.0:ek:ně
Kmen: nepořád
Suffixy: ['ek', '', '', '']
Koncovky: ['', 'nost', 'ný', 'ně']

neposeda -> neposedný 2.0.5:a:ný
         -> neposedně 6.0.0:a:ně
Kmen: neposed
Suffixy: ['', '', '']
Koncovky: ['a', 'ný', 'ně']

nepraktický -> neprakticky  6.0.1:ý:y
            -> nepraktičnost  1.1.0:cký:čnost
Kmen: neprakti
Suffixy: ['', '', 'č']
Koncovky: ['cký', 'cky', 'nost']
```

Výpis D.6: Ukázka výpisu ze souboru ssc.derivations_from_definitions.cdb, shodný formát se ssc.derivations.cbd.

```
1.1.1:mědiryt#bez:měditina#slina#hT
1.1.1:menševizmus#komunismus:menševictví#stavení
1.1.1:oceloryt#návod:ocelorytina#slina#hT
1.1.1:opruzení#stavení:opruzenina#slina#hT
1.1.1:ořeší#stavení:ořešina#slina#hT
1.1.1:předurčení#stavení:předurčenost#kost#hT
1.1.1:rokle#abatyše#hT:roklina#slina#hT
1.1.1:slepenec#válec:slepenina#slina#hT
1.1.1:stezka#aktovka#hT:pěšina#slina#hT
1.1.1:tábor#doktor:táborita#husita
1.1.1:taškařice#dělnice#hT:taškařina#slina#hT
```