Univerzita Karlova v Praze
Filozofická fakulta
Ústav teoretické a komputační lingvistiky

Jaroslava Hlaváčová

Formalizace systému české morfologie s ohledem na automatické zpracování českých textů

Formalization of the Czech Morphology System with Respect to Automatic Processing of Czech Texts

Disertační práce

Studijní program: Filologie
Studijní obor: Matematická lingvistika

Vedoucí práce: Doc. RNDr. Vladimír Petkevič, CSc.

Praha 2009

Prohlašuji, že jsem disertační práci vykonala samostatně s využitím uvedených pramenů a literatury.

Poděkování

Děkuji svému školiteli, docentu Petkeviči, za vlídné vedení a trpělivost.

Děkuji paní profesorce Panevové za konzultace ohledně lingvistických termínů a za připomínky k první verzi práce.

Děkuji své rodině za veškerou podporu. Svému muži Vaškovi za ohleduplnost. Dětem za výchovné pobídky k dokončení práce.

Práce byla podpořena grantem Informační společnosti č. 1ET100120503 poskytnutým Grantovou agenturou AV ČR a grantem č. 100008/2008 poskytnutým Grantovou agenturou UK.

Shrnutí

Přesný morfologický popis slovních tvarů je prvním předpokladem pro úspěšné automatické zpracování jazykových dat.

Systém kategorií a jejich hodnot, které se k popisu používají, jsou náplní první části práce.

Základním principem je tzv. **Zlaté pravidlo morfologie**, které říká, že každý slovní tvar by měl být v systému popsán jednoznačně. Existence variant na úrovni slovních tvarů i celých paradigmat však splnění tohoto pravidla komplikuje. Koncept variant rozšiřujeme na tzv. **mutace**, mezi které řadíme i jiné množiny slovních tvarů se stejným popisem (např. víceré tvary osobních zájmen). Mutace dělíme na **globální** pro popis na úrovni paradigmat a **flektivní** pro popis jednotlivých slovních tvarů. Toto rozdělení nám umožňuje postihnout jejich časté kombinace. Upouštíme od dělení variant (mutací) podle stylového příznaku jako neobjektivního kritéria. Při důsledném využívání hodnot kategorií `Flektivní mutace` a `Globální mutace` zůstane Zlaté pravidlo morfologie vždy splněno.

V kapitole o lemmatizaci zavádíme **vícenásobné lemma** pro popis variantních lemmat.

Podrobně se zabýváme popisem tzv. **složenin**, tedy slovních tvarů typu *zač, proň, koupilas, koliks*. Pro jejich lemmatizaci rovněž využíváme konceptu vícenásobného lemmatu. Podle slovních druhů jejich složek je dělíme na několik typů. Zabýváme se též problémem jejich vyhledávání v jazykových korpusech.

Druhá část práce popisuje systém vzorů pro popis slovních tvarů jednotlivých slovních druhů. U každého vzoru uvádíme sadu parametrů, které umožní postihnout velkou variabilitu v tvoření konkrétních paradigmat. Věnujeme se i pravidelnému odvozování příbuzných slov pomocí sufixů.

Abstract

Detailed morphological description of word forms represents one of the most important conditions of a successful automatic processing of linguistic data.

The system of categories and their values which are used for the description are the subject of the first part of the thesis.

The basic principle, so-called **Golden rule of morphology**, states that every word form has to be described by the system unambiguously. The existence of variants of word forms and whole paradigms, however, complicates the accomplishment of this rule. We introduce so called **mutations** as an extension of the variants to be able to include other sets of word forms with the same description (for instance multiple word forms of Czech personal pronouns). We divide mutations into two parts — **global** ones describing all word forms of a paradigm, and **inflectional** ones for the description on the word form level. This division enables us to express their various combinations. We do not use features of style for the mutation division, for they are subjective.

With a consistent use of the categories called Inflectional Mutation and Global Mutation, the Golden rule of morphology will always be valid.

The concept of multiple lemma is introduced in a chapter dealing with lemmatization. It describes lemma variants.

We give a detailed description of so-called **compounds**, which incorporate word forms of the type *zač, proň, koupilas, koliks*. The concept of multiple lemma is also used for their lemmatization. According to the word class of their components we divide the compounds into several types. We also deal with the problem of their searching in language corpora.

The second part of the thesis describes a system of patterns for word description. It is divided according to the part of speech. Each pattern has a special set of parameters that allow to grasp a large variability in word formation. We also deal with regular derivations of related words using suffixes.

Obsah

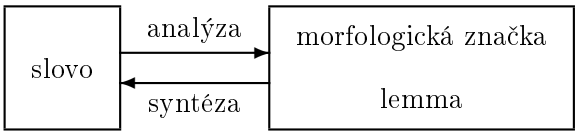
1 Úvod

Základem automatického zpracování jazyka je většinou rozsáhlý morfologický slovník, který popisuje slova daného jazyka. Slovník se využívá v zásadě k řešení dvou duálních úkolů. Prvním je rozpoznávání slov v textu, druhým vytváření slov (do textu).

Rozpoznání slova znamená jeho popis pomocí nějakých vlastností. Jednou ze základních vlastností každého slova je jeho základní tvar, neboli lemma, a slovní druh. Další vlastnosti se u flektivních jazyků potom liší, především podle slovního druhu. Každá vlastnost má nějaké hodnoty, které dohromady vytvářejí tzv. morfologickou značku. Rozpoznání slova tedy znamená určení lemmatu a morfologické značky, která toto slovo popisuje. Tomuto procesu se říká **morfologická analýza**[1]. Morfologická analýza je základem jakéhokoli dalšího (automatického) zpracování textu. Bez ní, a potažmo bez morfologického slovníku, se sotva obejde náročnější jazyková aplikace, ať už je to strojový překlad, rozpoznávání mluvené řeči, dialogové systémy, či další složité úlohy.

Vytváření slov jazyka, neboli **morfologická syntéza** nebo také **generování**, je opačný proces než morfologická analýza. Slova se generují na základě lemmatu a morfologické značky.

Vztah morfologické analýzy a generování ilustruje schéma na obrázku 1.1.



Obrázek 1.1: Dualita morfologické analýzy a generování

Morfologický slovník, o kterém jsme se zmínili hned na začátku, by měl obsahovat popis co nejvíce (nejlépe všech) slov jazyka. Způsob popisu může být různý, většinou se používá nějaký systém vzorů.

Pro český jazyk existuje několik automatických popisů, z nichž nejznámější a zejména v akademickém světě nejpoužívanější jsou systémy pražský a brněnský. Každý z nich používá vlastní morfologický slovník. Existuje snaha oba systémy sjednotit, což znamená nejen spojit oba slovníky, ale především jednoznačně definovat morfologické kategorie a jejich hodnoty. Výsledkem by měl být takový popis české morfologie, který bude co nejúplnější[2], ale zároveň nebude zbytečně přegenerovávat, tzn. nebude obsahovat slova v češtině neexistující.

Přesto, že oba zmíněné systémy již léta slouží lingvistům (zejména v podobě automaticky označkovaných jazykových korpusů) i informatikům (jako základ

[1] Někdy se za morfologickou analýzu považuje jen přiřazení morfologických značek.

[2] Vzhledem k obecné povaze jazyka (a nejen českého) nemůže být úplný nikdy. Otázka, co do jazyka patří a co už nebo ještě ne, je velmi subjektivní.

nejrůznějších aplikací), stále je co vylepšovat. Tato práce je pokusem o takové vylepšení.

Naším cílem tedy je vytvořit rámec pro popis slovních tvarů českého jazyka, tzn. přesně definovat kategorie, které se k popisu používají, a stanovit pravidla, podle kterých se slovním tvarům přiřazuje jejich základní tvar, čili lemma. Zdůrazňujeme, že cílem je popis, nikoli vysvětlování ani zdůvodňování jazykových jevů. Tam, kde to je možné, používáme samozřejmě zavedených a odůvodněných lingistických popisů, občas jsme však v zájmu jednoduchosti a jednoznačnosti popisu odhlédli od lingvistických hledisek a použili lingvisty neoblíbené „technické“ řešení.

Nezabýváme se tedy konkrétními „slovíčky“. Navrhujeme systém, který nejen umožní přesný popis pravidelných morfologických jevů, ale bude schopen konzistentně pojmout i výjimky. Na některé upozorňujeme v textu, ale zcela jistě ne na všechny.

Při popisu jednotlivých morfologických kategorií budeme vycházet ze závěrů tzv. Morfologického konkláve (dále jen Konkláve), které se sešlo 21.–23. října roku 2005, aby definovalo jednotlivé morfologické kategorie (viz kopie účastnického slibu v příloze B na straně 144). Jeho závěry však nikdy nebyly dovedeny ke zdárnému konci. Složení Morfologického konkláve bylo (podle abecedy, bez titulů): Jan Hajič, Jaroslava Hlaváčová, Karel Oliva, Klára Osolsobě, Karel Pala a Vladimír Petkevič. Dokument, který na jeho podkladě vznikl, nebyl nikdy publikován. Budeme se snažit pečlivě oddělit výsledky Konkláve od vlastních řešení, ať už pouze doplňujících, nebo zcela odlišných.

Jsme si vědomi toho, že některé kategorie, které používáme, mají spornou definici. Někdy záměrně směšujeme více jevů do jedné kategorie (např. mutace), jindy naopak rozdělujeme zavedené kategorie a jejich hodnoty (např. kategorie `Číslo` a kategorie `Duál`). Primárním důvodem je vytvořit systém kategorií takový, aby se pomocí jejich hodnot daly popsat slovní tvary alespoň v takovém rozsahu, v jakém se dají popsat pomocí zmiňovaných systémů nyní.[3] Navíc se snažíme odstranit nedostatky, které oba systémy mají a jichž si jsou jejich uživatelé, ale i tvůrci, vědomi. Snažíme se veškeré návrhy dobře popsat a zdůvodnit. K tomu využíváme příkladů z korpusu, z internetu i z vlastních pozorování. Příklady jsou průběžně číslovány. Není-li uvedeno jinak, jsou z korpusu SYN.

1.1 Základní definice

Klíčovými termíny pro popis systému morfologie jsou: slovní tvar, lemma, morfologická značka. Ač jsou tyto koncepty na první pohled jednoduché, při podrobné práci s nimi se dostáváme do situací, které nelze řešit bez pečlivě rozmyšlených definic a pravidel.

Slovo je řetězec písmen, který je na začátku i na konci ohraničen oddělovačem.

Oddělovač je řetězec nealfanumerických znaků. Většinou jde o mezery, o jiné tzv. „bílé znaky“ (white spaces) a znaky z množiny {(),.;':!?„“<>}, případně

[3] Výjimkou jsou pražské kategorie `Přivlastňovací číslo` a `Přivlastňovací rod`, které se patrně nevyužívají, a hodnota „Rodina“ kategorie `Rod` používaná v Brně, která by si zasloužila podrobnější rozbor.

další znaky. Oddělovače nejsou součástí slova, řadíme je do **interpunkce**. S interpunkcí se v českých korpusech zachází jako se zvláštním typem slovního tvaru. V pražském systému tvoří jednu z hodnot kategorie `Slovní druh`, podle níž ji lze vyhledávat. Interpunkce je zpracována dostatečně, není proto důvod se jí znovu zabývat[4].

Často se uvádí, že slovo musí mít nějaký význam. Nepovažujeme tento požadavek za zásadní, a to ze dvou důvodů:

1. V textu se můžeme setkat s řetězci písmen, jejichž významu nerozumíme, buď proto, že žádný význam nemají, nebo proto, že ho nechápeme. Může jít třeba jen o překlepy (známá chyba vzniklá posunutím ruky písařky na klávesnici), které učiní dané slovo nesrozumitelným. Chceme-li o takovém řetězci mluvit, používáme i v tomto případě termínu slovo. Slova jsou i záměrně zašifrované řetězce.

2. Existují věty, i správně syntakticky utvořené, které nemají žádný význam, např. *Rychlý strom respektuje rozpustnosti.*, abychom nepoužívali stále stejnou Chomského větu o zelených myšlenkách. Podle Patricka Hankse a jeho teorie „Corpus Pattern Analysis“ (viz (Hanks – Pustejovsky, 2004)) nemají jednotlivá slova význam, ten dostávají až v kontextu, ve kterém se ocitají (viz též lexikální sémantika Cruse v (Cruse, 1986)). Slova v nesmyslném kontextu tedy mají nesmyslný význam, což se dá také říci tak, že význam nemají žádný.

Poznámka k významu slova „význam“ Už v úvodu, při první definici, jsme se dostali k termínu „význam“, který vůbec není jednoduchý. Mnoho lingvistů, filosofů, matematiků i jiných odborníků se pokoušelo význam definovat. Hned na začátek předesíláme, že se významem slova „význam“ nebudeme zabývat. To ovšem neznamená, že ho nebudeme používat. Naopak, používat ho musíme, neboť bez něj bychom nebyli schopni definovat další pojmy. Termín význam budeme tedy chápat velmi intuitivně takto: Slovo má význam, jestliže existuje kontext, ve kterém něco označuje. Slova z předchozí poznámky 2 tedy přece jen v tomto smyslu význam mít mohou.

Slovní tvar je slovo, které má význam.

Rozlišujeme tedy slovo jako řetězec písmen a slovní tvar jako řetězec písmen s významem. Tím se lišíme např. od definice Havránka a Jedličky (Havránek – Jedlička, 1981): „Slovo je skupina hlásek, která má zřejmý význam.“ Za slovní tvary nepovažujeme interpunkční znaménka. Ta patří mezi oddělovače.
Množinu slovních tvarů jazyka budeme značit S.
Příklady: *nějakou, bývala, stolečku.*

Lemma je základní slovní tvar. Ve slovnících se používá jako slovníkové heslo.
Množinu lemmat jazyka budeme značit L.
Příklady z předchozího odstavce mají lemmata *nějaký, bývat, stoleček.*

[4] Uvažujeme o vytvoření klasifikace funkcí jednotlivých interpunkčních znamének.

Všechny slovní tvary, které lze vytvořit z jednoho lemmatu pomocí skloňování, časování nebo stupňování (obecně ohýbání), tvoří tzv. **paradigma**[5]. Můžeme také říci, že paradigma je množina slovních tvarů, které náležejí danému lemmatu. Na rozdíl od většiny klasických mluvnic zahrnujeme do paradigmatu i nespisovné (nekodifikované) slovní tvary.

Morfologická kategorie je vlastnost slovních tvarů.

Každá morfologická kategorie má předem definovanou konečnou množinu hodnot, kterých může nabývat. Jestliže pro nějaký slovní tvar daná morfologická kategorie nenabývá žádné hodnoty, řekneme, že tato kategorie není pro tento slovní tvar relevantní. Můžeme také říci, že tato kategorie není relevantní pro celé paradigma nebo pro dané lemma jako jeho reprezentanta. A konečně, relevantnost morfologických kategorií se může týkat celých tříd lemmat se společnými vlastnostmi.

Morfologickou kategorii, která popisuje množinu slovních tvarů, budeme nazývat **relevantní morfologickou kategorií** této množiny. Množina může být jednoprvková, tedy jeden konkrétní slovní tvar, nebo víceprvková. V tom případě jde většinou o celé lemma, množinu lemmat stejného slovního druhu, případně i poddruhu. Obecně může jít o množinu libovolnou.

Hodnota kategorie, která není relevantní pro daný slovní tvar, lemma nebo třídu lemmat, má nedefinovanou hodnotu (UNDEF), tedy např. stupeň podstatného jména má hodnotu UNDEF.

Jako příklad morfologické kategorie uveďme slovesný vid, který má tři hodnoty: dokonavý, nedokonavý, obouvidý. Tato kategorie je relevantní pro slovesa a deverbativa. Není relevantní např. pro předložky.

V našem návrhu budeme hodnotám morfologických kategorií přiřazovat kódy, a to tak, aby byly co nejvíce v souladu s kódy používanými v současných morfologických systémech, ne vždy však bude možné shodu dodržet.

K popisu slovního tvaru je třeba většinou více morfologických kategorií. Jejich kódy potom vytvářejí morfologickou značku. Mluvíme-li tedy o morfologické značce, máme na mysli hodnoty relevantních morfologických kategorií daného slovního tvaru.

Hodnoty kategorií se určují v závislosti na ostatních kategoriích, především na kategorii slovního druhu a poddruhu, viz tabulku 4.7 na str. 52.

Za morfologické značky však považujeme jen takové řetězce, které kódují hodnoty všech relevantních morfologických kategorií pro daný slovní tvar (např. rod, číslo a pád pro podstatná jména). Takto chápaná morfologická značka vlastně popisuje slovní tvar obecně (např. podstatné jméno rodu ženského ve třetím pádě jednotného čísla), ve spojení s lemmatem potom popisuje konkrétní slovní tvar. Jinými slovy, máme-li lemma a morfologickou značku, můžeme vytvořit jednoznačně slovní tvar (tzv. Zlaté pravidlo morfologie, viz oddíl 1.2).

Značky, které jsou podspecifikované, tedy nemají vyplněné všechny relevantní kategorie, nepovažujeme za morfologické značky. Podle takové značky bychom totiž nebyli schopni pro dané lemma vygenerovat jednoznačný slovní tvar. Např. hodnota morfologické kategorie `Číslo` pro lemma *jarní* vygeneruje

[5] Termín paradigma se někdy používá ve významu „vzor“. My tyto dva termíny rozlišujeme (paradigma a vzor).

množství tvarů, které mají různé rody, pády, stupně. Zakódování této hodnoty samostatně tedy pro nás není morfologickou značkou[6].

Takto tedy vypadá definice morfologické značky:

Morfologická značka je řetězec znaků, který kóduje hodnoty všech relevantních morfologických kategorií pro nějaký slovní tvar nějakého lemmatu.

Množinu morfologických značek budeme značit M. Této množině se říká (morfologický) **tagset**, my se ale budeme snažit tomuto převzatému termínu vyhnout.

Hodnoty morfologických kategorií popisují slovní tvary, proto můžeme i o morfologické značce říci totéž.

Dva nejpoužívanější české systémy morfologických značek jsou systém pražský (viz (Hajič, 2004)) a brněnský (viz (Sedláček, 1999)). O jejich přednostech, záporech i rozdílech mezi nimi se už hodně mluvilo. Nejpřehlednější porovnání provedla Klára Osolsobě (Osolsobě). Budeme se o nich zmiňovat jen tehdy, jestliže bude nutné poukázat na nějaké rozdíly mezi řešením navrhovaným a již existujícím.

Pražský morfologický systém používá tzv. pozičního systému značek, kde každá pozice kóduje určitou kategorii, i když je pro daný slovní tvar nerelevantní, brněnský systém používá kompaktní značky uvádějící kódy jen relevantních morfologických kategorií. Oba systémy však mohou být ekvivalentní (bohužel nejsou — to ale není chyba kódování).

Při popisu morfologických kategorií a jejich hodnot zavádíme kódy především proto, abychom s nimi mohli v této práci dále pracovat, zejména pomocí nich vytvářet dotazy.

V kapitole 4.3 navrhujeme způsob, jak vytvořit morfologickou značku. Většinou však v celé práci pracujeme jen s hodnotami jednotlivých kategorií, protože konkrétní tvar morfologické značky není podstatný. Podstatné je pouze to, aby obsahovala všechny relevantní hodnoty.

Lemmatizací rozumíme zobrazení, které každému slovnímu tvaru přiřadí množinu jeho lemmat (viz (Hajič, 2004)):

$$\lambda: S \rightarrow 2^L$$

kde S je množina slovních tvarů a L množina lemmat.

Obvykle se lemmatizací rozumí zobrazení, přiřazující slovnímu tvaru jeho (jedno) lemma. Existují však lemmata, která jsou navzájem ortografickými variantami (*univerzita* — *universita*). V takových případech sdružujeme varianty pod společné lemma. Dalším důvodem k zavedení vícenásobných lemmat jsou tzv. složeniny[7], které zavádíme v kapitole 4 jako zvláštní slovní druh. Složeniny nemají jednoduché lemma, protože jsou složeny z více slov, každé s jiným lemmatem. Určit jednoslovné lemma slovních tvarů jako *zač*, *každémus* nebo *proňs*

[6] Přesto se částečné kódování jen některých kategorií může pro určité aplikace hodit. Často používané je značkování pouze podle slovního druhu.

[7] Tímto termínem nemyslíme kompozita (slova složená), ta lemmatizujeme jako jedno slovo, neboť to nepřináší žádné problémy, např. *spolupořadatel*, *koupěschopný*. Termín se může zdát nevhodný, lepší jsme však ani po konzultacích s odborníky nevymysleli. Konkláve používalo termín ještě nevhodnější.

není jednoduché. Zde dobře poslouží množina lemmat jednotlivých složek složeniny, tedy vícenásobné lemma. O složeninách více v kapitole 6, o lemmatizaci v kapitole 2.

Morfologická analýza je zobrazení, které každému slovnímu tvaru přiřadí množinu dvojic ⟨lemma, morfologická značka⟩:

$$\mu: S \rightarrow 2^{L \times M}.$$

Lemmatizace je tedy součástí morfologické analýzy.

U homonymních slovních tvarů dostáváme dvě i více lemmat, např.

$$\lambda(pekla) = \{peklo,\ péci\}.$$

Každé z lemmat navíc může (a v uvedeném příkladě to tak opravdu je) být členem více než jedné dvojice ⟨lemma, morfologická značka⟩.

1.2 Zlaté pravidlo morfologie

Fakt, že jednomu slovnímu tvaru přiřadí zobrazení μ více různých hodnot, nevadí. Mnohem více vadí v jistém smyslu duální skutečnost, že jedné dvojici ⟨lemma, morfologická značka⟩ může odpovídat více než jeden slovní tvar. Např. 6. pád podstatného jména *hrad* v jednotném čísle může být slovní tvar *hradu* i *hradě*.

Bylo by výhodné, kdyby každá dvojice ⟨lemma, morfologická značka⟩ jednoznačně popisovala nejvýše jeden slovní tvar. Tomuto požadavku říkáme Zlaté pravidlo morfologie.

Při různých automatických aplikacích, které využívají generování slovních tvarů, je totiž těžké rozhodování, která z variant se má vybrat. Příkladem takové aplikace je třeba strojový překlad do češtiny, který v určité fázi musí vybírat v cílovém jazyce správný slovní tvar. Jestliže lemma i morfologická značka jsou pro dva tvary stejné, zodpovědný výběr je prakticky nemožný.

Je tedy třeba popis variantních slovních tvarů rozšířit tak, aby dvojice ⟨lemma, morfologická značka⟩ byla pro každý slovní tvar jednoznačná.

To lze udělat několika způsoby. Je možné zahrnout informaci o variantách do lemmatu nebo do morfologické značky, nebo vyčlenit tuto kategorii jako další atribut slovního tvaru. Poslední řešení jsme zvolili my. Z důvodů, které uvedeme dále v kapitole 3, jsme tuto kategorii nazvali `Mutace`.

Zlaté pravidlo morfologie tedy vypadá schematicky takto:

lemma + morfologická značka + mutace = jednoznačný slovní tvar

2 Lemma a lemmatizace

Základní jednotkou morfologického slovníku je lemma, které zastupuje celé paradigma slovních tvarů.

Lemmatizaci chápeme jako zobrazení λ z množiny slovních tvarů do množiny lemmat. Již v úvodu jsme naznačili, že zobrazení λ obecně nepřiřazuje jediné lemma, ale množinu lemmat: $\lambda: S \rightarrow 2^L$, kde 2^L označuje množinu podmnožin množiny L.

Každému slovnímu tvaru přiřazuje zobrazení λ alespoň jedno lemma (např. $\lambda(okna) = \{okno\}$). Případy, kdy toto zobrazení není jednoznačné, nejsou v češtině řídké (např. $\lambda(pekla) = \{peklo,\ péci\}$). Je to způsobeno vysokou slovnědruhovou a morfologickou homonymií českého jazyka. Homonymie se bez kontextu (a někdy ani s ním) zbavit nelze.

2.1 Vícenásobné lemma

Kromě homonymie však existuje ještě jeden problém, který s lemmatizací souvisí. Jsou to varianty. Vezměme si např. slovní tvary *diskuze* a *diskuse*. Máme je analyzovat jako dvě různá lemmata, nebo varianty lemmatu jednoho?

Tato otázka má závažné praktické pozadí. Jestliže bude např. uživatel korpusu vyhledávat slovní tvary lemmatu *diskuze*, mají se zobrazovat jen slovní tvary se *-z-*, nebo i ty se *-s-*? S problémem se potýká i syntéza. Podle čeho se má z více vygenerovaných slovních tvarů se stejnými charakteristikami vybrat jeden?

Dosavadní morfologické slovníky se tímto problémem příliš nezabývají, a tak můžeme nalézt varianty na úrovni lemmatu, které jsou v pražském morfologickém slovníku zahrnuty pod jedno společné lemma (např. varianty *diskuze* i *diskuse* mají jediné společné lemma *diskuse*), i takové varianty, které jsou rozlišeny jako dvě různá lemmata (*citron* a *citrón*).

Ideální by bylo, kdyby lemma vždy odpovídalo slovnímu tvaru, ale kdyby se zároveň všechny varianty jednoho lemmatu sdružily, aby se daly například snadno vyhledat v korpusech. Toho lze dosáhnout zavedením konceptu vícenásobného lemmatu. Vícenásobným lemmatem z našeho příkladu jsou tedy dvouprvkové množiny $\{diskuze,\ diskuse\}$ a $\{citron,\ citrón\}$. Prvkům této množiny budeme říkat **variantní lemmata**. [1]

Vícenásobné lemma zavádíme i pro taková variantní lemmata, která jsou nespisovná, zastaralá nebo jinak příznaková. Máme tedy např. i vícenásobné

[1] Vícenásobnými lemmaty se zabýval také Karel Kučera (viz (Kučera, 2007)), ovšem z diachronního hlediska. Vzhledem ke změnám pravopisu slov v průběhu dějin potřeboval sdružit slova v různých etapách vývoje jejich zápisu. Na rozdíl od našeho řešení však zvolil tzv. hyperlemma jakožto zástupce množiny (historických) lemmat se stejným významem. Jeho hyperlemma je jediné a vybírá se ze současné slovní zásoby (pokud takové příslušné lemma existuje). Naše vícenásobné lemma je pohled na stejnou problematiku z hlediska synchronního.

lemma {*otevřít, votevřít*}, {*okénko, okýnko, vokýnko*}, {*blůza, bluza, blusa, blůsa, blúza*}.

Množina lemmat přiřazená homonymním slovním tvarům však vícenásobné lemma není, přestože je to také výsledek zobrazení λ.

Vícenásobné lemma je množina lemmat se stejným významem lišících se pouze zápisem (ortografické varianty).

Pro vícenásobné lemma definujeme ještě tzv. **rozšířené paradigma** jako sjednocení paradigmat jednotlivých variantních lemmat.

Kdykoli budeme v následujícím textu mluvit o lemmatu, budeme mít na mysli i případné vícenásobné lemma, pokud neuvedeme jinak. Pojem lemma tedy rozšíříme i na vícenásobná lemmata. Je to v souladu s naší definicí lemmatizace, tedy zobrazení λ, které nepřiřazuje slovním tvarům jednotlivá lemmata, ale množiny lemmat (i jednoprvkové).

Přesto však v případě, že lemma není vícenásobné, nebudeme nadále pro jeho vyjádření používat množinový zápis. Tedy např. $\lambda(polévkou) = polévka$. Stejně tak upustíme od množinového zápisu v případě lemmatu sloves, které, bráno zcela striktně, je také vždy vícenásobné, neboť infinitiv má vždy dva tvary, jak ukazuje příklad s vícenásobným lemmatem {*péci, péct*}.

Vícenásobné lemma jakožto množina lemmat přiřazená jednomu slovnímu tvaru poslouží i v případě složenin, u nichž není možné přirozeně jednoznačně a jednoduše zavést základní slovní tvar. Blíže se budeme lemmatizaci složenin věnovat v kapitole 6 o složeninách.

Pro složeniny tedy definujeme vícenásobné lemma poněkud odlišně:

Vícenásobné lemma složeniny je množina lemmat jednotlivých složek složeniny.

2.2 Vybrané problémy lemmatizace

2.2.1 Lemmatizace sloves

Lemmatem slovesa je jeho infinitiv. Toto jednoduché tvrzení je třeba podrobit detailnějšímu zkoumání pro některé speciální jevy.

2.2.1.1 Zvratná slovesa

Spory vzbuzuje dosavadní praxe přiřazovat infinitiv vždy bez zvratné částice, a to i v případě, že se jedná o reflexivum tantum. Námitka, že *smát* je nesmyslné slovo, je jistě správná, lemma by mělo být *smát se*, tedy dvě slova.

Taková lemmatizace by však mohla znamenat komplikace. V tom případě by totiž bylo logické, aby se stejné lemma přiřadilo nejen slovním tvarům paradigmatu *smát*, ale i vlastní zvratné částici. Pomiňme nyní nesmírně obtížné rozpoznávání zvratné částice, která ke slovesu patří, ve složitějších kontextech, kdy mohou být obě části zvratného slovesa od sebe vzdáleny, a to libovolným počtem slovních tvarů, dokonce na obě strany. Pomiňme i z toho vyplývající fakt, že takováto lemmatizace by nebyla možná v okamžiku morfologické analýzy, ale až po nějaké formě desambiguace, což by v důsledcích pravděpodobně znamenalo potřebu radikálně změnit zavedený postup automatického zpracování textů.

Ani samo přiřazení slovesného lemmatu zvratné částici se nám však nejeví jako rozumné z hlediska automatického zpracování a posléze využívání lemmatizovaných korpusů. Kdybychom totiž důsledně zařazovali zvratnou částici do lemmatu příslušného slovesa, bylo by potom asi také třeba obě části sdruženého lemmatu jednotně značkovat, tedy přiřadit i zvratné částici hodnoty slovesných kategorií osoby, čísla, vidu, slovesného tvaru a dalších (viz kap. 4). Slovní poddruh zvratné částice by potom ztratil význam a neznačkoval by se, což by mohlo přinést problémy při jejich vyhledávání v korpusech. Navíc takové značkování není obvyklé. Samozřejmě není třeba držet se zaběhaných postupů, jestliže zjistíme, že už nevyhovují. V tomto případě však takový krok není nutný.

Je tu i možnost, že by se zvratná částice značkovala nezávisle na zbytku zvratného slovesa, přičemž toto sloveso by se lemmatizovalo i se zvratnou částicí, tedy např. $\lambda(směje) = \{smát\ se\}$ a $\lambda(se) = \{se\}$. Tato alternativa nám připadá nekonzistentní, a navíc zbytečná. Tím, že do lemmatu přidáme další slovo, totiž zvratnou částici, pouze upozorňujeme na fakt, že jde o reflexivum tantum (jiná reflexiva by se, vzhledem k jejich nejasným vymezením, jako reflexiva značit neměla). Vyhledávání lemmatu *smát se* v korpusech by stejně mohlo přinést jako výsledek pouze tvary slovesa *smát*, bez zvratné částice, a to ani kdyby se nacházela v těsné blízkosti. Takové řešení nám připadá nelogické a zbytečně složité.

Mnohem jednodušší je považovat lemma za technickou entitu, která reálně v jazyce nemusí existovat, ale která slouží k identifikaci celého paradigmatu. Přestože tedy samostatný tvar *smát* neexistuje, je možno pracovat s tímto tvarem jako s lemmatem na morfologické rovině. Teprve další roviny jazykového popisu sdruží jednotlivé slovní tvary ve správné spojení *smát se*.

2.2.1.2 „Stupňování“ sloves

Mnoho nedokonavých (avšak ne iterativních) sloves má schopnost spojovat se s některými speciálními předponami a se zvratnou částicí *se* nebo *si*, a tím vytváří celé paradigma nových slovních tvarů s poměrně přesně definovaným významem. Předpony, k nim příslušející zvratné částice a význam celého prefigovaného slovesa ukazuje tabulka 2.1.

Předpona	Sloveso	zvratná částice	Význam
roz-	X	*se*	začít X
po-	X	*si*/*se**	X v klidu, většinou příjemně
za-	X	*si*/*se**	X po delší dobu a užít si to
na-	X	*se*	hodně X
vy-	X	*se*	hodně X a být s tím spokojen
u-	X	*se*	X až do vyčerpání

Tabulka 2.1: Přehled stupňovacích slovesných předpon. Hvězdička (*) v tabulce znamená, že pokud jde reflexivum tantum, zůstává i po přidání těchto prefixů zvratná částice *se*.

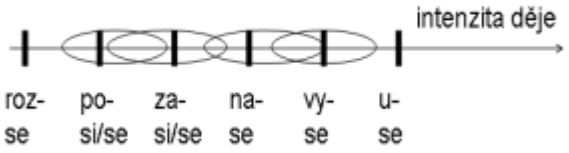
Dosadíme-li v tabulce místo X např. sloveso *mávat*, dostaneme sadu nových paradigmat.

Uvádíme vymyšlené příklady uspořádané do malého příběhu, ze kterého by měl být zřejmý význam:

Včera se na nádraží pořádal kompars na nový film. Měli jsme za úkol mávat na odjíždějící vlak. Když dal režisér pokyn, rozmávali jsme se. Nejdřív to vypadalo, že si pomáváme a půjdeme domů. Scéna s máváním se však mnohokrát opakovala, takže jsme si zamávali víc, než se nám líbilo. Namávali jsme se opravdu hodně, vymávali jsme se do sytosti. Měli jsme strach, že se umáváme k smrti. (1)

Tímto způsobem jsme vytvořili sadu paradigmat s povinnou zvratnou částicí.

Jednotlivá prefigovaná zvratná slovesa je možno s určitou tolerancí uspořádat podle intenzity děje (viz obr. 2.1).



Obrázek 2.1: Osa s naznačenou posloupností předpon podle stupňující se intenzity děje (zleva doprava).

Krajní body tvoří předpony *roz-* a *u-*, uprostřed je podle intenzity posloupnost *po-, za-, na-* a *vy-* s vágním až překrývajícím se rozsahem. Z tohoto důvodu nazýváme tento způsob tvoření s jistou nadsázkou pracovně „stupňování" intenzity slovesného děje.

O všech těchto případech píše Kopečný v (Kopečný, 1962b), v poslední kapitole o českých slovesných předponách, není zde však učiněn závěr o „stupňování". Uvádíme stručné charakteristiky z tohoto textu, u každého několik příkladů z korpusu SYN na dokreslení. Zmiňuje se o nich i Šmilauer v (Šmilauer, 1971), který některá tato slovesa řadí do skupiny sloves vyjadřujících velkou „míru děje".

2.2.1.2.1 *roz-* O jednom z významů předpony *roz-* Kopečný píše: „rozproudění činnosti, obyčejně až po dosažení náležité míry".

Někde v tom mlází se znova rozťukal datel (2)
Promnul si prsty, samým vzrušením se mu rozbrněly. (3)

2.2.1.2.2 *po-* Krátce si všímá i námi uvedeného významu předpony *po-*, o kterém píše, že může znamenat „i velkou míru děje", což je jen částečně ve shodě s naším pozorováním.

místečko ve stínu, kde by si každý druhý pejsek rád pěkně pochrupkal (4)
pohrál si se startovacími klapkami na křídlech (5)

2.2.1.2.3 *za-* Kopečný se zmiňuje i o předponě *za-*, jíž přiřazuje „význam vzplanutí děje, jeho začetí“ a spojuje ji „s pocitem malé míry“, což nám nepřipadá zcela přesné, viz příklad ve vymyšleném příběhu. Ani ostatní uvedené příklady ze současné češtiny (6)–(8) tomuto hodnocení nenasvědčují.

o jejich tématech si budou moci rovnou i zachatovat. (6)
Poté zašel do tělocvičny, aby si podle svého rituálu ještě zaposiloval. (7)
Na druhou stranu : trochu si čas od času zašílet v takovéto nevinné záležitosti patří v této nespravedlností a stresem napěchované době skoro k léčebným procedurám. (8)

2.2.1.2.4 *na-* Výrazům s předponou *na-* (*namodlit se, nasmát se* apod.) říká Kopečný „augmentativnost“, případně „také intenzitivnost“.

Člověk se hrozně naběhá. (9)
matinka zatim doma vykládala, co se nastará a naběhá,... (10)
Co jsem se jen natančila to léto (11)

2.2.1.2.5 *vy-* U předpony *vy-* uvádí ještě případ zdvojené předpony *vyna-*, která však nezapadá zcela do naší pomyslné škály.

vyplačte se do sytosti (12)
V pohodě se tam celé dny můžeme do sytosti vyjezdit (13)

2.2.1.2.6 *u-* Předponu *u-* hodnotíme shodně jako Kopečný, když říká: „Reflexivní typ *upracovat se* je téměř paradigmatický.“ My považujeme za paradigmatické všechny právě vyjmenované typy.

To by se asi norští fanoušci uslavili k smrti. (14)
jak bylo zjištěno, unudit se nikdo nemůže (15)
málem se uštěkal (16)

Přestože se pomocí prefixu a reflexiva vytvoří nové sloveso s celým paradigmatem, umisťujeme tato slova do paradigmatu neprefigovaného (základního) slovesa. Pokládáme zde prefixaci za tvoření slovesného tvaru, nikoli za slovotvorbu. Jinými slovy: tvrdíme, že všechny takto utvořené tvary mají společné lemma, a to X, tedy v našem příkladě *mávat*.

Důvodů je hned několik.

Předně je to velká produktivita. Není sice pravda, že takto lze vytvářet celou sadu od každého nedokonavého slovesa (protipříkladem budiž třeba tvar **zadotýkat se*), přesto lze takto vytvořit velké množství „nových“ slov. Navíc předpony v těchto slovech mají VŽDY stejný význam. Tento význam je naznačen v tabulce 2.1 a vyplývá též z příkladu (1).

Kdybychom tyto slovní tvary lemmatizovali jako samostatné předponové sloveso, museli bychom pro každé takové slovo zavést nové lemma, někdy homonymní s lemmatem již existujícím.

Některá takto vytvořená slova již ve slovní zásobě existují, ale mají jiný význam. V našem příkladě je to tvar *zamávat*, ovšem bez zvratné částice. Příkladem úplné homonymie, včetně zvratné částice, je *vysmát se*. Běžný význam je zřejmý z příkladu (17):

Budu na něj hodná a on se mi pak vysměje. (17)

Ale vyskytují se i příklady ve významu, který popisujeme zde:

Stavil jsem se tady jen proto, že se tady člověk může v klidu vysmát. (18)

I z těchto případů je zřejmý rozdíl významu. V první větě jde o *výsměch*, zatímco ve druhé o *smích*. Další rozdíl spočívá ve valenci. Zatímco v prvním příkladě jde o sloveso s akuzativní valencí, druhý příklad je intranzitivní. To však není pravidlem v jiných případech.

Podobné je *usmát se*, které může vypovídat buď o *úsměvu*, nebo opět o *smíchu*. I zde je rozdíl ve valenci (*usmát se* na koho — *usmát se* (bez valence)).

Výskyt stupňované intenzity není zpravidla vysoký, ale najdou se výjimky. Některá takto vytvořená slova jsou naopak velmi běžná, i s uvedeným významem, např. *rozesmát se*. V těchto případech je ovšem rozumné předponovou odvozeninu přímo zahrnout do slovníku.

Rozhodnutí, která prefigovaná (stupňovaná) slovesa jsou běžná, lexikalizovaná, a kde jde jen o okazionalismy, je samozřejmě velmi obtížné. Pravděpodobně v tom nebude panovat shoda, navíc se názory budou měnit v čase. Pro současný morfologický slovník je nejspíš nejlepší konzervativní řešení, tedy ponechat ve slovníku ta paradigmata, která tam jsou, včetně zavedené lemmatizace, ale nepřidávat globálně nová. Vycházíme z toho, že současný slovník již naprostou většinu běžných slov obsahuje.

Rozhodně do morfologického slovníku patří ta slovesa, která nejsou zvratná nebo jsou tranzitivní, a dále potom tzv. odvozená reflexiva. Např. *utancovat se* je odvozené reflexivum od nezvratného tranzitivního slovesa *utancovat* (koho). Lemma *utancovat* tedy do morfologického slovníku zahrnujeme.

Na závěr této kapitoly je třeba dodat, že naznačené uchopení zvratných sloves s předponami *roz-, po-, za-, na-, vy-* a *u-* je třeba zpracovat ještě mnohem důkladněji. Tato kapitola nechť slouží jen jako upozornění na velmi pravidelný, byť poměrně řídký úkaz ve vyjadřování expresivity sloves.

2.2.2 Záporná lemmata

Tato kapitolka se týká sloves, přídavných jmen a příslovcí. Začneme slovesy.

Předpona záporu *ne-* se obvykle při lemmatizaci odstraňuje, takže lemma je kladné, např. $\lambda(nenechal) = nechat$. Existují však slovesa, která se v kladném smyslu vůbec nepoužívají. Mohlo by se zdát, že v těchto případech by mělo smysl ponechat lemma záporné. Podívejme se však na tyto případy blíže.

Vezměme si příklad slovesa *nedutat*, které je lemmatizováno jako *dutat*. Tvrdíme, že tato lemmatizace je správná, přestože se sloveso *dutat* v kladném smyslu téměř nevyskytuje. Pro to, zda lemmatizovat sloveso se záporkou či nikoli, navrhujeme test pomocí opisného budoucího času[2]:

[2] Tento test je možný jen u nedokonavých sloves, která tvoří budoucí čas s pomocným slovesem *být* (*bude*,...).

V případě, že zápor budoucího času daného slovesa převede záporku *ne-* k pomocnému slovesu, je třeba lemmatizovat sloveso kladně. V opačném případě tvrdíme, že kladný význam skutečně neexistuje.

Sloveso *nedutat*: říkáme *nebudu dutat*, nikoli **budu nedutat*. Zde je lemmatem *dutat* bez záporky.

Komu se nelení, tomu se zelení. (19)

Lemma z příkladu (19) je *lenit*, nebo *nelenit*? Budoucí čas: *nebude lenit*, nikoli **bude nelenit*. Lemma je tedy *lenit*, a to přesto, že se v kladném smyslu sloveso *lenit* téměř neužívá.

Opačným příkladem je sloveso *nenávidět*, kde nelze říci **nebudu návidět*, vždy jen *budu nenávidět*. Zde je lemmatem *nenávidět* se záporkou. Podobné je sloveso *nedoslýchat*, kde sice máme sloveso *doslýchat se*, ovšem s odlišným významem, navíc zvratné.

Stejně jako existují záporná slovesa, existují i záporná přídavná jména a příslovce, v menší míře i záporná podstatná jména. U nich neumíme vytvořit podobně jednoduchý test, jako u sloves test s budoucím časem. Pouze přídavná jména, podstatná jména a příslovce od sloves odvozená převezmou záporné nebo nezáporné lemma podle (ne)zápornosti základního slovesa.

V případě, že lemma se záporkou má odlišný význam než zápor lemmatu bez záporky, je lemmatem tvar se záporkou. Např. lemma *nesmyslný* není záporem lemmatu *smyslný*, jde tedy o dvě lemmata — *smyslný* a *nesmyslný*. První z nich teoreticky může přibrat záporku a popisovat někoho, kdo není smyslný, druhé však už se záporkou možné není (**nenesmyslný*).

Bohužel však existují slova, u nichž nepanuje shoda. Např. příslovce *kale*, které se běžně používá ve východních Čechách (viz příklad (20)), mnozí mluvčí jako nezáporné vůbec neznají a tvrdí, že jediné přípustné použití je se záporkou, tedy *nekale* (příklad (21)).

Kale tomu nerozuměl. (odposlechnuto) (20)
Obě lupičské tlupy si nekale konkurovaly... (21)

Sporné je např. i podstatné jméno *nekřtěňátko*. V kladné podobě, tedy jako *křtěňátko*, je v korpusu SYN obsaženo jen jednou z 22 vyskytů. V pražském systému je lemma bez záporky. Ve všech sporných případech musí rozhodnout tvůrce slovníku, kterou alternativu zvolit.

U rčení typu *hlava nehlava* lemmatizujeme obě složky stejně, a to jako nezáporná lemmata (v našem příkladě tedy *hlava*). Podobných rčení lze nalézt v korpusech celou řadu, jak ukazují příklady (22) a (23) z korpusu SYN.

režisér musí tvrdě zakročit, hvězda nehvězda (22)
Sochařka nesochařka, nejspíš bydlí v Greenwich Village. (23)

2.2.3 Slovní tvary „bez lemmat“

Na závěr se ještě dotkneme problematických slovních tvarů, které nemají jasně definované lemma.

Takovým slovním tvarům sice lze nějaké lemma přiřadit, ale stává se, že v tomto tvaru lemma vůbec neexistuje.

Např. slovo *krážem* ze spojení *křížem krážem* je sporné. V analogii s křížem by se mohl nabízet tvar **kráž*, avšak takový slovní tvar (současná) čeština nemá. V pražských korpusech je *krážem* lemmatizováno jako příslovce *krážem*, což je logické, ovšem předchozí *křížem* má lemma *kříž*. V tomto konkrétním případě by i slovní tvar *křížem* měl být lemmatizován nejspíš jako příslovce s lemmatem *křížem*. Samozřejmě by bylo nejlepší v tomto případě tato slůvka vůbec od sebe neoddělovat a lemmatizovat je dohromady. Taková spojení však zásadně necháváme na zpracování ve vyšších rovinách lingvistického popisu textu. Navíc jsme v korpusu SYN našli i příklad (24), i když zřejmě velmi netypický, kdy by takovéto spojení nebylo vhodné.

prošed křížem a krážem město (24)

Dalším takovým slovním tvarem je *bycha* ze spojení *pozdě bycha honit*. Lemmatem zřejmě nemůže být *bych* jako podstatné jméno, neboť takové neexistuje. Pražské řešení $\lambda(bycha)$=*bycha* nám připadá rozumné, včetně přiřazení akuzativu namísto pro lemma obvyklého nominativu.

Takových slovních tvarů je celá řada. Jejich lemmatizaci je třeba řešit individuálně. Považujeme za rozumné použít jako lemma sám slovní tvar (jako v příkladě *bycha*), i když nesplňuje podmínku obvykle na lemmata kladenou, totiž že se má jednat o infinitiv, resp. nominativ. My tuto podmínku od lemmat nevyžadujeme (viz definice lemmatu na str. 3).

3 Mutace

Již v úvodu jsme narazili na problémy týkající se variant. V kapitole 2 o lemmatech a lemmatizaci jsme zavedli pojem vícenásobné lemma s variantními lemmaty.

Problém variant je však mnohem širší, dokonce ani sám termín varianta není jednoduchý. Jeho pojetí v lingvistické literatuře totiž zdaleka není jednotné, navíc se používá ještě termín dubleta, s podobným významem. Přehled různých chápání obou termínů je stručně podán v (Tušková, 2006). Zde si autorka vzala za východisko Mluvnici češtiny (viz (Akademická mluvnice)), která dělí varianty na rovnocenné a diferencované. Rovnocenné jsou ty, které jsou „rovnocenné významově, funkčně i stylově, a jsou navzájem volně zaměnitelné“, zatímco diferencované nelze volně zaměňovat kvůli stylovému zabarvení, dobové vázanosti, frekvenci nebo různému významu. Jak je vidět, toto dělení není zdaleka jednoznačné ani objektivní.

Patrně by nebylo těžké zavést nějakou formální definici založenou na ortografických rozdílech v zápisu, ale je jisté, že s takovou definicí by nikdy nebyli spokojeni všichni. Ukazuje se totiž, že varianta je pojem velmi různorodý a většinou je spojen s nějakým typem hodnocení — stylu, časového zařazení, dialektu, a podobně.

Při formálním morfologickém popisu jednotlivých slovních tvarů nás však tato hodnocení nezajímají. Naopak bychom se jim chtěli vyhnout, protože často nemají jednoznačná kritéria. Nějakou kategorii, která rozliší slovní tvary se stejnými hodnotami morfologických kategorií, však potřebujeme, abychom mohli vždy zachovat platnost Zlatého pravidla morfologie. Protože však termín „varianta“ má již v lingvistice svůj význam (i když ne zcela přesně vymezený), navrhujeme pro naše účely termín jiný, a to mutace. Jeho vymezení je čistě technické:

Mutace jsou takové dvojice slovních tvarů, které mají stejné lemma a které nelze rozlišit hodnotou žádné jiné morfologické kategorie. Jinými slovy jsou to takové dvojice slovních tvarů, pro které mají všechny morfologické kategorie stejnou hodnotu.

Pojem mutace je širší než varianta, mezi mutace řadíme totiž nejen varianty (v obvyklém významu), můžeme mezi ně zařadit např. i dvojici vokalizované a nevokalizované předložky, které se za varianty nepovažují. Také rozdílné tvary osobních zájmen, např. *jeho, ho, něho, něj, jej* nejsou pravými variantami, přestože hodnoty všech jejich klasických morfologických kategorií jsou stejné. V takových případech bychom sice mohli zavést nové kategorie se speciální sadou hodnot, které uvedené tvary rozliší, ale zavedení kategorie `Mutace` umožní vyřešit tento problém pro všechny podobné případy najednou.

Zavádíme tedy novou kategorii `Mutace`, která svými hodnotami rozliší mutace, jak jsme je vymezili v předchozích odstavcích.

Nejsnadnější je vymezení tzv. flektivních mutací, které se liší v zakončení (např. *hradu — hradě*). Zahrnujeme mezi ně i nespisovné varianty.

U ostatních (globálních) mutací, které se projevují v celém paradigmatu, je situace složitější. Někdy jde o pouhé ortografické varianty (*atomismus — atomizmus*), někdy o různou výslovnost (*citron — citrón*), případně ovlivněnou obecnou češtinou (*mýdlo — mejdlo*), jindy o různé způsoby tvoření (*brzy — brzo*).

V této kapitole popíšeme právě zavedenou kategorii `Mutace` v souvislosti s nevyhovujícím řešením variant v současných morfologických systémech.

Hlavním myšlenka spočívá v rozdělení mutací na dvě skupiny, globální a flektivní, podle toho, zda se týkají celého paradigmatu, nebo jen některých jeho slovních tvarů.

3.1 Motivace

Představme si, že bychom chtěli vyhledat v nějakém korpusu všechny slovní tvary náležející lemmatu *okénko*. V lemmatizovaném korpusu to uděláme snadno dotazem např. [lemma="okénko"]. Mělo by nás však také napadnout, že lemma *okénko* má variantu *okýnko*, která, ač ji můžeme chápat jako „míň spisovnou", je uváděna jako rovnocenná (SSJČ). A pokud bychom chtěli být opravdu důslední, měli bychom zahrnout i nespisovnou variantu *vokýnko*. Výsledný dotaz na korpus by potom mohl vypadat např. takto: [lemma="v?ok[éý]nko"]. Přitom předpokládáme, že slovník, podle kterého se náš korpus lemmatizoval a značkoval, obsahuje lemmata *okénko*, *okýnko* a *vokýnko* (zadaný dotaz připouští i možnost **vokénko*, která však zřejmě neexistuje).

Vzhledem k tomu, že se jedná o varianty jednoho lemmatu, není asi rozumné vytvářet tři různá lemmata. Mnohem přirozenější by bylo, kdyby všechny tvary všech tří variant měly jedno společné lemma. Rozhodněme tedy, že společným lemmatem uvedených tří variant bude např. *okénko*. Budeme-li nyní chtít vyhledat v korpusu např. 7. pád množného čísla lemmatu *okénko*, vyhledají se nám tyto tvary:

1. *okénky*
2. *okýnky*
3. *vokýnky*
4. *okénkama*
5. *okýnkama*
6. *vokýnkama*

První dvě varianty jsou spisovné a víceméně rovnocenné, zbylé jsou nespisovné, přičemž varianta 3 (?*vokýnky*) je podivná, neboť má nespisovný kmen a spisovnou koncovku. Mnohem přirozenější je *vokýnkama* (6). S takovou vyhledávkou budeme spokojeni.

Co však s opačnou úlohou, kdybychom chtěli 7. pád množného čísla lemmatu *okénko* vygenerovat? Dostaneme 6 různých slovních tvarů. Mohli bychom je rozlišit na úrovni stylových příznaků, avšak vidíme, že spisovných i nespisovných tvarů máme více, a to hned ze dvou důvodů. V příkladech 3 a 6 jde o nespisovnost způsobenou protetickým *v-*, v příkladech 4, 5 a 6 je nespisovnost způsobena hovorovou koncovkou *-ama*. My bychom ale chtěli, aby každá relevantní kombinace hodnot morfologických kategorií pro dané lemma vygenerovala jednoznačný slovní tvar (Zlaté pravidlo morfologie — viz 1.2). Jinými

slovy, aby se popis uvedených šesti slovních tvarů v něčem lišil: v lemmatu nebo v hodnotě nějaké morfologické kategorie. Tuto kategorii jsme nazvali Mutace.

3.2 Rozdělení mutací

Jak už jsme si všimli v předchozím oddíle, ne všechny mutace jsou stejného typu. Můžeme rozlišovat mutace na úrovni jednotlivých slovních tvarů, vyjadřujících určitou kombinaci hodnot morfologických kategorií, např. *okýnky* a *okýnkama* pro 7pl, a na úrovni celých paradigmat, např. dvojice *okýnko* — *vokýnko*, *okýnka* — *vokýnka* atd. V prvním případě jde o spisovnou a nespisovnou koncovku, které mohou být připojeny k libovolnému podstatnému jménu rodu středního, které se skloňuje podle vzoru město (*městy* — *městama*), zatímco ve druhém případě jde o nespisovnost týkající se libovolného tvaru většiny lemmat začínajících na *o-*[1].

Ve slovním tvaru *vokýnkama* se potom spojují oba druhy nespisovnosti, tento tvar je tedy nespisovný „na druhou“.

Některé slovní tvary však vykazují mutace, které nejsou systematické vůbec, týkají se pouze a jenom příslušného lemmatu. Snad nejznámějším případem je dvojice *brzy* a *brzo*.

Mutace tedy mohou být dvojího typu, přičemž jeden nevylučuje druhý:

1. mutace týkající se celého paradigmatu, tj. všech slovních tvarů,
2. mutace týkající se jen některých tvarů daného paradigmatu.

První typ budeme nazývat mutací globální, druhý typ mutací flektivní.

Flektivní mutace jsou takové mutace, které se projevují jen v některých tvarech paradigmatu.

Globální mutace jsou takové mutace, které se projevují ve všech tvarech paradigmatu, a to všude stejně.

Obrázek 3.1 ukazuje, jak se mohou globální i flektivní mutace kombinovat.

globálně / flektivně	spisovné	nespisovné
spisovné	*okny*	*vokny*
nespisovné	*oknama*	*voknama*

Obrázek 3.1: Příklad kombinace globálních a flektivních mutací

[1] Zdaleka ne všechny však mohou protetické *v-* přijímat. Protipříklady: **voficiální, *vovšem, *vondatra, *vovoce, *votec*, ale *vocet* ano.

3.3 Dosavadní pojetí variant v pražském a brněnském systému

Současné systémy se problému variant (mutací) věnují jen okrajově, a proto v nich také často nebývá splněno Zlaté pravidlo morfologie. V pražském systému je variantám vyhrazena 15. pozice v morfologické značce (viz (Hajič, 2004)). Pro ni je vytvořen číselník, který kóduje stylové příznaky. Globální a flektivní mutace zde nejsou rozlišeny, což vede k velké nekonzistenci v popisu.

Většinou se rozlišují mutace flektivní, ale ne vždy. Flektivní mutace, které jsou systematické, bývají obvykle zahrnuty ve vzoru, podle kterého se generují všechny tvary daného paradigmatu, ale značí se i mutace nesystematické, např. mutace *dýchá* se značkou `VB-S---3P-AA---` / *dýše* se značkou `VB-S---3P-AA--2`.

Některé globální mutace se popisují samostatnými lemmaty, např. *busola* — *buzola*, jiné se sice sdružují pod společné lemma, ale potom nastává situace, kdy se k danému lemmatu a morfologické značce vygenerují dva různé slovní tvary, např. lemma *klauzule*, ke kterému existuje (zastaralá) mutace *klausule*. Ve slovníku jsou zahrnuty obě mutace pod společným lemmatem *klauzule*, ale nejsou jako mutace označeny. Každá morfologická značka tedy vytvoří pro toto lemma dva slovní tvary, jeden se *-s-* a druhý se *-z-*, čímž porušuje Zlaté pravidlo morfologie.

Brněnský systém se mutacemi zabývá ještě méně. Různé globální mutace jednoho lemmatu se považují za lemmata dvě. Např. *busola* i *buzola* mají stejné morfologické značky ve všech příslušných dvojicích svých tvarů, ale lemmata jsou různá. Svým způsobem je toto řešení lepší než pražské, protože se tím neporuší Zlaté pravidlo morfologie.

Zato flektivní mutace v brněnském systému Zlaté pravidlo porušují často. Existuje sice gramatická kategorie „Stylistický příznak tvaru“, ale v korpusu DESAM je využívána minimálně. Např. slovní tvary *soudci* i *soudcové* mají stejnou značku `k1gMnPc1`, zatímco v pražském systému dostávají *soudci* značku `NNMP1-----A----` a *soudcové* `NNMP1-----A---1`, jsou tedy rozlišeny na 15. pozici jako rovnocenné varianty.

V přehledu morfologických kategorií (viz kap. 4) jsme umístili flektivní mutaci (FMU) mezi flektivní kategorie a globální mutaci (GMU) mezi kategorie globální. Důvod je zřejmý. Globální mutace popisuje celé paradigma, zatímco flektivní mutace se týká jen některých jeho slovních tvarů. Oba typy mutací se mohou navíc kombinovat, a na to by jediná kategorie nestačila.

Zbývá nyní vyřešit otázku hodnot kategorií FMU a GMU.

3.4 Diskuse o hodnotách kategorie Mutace

Vzhledem k tomu, že každý slovní tvar může vykazovat jen malé množství mutací, ať už globálních, nebo flektivních, stačilo by oba typy mutací v rámci jednoho slovního tvaru jednoduše číslovat. Neznáme případ, kdy by slovní tvar měl více než 10 mutací (platí i pro globální a flektivní mutace zvlášť), takže by k popisu stačily číslice 0 až 9.

Nejjednodušší by bylo použít číslice pouze k formálnímu odlišení mutací. Na druhou stranu jsou ale uživatelé korpusů zvyklí na to, že hodnota mutace, když

už je někde uvedena, také o něčem vypovídá. V současných systémech se její hodnota využívá k odlišení stylových příznaků mutací.

V příspěvcích (Hlaváčová, 2008) a (Hlaváčová – Lopatková, 2008) jsme proto navrhli — ve shodě s dosavadní praxí — zakódovat do hodnot FMU i GMU stylový příznak. Tato vlastnost je velmi subjektivní, proměnlivá v čase i místně, takže je velmi obtížné ji stanovit. Pokud by záleželo skutečně jen na rozlišení jednotlivých mutací kvůli jednoznačné morfologické značce, na konkrétní hodnotě (čísle) by nemuselo záležet. Přesto jsme navrhli kódovat mutace pomocí pomyslné škály, kdy 0 by ležela uprostřed jako nejběžnější synchronní mutace. Číslice 1, 2,... by vyjadřovaly mutace rovnocenné, přičemž řada by směřovala „do budoucnosti“. To znamená, že tyto hodnoty by dostaly mutace nespisovné, případně hovorové či obecně české (budoucnost je tu míněna jako potenciální čas, kdy by nespisovné mutace eventuálně mohly být uznány jako spisovné). Číslice 9, 8, 7,... by potom směřovaly „do minulosti“, což znamená, že by popisovaly mutace už nepoužívané, zastaralé, archaické, případně nářeční.

Je jasné, že taková škála má velmi mnoho nevýhod, a je tedy snadno napadnutelná. Je velmi abstraktní, časové vztahy jsou, zvláště co se týče budoucnosti, často jen hypotetické. Některé mutace, např. nářeční, na takovou škálu ani nepatří, potřebovaly by svou samostatnou škálu.

Další námitkou proti tomuto řešení je fakt, že stylové příznaky nejsou přijímány celou lingvistickou komunitou jednoznačně. Spíše by měly být (a také jsou) cílem výzkumu (viz např. (Křístek, 2002)). Morfologický slovník by měl být na subjektivních názorech jednotlivých badatelů nezávislý. Z toho důvodu bychom se neměli snažit hodnoty morfologických kategorií, tedy ani mutací, jakkoliv hodnotit.

Z výše uvedených důvodů od takového návrhu upouštíme a přiřazujeme kategoriím `Flektivní mutace` a `Globální mutace` nezávislou sadu hodnot, prostou jakéhokoli hodnocení. Pomocí těchto hodnot se snažíme vyjádřit obecné vlastnosti mutací. Nejběžnější dvojice hodnot obou kategorií uvádí tabulka 3.1. Z uvedených příkladů vidíme, že tyto hodnoty se mohou týkat nejrůznějších typů slovních tvarů (v případě FMU) i celých paradigmat (v případě GMU).

Dvojice hodnot	Vysvětlení	Příklad
D — K	delší — kratší (počet písmen)	*pustější — pustší* *skáčeme — skáčem* *vracejí — vrací*
d — k	dlouhá — krátká (samohláska)	*musím — musim* *zavříno — zavřeno* *tráv — trav* *salón — salon*
t — m	tvrdá — měkká	*vlaštovka — vlasťovka* *student — študent* *mazám — mažu*

Tabulka 3.1: Základní hodnoty kategorií `Flektivní mutace` a `Globální mutace`

Již v úvodní kapitole jsme naznačili, že mutace, tzn. kategorie FMU ani GMU, nebudeme zahrnovat do morfologické značky. Vyčleňujeme je jako další, speciální atributy popisu a souhrnně jim říkáme `Mutace`. V rámci této kate-

gorie lze libovolně kombinovat jednotlivé typy mutací, globálních i flektivních. Každý slovní tvar je tedy jednoznačně popsán hodnotami svého lemmatu, morfologické značky (bez mutací) a hodnotou kategorie `Mutace`.

Toto řešení nám připadá velmi slibné, neboť jím lze zachytit neomezené množství kombinací jednotlivých mutací. Vyhneme se tím také jednomu z obvyklých požadavků, totiž stanovení „základní mutace“, která by měla být nejběžnější, což je ovšem většinou velmi těžké stanovit. Kód, který zaznamená hodnoty kategorie `Mutace`, zapíšeme pomocí regulárního výrazu takto:

$$\mathrm{MUT} = \mathrm{F}.+\mathrm{G}.+$$

Znaky zapsané na místě tečky za písmenem F kódují typy flektivní mutace, znaky za G typy mutace globální.

Kromě všeobecných hodnot z tabulky 3.1 existuje celá řada mutací typických pouze pro některé kombinace hodnot morfologických kategorií. Množinu hodnot typických pro kategorii `Globální mutace` uvádíme v tabulce 4.4 na str. 34. O kategorii `Flektivní mutace` pojednáme blíže v oddíle 4.2.12, o jednotlivých hodnotách potom v kapitolách 9 až 13, kde popisujeme vzory konkrétních slovních druhů.

4 Morfologické kategorie

Nejprve vytvoříme množinu všech morfologických kategorií[1], které se k popisu slovních tvarů používají. Množinu hned zpočátku rozdělíme na dvě podmnožiny. První bude obsahovat tzv. globální kategorie, druhá kategorie flektivní.

Globální morfologická kategorie je taková kategorie, jejíž hodnota je stejná pro celé paradigma. Příkladem globální morfologické kategorie je slovní druh.

Flektivní morfologická kategorie je taková kategorie, jejíž hodnoty se pro jednotlivé slovní tvary jednoho paradigmatu liší. Příkladem flektivní morfologické kategorie je pád. Také kategorie rod je flektivní, ovšem jen v rámci přídavných jmen a některých zájmen a číslovek.

U každé kategorie uvedeme:

1. hodnoty, kterých může nabývat;
2. pro jaké druhy slovních tvarů je relevantní; nemá např. smysl určovat rod příslovcí, nebo stupeň přivlastňovacích přídavných jmen. Pro jednoduché formulace v následujícím textu si situaci zjednodušíme tak, že uvažujeme každou kategorii pro každý slovní tvar s tím, že jedna z hodnot může být UNDEF.

Každé kategorii přidělíme zkratkové jméno.

Každé hodnotě přiřadíme jednoduchý symbol. Jak už bylo uvedeno, tento symbol může být v různých systémech různý. Zde ho uvádíme proto, aby bylo možno jednoduše tvořit dotazy a nezaplést se do zbytečně zdlouhavých popisů. Hodnoty morfologických kategorií vycházejí převážně ze symbolů pro tytéž nebo podobné hodnoty v pražském systému (viz (Hajič, 2004)).

Hodnoty jednotlivých kategorií jsou v ideálním případě ekvivalencí na množině všech slovních tvarů. Tato ekvivalence rozděluje slovní tvary na navzájem disjunktní třídy, které celou množinu pokrývají. To mimo jiné znamená, že hodnoty každé kategorie jsou „vyčerpávající“, tzn. že neexistuje slovní tvar, který by neměl přiřazenou hodnotu každé kategorie, počítáme-li i hodnotu UNDEF.

V praxi ovšem často dochází k případům, že jeden slovní tvar lze popsat více hodnotami jedné kategorie (např. kategorie `Pád` pro slovní tvar *stavení*). Při postupném přechodu do dalších rovin lingvistického popisu se však množina hodnot jednotlivých kategorií obvykle zmenšuje. Už syntaktický rozbor věty vybere z množiny všech možných hodnot dané kategorie většinou hodnotu jedinou. V některých případech je třeba pro desambiguaci zkoumat další roviny, sémantickou, případně pragmatickou, které berou v úvahu kontext širší než jedna věta, nebo i mimojazykové znalosti. Některé věty zůstávají víceznačné i po podrobných rozborech.

[1] Mluvíme o morfologických kategoriích, přestože některé z nich ze striktního pohledu čistě morfologické nejsou. Už sama nejdůležitější kategorie, tedy slovní druh, nevyjadřuje jen morfologické vlastnosti.

Na úvod ještě poznamenejme, že většina kategorií a jejich hodnot je tradiční, některé jsou však nové, vytvořené speciálně pro snadnější automatické zpracování textů. Tradiční lingvisté se mohou nad některými návrhy pozastavovat. Budeme se však snažit nové přístupy vždy důkladně odůvodnit.

Ve zbytku kapitoly představíme všechny morfologické kategorie, které se používají k popisu českých slovních tvarů. U nových kategorií nebo hodnot, které vyžadují podrobnější vysvětlení, necháváme rozbor na zvláštní kapitoly.

4.1 Globální morfologické kategorie

Připomeňme si definici globální kategorie:

Globální morfologická kategorie je taková kategorie, jejíž hodnota je stejná pro celé paradigma.

Zde je výčet všech globálních kategorií

- `Slovní druh` — POS
- `Poddruh` — SUB
- `Funkce` — FCE
- `Vid` — ASP
- `Zkratka` — ABR
- `Globální mutace` — GMU

4.1.1 Slovní druh (POS)

Tato kategorie je základní v tom smyslu, že má definovanou konkrétní hodnotu pro každý slovní tvar. Jinými slovy, v této kategorii nezavádíme hodnotu UNDEF. Tento požadavek se dá chápat také tak, že kategorie slovní druh rozkládá množinu všech slovních tvarů na podmnožiny, které celou množinu pokrývají. Ideální by bylo, kdyby tyto množiny mohly být disjunktní, takže by se o každém slovu dalo jednoznačně říci, jaký má slovní druh. To bohužel neplatí, a to z několika důvodů:

- Homonymie

  Homonymní slovní tvary mají více slovnědruhových interpretací, např. *drát* jako podstatné jméno a *drát* jako sloveso. Tento problém se snadno vyřeší tím, že se víceznačná lemmata nějakým způsobem odliší. V našem příkladě budeme mít potom dvě různá lemmata *drát*: *drát-1* a *drát-2*.

- S-formy

  Název jsme převzali od brněnských kolegů. Jde o skupinu slov, která nejsou jednoznačně zařaditelná. Může to být proto, že v různých kontextech nabývá slovní druh různých hodnot, přestože můžeme říci, že jejich význam je stále týž. Často se ani renomovaní lingvisté neshodnou, jaká hodnota by to v daném kontextu měla být. Většinou jde o slova synsémantická (funkční). I tento problém lze vyřešit zmnožením jednotlivých

lemmat. V tomto případě je však podstatné povolit při morfologické analýze více možných hodnot, potažmo více morfologických značek. Příklady s-forem: *dokud* (příslovce, spojka), *tedy* (spojka, částice), *jak* (příslovce, spojka, částice i podstatné jméno), *jednou* (číslovka, příslovce, částice).

Hodnoty kategorie `Slovní druh`

- N: podstatné jméno
- A: přídavné jméno
- P: zájmeno
- C: číslovka
- V: sloveso
- D: příslovce
- R: předložka
- J: spojka
- I: citoslovce
- T: částice
- F: cizí slovo (K)
- G: prefixový segment (K)
- S: složenina (K)
- X: neznámé slovo

Slovní druhy označené znakem (K) v závorce byly nově zavedeny na zasedání Konkláve. Znak X označuje slovní druh „neznámé slovo“, které sice k tradičním slovním druhům nepatří, ale lingvisté se s ním již dávno setkávají při práci s jazykovými korpusy. Ostatní slovní druhy jsou tradiční, budeme je nadále považovat za známé a jejich definice uvádět nebudeme.

Uveďme si charakteristiky nových slovních druhů.

4.1.1.1 Cizí slovo

je slovo, které nepodléhá české flexi a nemá v češtině vlastní význam.

Příklady cizích slov: *the, you, der, di, du, to, company,....*

Nepatří sem však nesklonná slova, která jsou součástí české slovní zásoby, jako např. *kupé*[2], *lila* (barva). Ta sice české flexi také nepodléhají, ale mají v češtině jasný význam. Může se stát, že cizí slova, zvláště ta krátká, jsou homonymní se slovy českými (z uvedených příkladů jsou to slova *der* — imperativ slovesa *drát*, *di* a *du* jako nespisovné tvary slovesa *jít*, i slovní tvar *to* homonymní s českým ukazovacím zájmenem). Lemma cizího slova je vždy stejné jako slovo samo.

Zavedení slovního druhu cizí slovo velmi usnadní automatické zpracování.

[2] Děkuji vedoucímu práce za upozornění na existenci neobvyklého tvaru *kupém*, který však spíše potvrzuje, že toto slovo nepatří mezi slova cizí, neboť jakési české flexi zdá se podléhá.

4.1.1.2 Prefixový segment

je začátek slova, který stojí samostatně, a teprve někde dál v textu je doplněn na plnovýznamové slovo. Většinou to bývá předpona. Jde-li o jiný řetězec, vždy se v daném kontextu jako předpona chová.

Příklady: *česko a rusko - německý*, *tří až čtyřprocentní*. Spojovník, který případně může stát bezprostředně za prefixovým segmentem, ať už s mezerou, nebo bez ní, není součástí ani slova, ani lemmatu.

Lemma prefixového segmentu se shoduje se slovním tvarem. V uvedených třech příkladech jsou tedy lemmaty slova *česko*, *rusko* a *tří*.

4.1.1.3 Složenina

popisuje slovní tvar, který vznikl z více slov (většinou různých slovních druhů) a určení jeho slovního druhu je problematické. Jejich vymezení a podrobný popis jsou uvedeny v kapitole 6.

Příklady: *zač, oň, byls*.

Pod termínem „složenina“ chápeme tedy něco jiného než např. Šmilauer v (Šmilauer, 1971), který pokládá termíny „složenina“ a „složené slovo“ za synonyma. Složená slova (kompozita), sice také vznikla z více slov, avšak jejich zařazení mezi slovní druhy nečiní potíže (např. *středomořský, vlastizrádce, spoluviník*).

4.1.1.4 Neznámé slovo

je takové slovo, jehož slovní druh neumíme určit. Jsou to slova, která morfologická analýza nerozpozná.

Tato hodnota kategorie `Slovní druh` už v pražském systému existuje a uživatelé s ní pracují.

Při ruční anotaci se neznámým může stát pouze takové slovo, které anotátor nezná a nemůže určit. Může to být například nějaká šifra nebo naprosto nesrozumitelný překlep (např. poslední slovo před touto závorkou se posunutím pravé ruky na klávesnici může stát neznámým slovem *ořejkeoz*).

Neznámá slova nejsou součástí morfologického slovníku. Při automatické morfologické analýze je tedy tato hodnota přiřazena těm slovním tvarům, které nejsou rozpoznány. Z neznámého slova se může stát známé buď přidáním do slovníku, nebo dalšími metodami (guesser — viz oddíl 7.1.1).

4.1.2 Poddruh (SUB)

Kategorie `Poddruh` je relevantní pro všechny slovní druhy kromě citoslovcí, předložek, cizích slov, prefixových segmentů a neznámých slov.

Hodnoty kategorie `Poddruh` jsou závislé na slovním druhu. Některé slovní druhy mají dva poddruhy (zájmena, číslovky, příslovce). V tomto případě nazveme druhou kategorii poddruhu jinak (`Funkce`, viz 4.1.3), kvůli možné kombinovatelnosti.

Podívejme se tedy na jednotlivé slovní druhy. U každého slovního druhu uvedeme, jaké poddruhy se u něj rozlišují, a přiřadíme jim kromě názvu i jednopísmenný kód. Tento kód je pokud možno v souladu s kódem 2. pozice v pražském

systému morfologických značek, která se shodou okolností také nazývá Poddruh. My však tuto kategorii chápeme zcela odlišně. Poddruh v dosavadním pražském systému je směs hodnot popisující nejrůznější slovní tvary. Některé hodnoty se týkají jednotlivých slovních tvarů, jiné celých lemmat, není to tedy ani flektivní, ani globální kategorie. Náš `Poddruh` je kategorie striktně globální, je tedy relevantní vždy pro celé paradigma.

Oba poddruhy, náš globální, i dosavadní pražský, mají jednu stejnou podstatnou vlastnost: jejich kód je v rámci celé kategorie jedinečný. Znamená to, že z hodnoty kategorie `Poddruh` je možno určit slovní druh. Jak bylo řečeno výše, nezáleží na konkrétním kódu hodnoty, ale je důležité, aby tyto kódy byly různé.

Tento požadavek, zavedený Hajičem (viz (Hajič, 2004)), není na první pohled příliš důležitý. Kategorie poddruh je napříč slovními druhy velmi různorodá, u každého slovního druhu popisuje jiné vlastnosti, a tak by se mohlo zdát, že nezáleží na jednoznačnosti jejího kódu. Podstatně to však zjednoduší vyhledávání v korpusech podle této kategorie. Usnadní to posléze i kódování kategorie `Poddruh` v rámci složenin. Jednoznačnost kódu má tři výjimky:

1. Každý slovní druh, pro který je kategorie `Poddruh` relevantní, musí být hodnotami této kategorie zcela pokryt. U některých slovních druhů proto zavádíme hodnotu „Ostatní“, která je vždy kódována hodnotou 0 (nula), je tedy stejná pro více slovních druhů. Konkrétně jsou to podstatná jména, přídavná jména, slovesa, příslovce a částice. Tato hodnota vlastně vyjadřuje jen tu informaci, že dané slovo nemá žádnou z vlastností sledovaných v kategorii `Poddruh`. Proto není nutné tuto hodnotu dělit podle slovních druhů.

2. Poddruh „přivlastňovací“ sdílejí zájmena a přídavná jména. Jde totiž o velmi podobnou vlastnost u obou slovních druhů.

3. Poddruh „deverbativní“ sdílejí podstatná jména, přídavná jména a příslovce. I zde jde o vyjádření podobné vlastnosti, totiž blízké příbuznosti se slovesem.

4.1.2.1 Poddruh podstatných jmen

- S: deverbativní typu *věznění, pokrytí,...*
- 0: ostatní

Hodnota S náleží podstatným jménům odvozeným ze sloves, vyjadřující slovesný děj a končící na *-ní* nebo *-tí*. Tato podstatná jména se mohou chovat ve větě jinak než ostatní, primární podstatná jména. Jedině ona totiž mohou mít zvratnou částici a adverbiální rozvití adverbii pravidelně odvozenými od přídavných jmen (*zpívání falešně*). Viz též příklad (25).

Mužské bádání, v ženském jazyce nazývané též štítění se práce, dokáže jít rozumnému člověku (míním tím ženu) na nervy. (25)

Kopečný je v (Kopečný, 1962a) považuje za slovesa pro jejich „naprostou totožnost významovou (obsahovou) a pro paradigmatičnost, s níž se od slovesných základů derivují.“ Syntakticky i morfologicky ovšem patří mezi podstatná jména, proto je mezi nimi ponecháme. K odlišení stačí vyjádření poddruhu.

Je třeba ovšem rozlišovat lexikalizovaná deverbativní podstatná jména, která nevyjadřují slovesný děj. Kopečný uvádí příklad *vázání ječmene — lyžařské vázání*. V tomto případě patří do slovníku dva záznamy, a to *vázání-1* s poddruhem 0 a *vázání-2* s poddruhem S. Dalším příkladem může být *krmení zvířat — krmení pro zvířata*. Najdou se i případy, kdy tato distinkce bude sporná.

4.1.2.2 Poddruh přídavných jmen

- U: přivlastňovací (*matčin, otcův,...*)
- G: od přechodníku přít. (*mající, sedící, beroucí,...*)
- M: od přechodníku min. (*ušedší, nakupovavší,...*)
- S: ostatní deverbativní (*namazaný, zemřelý, nakousnutý, namazán, nakousnut,...*)
- 0: ostatní (jarní, starý,...)

Přídavná jména utvořená od přechodníků nejsou opět podle Kopečného „normální adjektiva“, neboť neplní všechny své funkce. Konkrétně nemohou být použita ve funkci jmenného přísudku (**žák je sedící*). Tato námitka neplatí obecně, neboť některá tato přídavná jména se už lexikalizovala. Příkladem může být lemma *vynikající* v příkladě (26).

Ten čaj je vynikající. (26)

Vzhledem k jejich adjektivní flexi je považujeme za přídavná jména.

Rozhodnutí, která přídavná jména jsou deverbativní (SUB=S) a která ostatní (SUB=0), je někdy těžké. Nejednoznačné případy (např. *šílený, vleklý*) bude třeba řešit jednotlivě.

4.1.2.3 Poddruh zájmen

U zájmen je třeba rozlišovat na této pozici dvě kategorie; jedné necháváme název `Poddruh`, druhou nazýváme `Funkce`. Vzhledem k tomu, že `Funkce` je kategorie společná pro více slovních druhů, totiž pro zájmena, číslovky a příslovce, pojednáme o ní jako o zvláštní morfologické kategorii dále v oddíle 4.1.3.

Toto dělení jsme zavedli proto, že běžné dělení zájmen nebere v úvahu dvojí podstatu užívaných hodnot. Např. zájmeno *něčí* je současně přivlastňovací i neurčité. Dvojí dělení zájmen bylo prvně použito v brněnském systému a přejalo ho i Konkláve.

Hodnoty obou kategorií, `Poddruh` i `Funkce`, se samozřejmě mohou kombinovat (proto byly zavedeny), ovšem ne zcela libovolně. Možné kombinace jsou uvedeny v tabulce 4.1 na str. 32, která by měla obsahovat úplný výčet všech českých zájmen.

Hodnoty kategorie `Poddruh` pro zájmena jsou:

- Z: substantivní (*já, kdo, nikdo, oni,...*)
- U: přivlastňovací (*můj, čí,...*)
- D: ukazovací (*ten, takový,...*)
- V: vymezovací (*každý, všechen, týž, sám*)
- 0: ostatní

Až na zájmena substantivní jde o klasické hodnoty.

Zájmena substantivní nahrazují osobní zájmena, ale navíc ještě některá zájmena neurčitá (*někdo, kdosi,...*), tázací (*kdo, co*) a záporná (*nikdo, nic*), která mají jednak substantivní skloňování a jednak podobné syntaktické postavení ve větě jako zájmena osobní (určitá).

Substantivní zvratná zájmena *sebou, sobě, sebe, se, si* nemají první pád, jejich lemma tedy položíme rovno slovnímu tvaru, jmenovitě *sebou, sobě, sebe, se, si*. Viz též oddíl 2.2.3, kde jsme se zamýšleli nad lemmatizací slovních tvarů, které nemají nominativ, resp. infinitiv.

Slovní tvary *se*, *si* jsou homonymní. Jde buď o zájmena, nebo o částice, podle kontextu. Tvar *si* je navíc obecně českou variantou 2. osoby jednotného čísla přítomného času lemmatu *být*.

Slovní tvar *toť*, který se tradičně řadí mezi ukazovací zájmena, je homonymní se složeninou. Z korpusů jsme totiž zjistili, že jako klasické zájmeno se vyskytuje jen ve velmi omezeném typu vět. Většinou spíše zastupuje frázi *to je*, případně *to jsou*, a je tudíž složeninou. Viz příklady (27) až (29):

Toť se ví, že... (zájmeno ukazovací)	(27)
Toť vše. (složenina v jednotném čísle)	(28)
... toť pomyslné kóty (složenina v množném čísle)	(29)

4.1.2.4 Poddruh číslovek

Slovní druh Číslovky je velmi různorodý. Viz např. nejnovější pojednání o číslovkách a kvantifikátorech (Šimandl, 2007) a (Jiranová, 2008).

Hodnoty kategorie `Poddruh`, které zde uvádíme, jsou ty, na kterých se shodlo Konkláve. Kromě nich lze porůznu v literatuře najít ještě další druhy číslovek. Jde např. o číslovky velikostní (*tisícový, osmitisícový*), které ovšem považujeme za přídavná jména.

Na druhou stranu číslovky úhrnné, souborové a druhové mají nejasně rozlišené skloňování (viz též tabulku 13.3 a zmínku v kapitole 13 na straně 130), takže by možná bylo rozumné je sdružit do poddruhu jediného. Tato otázka by zasloužila ještě podrobnější zkoumání.

Slova, která se občas označují jako dílové číslovky (např. *polovina, osmina, setina*) chápeme jako podstatná jména. Výjimku tvoří pouze tři dílové číslovky uvedené v tabulce 4.1. Stejně tak tradiční číslovky skupinové (např. *dvojice, pětka, tisícovka*) považujeme za podstatná jména. Je to proto, že se ve větě vždy syntakticky projevují jako klasická podstatná jména.

Slova typu *kopa, tucet, hromada, spousta, kus, kousek* jsou podstatná jména. Jejich kvantitativní vlastnost se v případě potřeby musí zachytit jinak, podle aplikace.

Uvědomujeme si, že s takovým rozdělením číslovek leckdo nemusí souhlasit. Číslovky jsou bohužel velmi těžko vymezitelná kategorie. Pro potřeby automatického zpracování však je potřeba vymezit číslovky jednoznačně a co nejjednodušeji.

Hodnoty kategorie `Poddruh` pro číslovky tedy jsou:

- 1: základní (*jedna, sto*,...)
- r: řadové (*druhý, pátý*,...)
- u: úhrnné (*dvé, patero*,...)
- s: souborové (*dvoje, patery*,...)
- d: druhové (*dvojí, paterý*,...)
- n: násobné (*dvakrát, pětkrát*,...)
- o: opakovací (*podruhé, popáté*,...)
- v: výčtové (*zadruhé, zapáté*,..., ale i *druhé* z dvojice *za druhé*,...)
- p: dílové (*půl, čtvrt, třeť*)

4.1.2.5 Poddruh sloves

- m: modální (*moci/moct, mít, mívat, muset, musívat, smět, chtít, hodlat, dát se, dávat se, dovést, umět*)
- f: fázová (*začít, začínat, přestat, přestávat, zahájit, skončit,...*)
- b: pomocná (*být, bývat, mít, dostat*)
- 0: ostatní (*navštívit, koupat se,...*)

Podle závěrů Konkláve se u pomocných sloves *být* a *bývat* jako pomocné značí jen tvoření minulého času, podmiňovacího způsobu a budoucího času. V tom případě jsou pro něj relevantní kategorie Osoba, Číslo i Slovesný tvar (indikativ *přišel* ***jsi***, příčestí činné ***byl*** *bys přišel* nebo budoucí čas ***budeš*** *skákat*). Totéž pro *bývat*, kde však může být pouze příčestí činné (*byl bys* ***býval*** *přišel*). Slovní tvary lemmatu *bývat* v indikativu prézentu nejsou nikdy pomocným slovesem a jednoduchý budoucí čas *bývat* nemá.

4.1.2.6 Poddruh příslovcí

- P: místní (*kudy, tudy, odkud, nikudy, nikam; daleko, nedaleko,...*)
- T: časová (*kdy, nikdy; včera, odpoledne,...*)
- J: způsobová (*jak, všelijak; krásně, velmi, široce,...*)
- R: predikativní (*jasno, možno, teplo, volno,...*)

První skupina příkladů (před středníkem) místních, časových a způsobových příslovcí se často řadí mezi tzv. zájmenná příslovce. U nich je možno určovat i kategorii `Funkce` s hodnotami určitá, neurčitá, záporná jako u číslovek a zájmen (viz dále, zejména tabulku 4.3). Pomocí kategorie `Funkce` lze též zájmenná příslovce vymezit. Jsou to ta příslovce, pro něž je kategorie `Funkce` relevantní, tedy POS=D a FCE≠UNDEF.

4.1.2.7 Poddruh spojek

- ^ (stříška): souřadicí (*a, ale, nebo,...*)
- , (čárka): podřadicí (*protože, když, až, -li,...*)
- * (hvězdička): matematické operace (*plus, minus/mínus, krát, děleno* — neplést s *děleno* jako jmenný tvar přídavného jména *dělený*, případně rod trpný od slovesa *dělit*)

Slovní tvary *abys, abyste, abychom, kdybych, kdybyste, kdybychom*, ale i nespisovné *abysem, abysme, kdybysem* a *kdybysme* jsou také spojky. Podrobněji o nich pojednáme v kapitole 5 o kondicionálu.

4.1.2.8 Poddruh částic

- 7: zvratné (*se, si*)
- c: kondicionálová (pouze *by*)
- 0: ostatní (*ba, ano, bože, nechť, ať,...*)

Bylo by třeba ještě zvážit, zda mezi zvratné částice nepočítat i tvar *sebou* ve spojení např. *hodit sebou*.

Částice by se mohly členit ještě více, např. podle (Akademická mluvnice). Částicemi se intenzivně zabývá i práce (Čermák, 2007). Inspirací může být také klasifikace slovenských částic (viz např. (Slovenská morfologie) a (Šimková, 2001)).

Oba tyto problémy necháváme zatím otevřené.

Ostatní slovní druhy, tedy předložky[3], citoslovce, cizí slovo[4], prefixový segment a neznámé slovo, poddruh nemají.

4.1.2.9 Poddruh složenin

Složeniny „dědí“ poddruh některé své složky. Složeninami včetně jejich poddruhů, se budeme zabývat podrobně v kapitole 6.

[3] Na Konkláve jsme místo Poddruhu zavedli pro předložky kategorii Vokalizace s hodnotami +/-, nový návrh ale toto rozlišení řeší pomocí kategorie Mutace (viz kap. 3).

[4] Výhledově by se mohlo uvažovat o přiřazení jazyka, ze kterého cizí slovo pochází, případně o jeho slovním druhu v daném jazyce. Pro zpracování českého textu ale zřejmě tyto údaje nemají význam.

4.1.3 Funkce (FCE)

Tato kategorie je relevantní pro zájmena, číslovky a příslovce, i když ne všechny vyjmenované slovní druhy mohou nabývat všech jejích hodnot.

Uvědomujeme si, že termín „funkce“ je již obsazen mnoha dalšími významy, přesto si myslíme, že tento název nejlépe vystihuje podstatu této kategorie. Na Konkláve byla tato kategorie nazvána „Neurčitost“ s tím, že jde pouze o pracovní název.

Hodnoty tázací a vztažná jsou v praxi velmi obtížně rozlišitelné (viz např. (Ševčíková, 2008)). Jedním z důvodů, proč je rozlišovat na morfologické rovině, spatřujeme v existenci slovních tvarů *jenž, kdož, jakž*[5] apod., které mohou být pouze vztažné, nikoli tázací. Podrobná diskuse o vhodnosti či nevhodnosti rozlišování těchto dvou hodnot je však mimo rámec této práce, proto obě hodnoty zachováváme, přičemž by nebyl problém je v konkrétní implementaci sloučit do hodnoty jediné.

- U: určitá (všechna osobní zájmena, určité číslovky, tady, teď,...)
- N: neurčitá (*někdo, čísi, několik, někdy,...*)
- Z: záporná (*nikdo, ničí, nijak,...*)
- T: tázací (*kdo, čí, kolik, kde,...*)
- V: vztažná (*kdo, čí, jenž, kdy,...*)
- S: zvratná (*se, si, sobě, sebe, sebou*)

Neurčitá funkce (FCE=N) se projevuje většinou pomocí speciálních předpon a přípon, které se připojují ke slovním tvarům s funkcí tázací (FCE=T). Kombinovatelnost ukazuje obrázek 4.1. Předpony a přípony z levého rámečku lze kombinovat se slovy v pravém rámečku, čímž vznikne slovo s neurčitou funkcí. Některé kombinace možné nejsou, např. nelze **něodkud*, dále se netvoří neurčité číslovky pomocí přípon, pouze pomocí předpon. Některé kombinace jsou velmi neobvyklé, ale netroufneme si tvrdit, že zcela nemožné, např. ?*kamžkolivěk*.

lec-, leda-, všeli-, ně-, bůhví-, čertví-, pánbůhví-, pámbuví-, kdoví-, nevím-, kde-, lec-, leda-, ledas-, málo-, všeli-, všelis-, zřídka-, sotva-, -koli, -koliv, -žkoli, -žkoliv, -kolivěk, -si	*kdo, co, čí, jaký, který, kde, kam, kudy, odkud, kdy, jak,* všechny tázací číslovky

Obrázek 4.1: Kombinovatelnost předpon s tázacími zájmeny, číslovkami a příslovci při tvoření slovních tvarů s funkcí neurčitou (FCE=N)

Uvádíme tři tabulky, ze kterých je vidět ortogonálnost kategorie `Poddruh` a `Funkce` pro zájmena (tabulka 4.1), číslovky (tabulka 4.2) a příslovce (tabulka 4.3). Seznam z pravého rámečku schématu 4.1 lze v tabulkách 4.2 a 4.3 doplnit

[5] Jde o příklady typu ... *uhlazujete povrch štětkou, jakž je zvykem lepičů plakátů.*, nikoli o výrazy *jakž takž*.

na místa neurčitých číslovek resp. příslovcí. V tabulce 4.1, která popisuje funkce zájmen, jsou možnosti vypsané explicitně.

Tabulku pro zájmena vytvořila autorka této práce, modifikoval Karel Oliva, doplnila opět autorka. Ostatní tabulky jsou autorčiny.

Tabulka 4.1 by měla být vyčerpávající, to znamená, že jsme se snažili, aby obsahovala veškerá zájmena. Jako zdroj posloužily dostupné slovníky a korpusy SYN2000 a SYN2005.

Prázdné závorky () je možno nahradit slovy ze stejné řádky, která jsou ve sloupci tázacích zájmen, konkrétně v řádku substantivních zájmen jsou to *kdo, co*, v řádku přivlastňovacích zájmen *čí* a v řádku ostatních zájmen *jaký* a *který*.

Hodnoty kategorie `Funkce` pro číslovky jsou stejné jako pro zájmena, ale ne všechny se pro určování číslovek uplatní. Zejména hodnota `Funkce` zvratná je pro číslovky nepoužitelná.

Tabulka 4.2 je analogická tabulce 4.1 pro zájmena, liší se však v jednom podstatném bodě, není totiž úplná. Přesněji, sloupec s funkcí určitá není úplný až na číslovky dílové a potom také políčko s číslovkami základními neurčitými. Ostatní neurčité číslovky získáme ze schématu 4.1 na str. 30, takže tento sloupec, i sloupec s číslovkami tázacími za úplný považovat můžeme. Neúplnost je naznačena třemi tečkami v příslušných políčkách.

Kategorii `Funkce` využíváme i pro specifikaci zájmenných příslovcí. Tabulka 4.3 ukazuje kombinovatelnost zájmenných příslovcí s různými hodnotami kategorie `Funkce`.

Hodnoty určitá, neurčitá a tázací je možné využít též pro bližší určení speciálních přídavných jmen s číselnou předponou (např. *pětičlenný, několikaletý, kolikawattový*). Dokonce můžeme určit tyto hodnoty kategorie `Funkce` i u podstatného jména *-násobek* s číselnými předponami (*pětinásobek, několikanásobek, kolikanásobek*). Nejzajímavější je u těchto slovních tvarů hodnota tázací, neboť může uvozovat vedlejší větu, což u „normálních“ podstatných a přídavných jmen není možné, viz příklady (30) a (31)[6].

Dopita ještě nemá jasno, kolikaletý kontrakt podepíše. (30)

... koeficienty, jež určují, kolikanásobek této základny budou ústavní činitelé dostávat. (31)

[6] Za upozornění na tyto typy přídavných a podstatných jmen děkuji Karlu Olivovi.

	určitá	neurčitá	záporná	tázací	vztažná	reflexivní
substantivní	*já, ty, on, ona, ono, my, vy, oni, ony.*	*leccos, ledacos, všelicos, ně(), bůhví(), pánbůhví(), pámbuví(), čertví(), kdoví(), nevím(), kde(), lec(), leda(), ledas(), málo(), všeli(), všelis(), zřídka(), sotva(), ()koli, ()koliv, ()žkoli, ()žkoliv, ()kolivěk, ()s, ()si.*	*nikdo, nic, pranic.*	*kdo, co, ()pak, ()že, ()ž.*	*jaký, který, jenž, kdo, co, an.*	*si, se, sebou, sobě, sebe.*
přivlastňovací	*můj, tvůj, jeho, její, náš, váš, jejich.*	*něčí, čísi, číkoli, číkoliv, číkolivěk, bůhvíčí, pánbůhvíčí, pámbuvíčí, čertvíčí, ledačí, ledasčí, lecčí, kdovíčí, nevímčí, kdečí.*	*ničí.*	*čí, čípak, číže.*	*jehož, jejíž, jejichž, čí.*	*svůj.*
ukazovací	*ten, tento, takový, tenhle, onen, onaký, týž, tentýž, takovýto, tenhleten, tamten, tuten, taký, tamhleten, tuhleten, tadyten, toť.*	--	--	--	--	--
vymezovací	*každý, sám, všechen, týž, tentýž, tatáž, totéž, titíž, tytéž, samý, všecek, všecken, všeliký, veškerý.*	--	--	--	--	--
ostatní	--	*ně(), bůhví(), čertví(), kdoví(), nevím(), kde(), lec(), leda(), ledas(), málo(), všelijaký, zřídka(), ()koli, ()koliv, ()žkoli, ()kolivěk, ()si, ()s.*	*nijaký, žádný, nižádný, pražádný.*	*jaký, který, ký, ()pak, ()že, ()ž.*	--	--

Tabulka 4.1: Kombinovatelnost hodnot kategorií Poddruh a Funkce u zájmen

	určitá	neurčitá	tázací
základní	*jedna, raz, dva, dvě, pět, sto, ...*	*několik, hodně, málo, poskrovnu, víc, dost ...*	*kolik*
úhrnné	*dvé, patero, tisícero, obé, ...*	*několikero*	*kolikero*
souborové	*dvoje, patery, tisícery, oboje, ...*	*několikery*	*kolikery*
druhové	*dvojí, paterý, tisícerý, obojí, ...*	*několikerý*	*kolikerý*
násobné	*dvakrát, pětkrát, (po)obakrát, ...*	*několikrát*	*kolikrát*
řadové	*druhý, pátý, ...*	*několikátý*	*kolikátý*
opakovací	*podruhé, popáté, ...*	*poněkolikáté*	*pokolikáté*
výčtové	*zaprvé, zadruhé, zapáté, ...*	*zaněkolikáté*	*zakolikáté*
dílové	*půl, čtvrt, třeť*	--	--

Tabulka 4.2: Kombinovatelnost hodnot kategorií Poddruh a Funkce u číslovek

	určitá	neurčitá	tázací a vztažná	záporná
místní	*tady, zde, tam, tudy, tamtudy, ...*	*někam, někudy, odněkud, ...*	*kam, kudy, odkud, ...*	*nikde, nikam, odnikud, ...*
časová	*teď, nyní, ...*	*někdy*	*kdy*	*nikdy*
způsobová	*tak, takto, ...*	*nějak*	*jak*	*nijak*

Tabulka 4.3: Kombinovatelnost hodnot kategorií Poddruh a Funkce u zájmenných příslovcí

4.1.4 Vid (ASP)

- D: dokonavý (*koupit, napsat, doručit, narodit se,...*)
- N: nedokonavý (*kupovat, psát, doručovat, chodívat,...*)
- O: obouvidý (*referovat, absolvovat, izolovat,...*)

Stejně jako dosavadní morfologické systémy, nezavádíme další hodnoty kategorie `Vid` pro iterativní slovesa, i když je, pokud je to možné, pravidelně vytváříme pomocí slovesných derivačních vzorů (viz 12.2.4). V případě potřeby by nebyl problém hodnoty doplnit.

V valenčním slovníku VALLEX (viz (Lopatková et al., 2006)) jsou vidové dvojice zpracovány jako jedno slovníkové heslo. V morfologickém slovníku to neděláme. Členy vidové dvojice považujeme za dvě různá slova. Bylo by však záhodno jejich slovníkové záznamy propojit zvláštním typem odkazu (viz také kapitolu 7).

4.1.5 Zkratka (ABR)

Tato kategorie je relevantní pro všechny slovní druhy.

Kategorie zkratka má pouze dvě hodnoty, a to:

- + : ano
- UNDEF

Kladnou hodnotu globální kategorie `Zkratka` dostávají zkratky, ostatní slovní tvary nemají tuto hodnotu definovanou. V brněnském systému je zkratka samostatným slovním druhem se značkou `kA`, v pražském systému se zkratky značkují pomocí 2. nebo 15. pozice.

Zkratka ve větě často zastupuje konkrétní slovní druh, a její prohlášení za samostatný slovní druh tak může komplikovat další zpracování. Pražské řešení nám proto připadá lepší, není však uplatňováno konzistentně, což je obecně slabá stránka obou zmiňovaných pozic v pražské morfologické značce.

Zkratka jako samostatná globální kategorie podle našeho názoru nejlépe postihne všechny možnosti, jak může zkratka vypadat.

Zkratka tedy může být libovolný slovní druh. V případě, že zkratka zastupuje jeden konkrétní slovní tvar, je slovním druhem této zkratky slovní druh tvaru, které zkratka zastupuje. Např. *č.* jako zkratka slovního tvaru *číslo* je podstatné jméno a jako takové jsou pro ni relevantní všechny kategorie relevantní pro ostatní podstatná jména. Zkratky, které nezastupují jedno slovo, ale celé slovní spojení, je třeba hodnotit individuálně. Zastupuje-li např. zkratka jmennou frázi (*USA, ODS*, apod.), je POS=N. U takových zkratek je možno stanovit i rod, číslo a pád podle toho, co zkratka zkracuje, nebo případně jak se používá. Je možno též využít hodnoty X (sdružená hodnota). Zastupuje-li zkratka složitější frázi (např. *atd., např.*), je POS=S (více viz zkratkové složeniny v kapitole 6).

4.1.6 Globální mutace (GMU)

Tato kategorie je relevantní pro všechny slovní druhy.

Podle naší definice se globální mutace projevují ve všech slovních tvarech paradigmatu. Z toho mimo jiné vyplývá, že globální jsou všechny mutace neohebných lemmat. Jde především o mutace příslovcí, která se nestupňují, tedy např. *zítra — zejtra*.

Mezi globální zahrnujeme i vokalizované mutace předložek (*od — ode*, *k — ke — ku*). Jsme si vědomi toho, že toto je zcela jiný typ mutací než např. mutace ortografické. Opět je třeba připomenout, že nám jde o co nejjednodušší popis, takže kategorie mutace, zde konkrétně mutace globální (GMU) využíváme k rozlišení slovních tvarů, jejichž ostatní relevantní kategorie nabývají hodnot totožných.

Zajímavější jsou slova ohebná. Mnoho globálních mutací je specifických, týkajících se jednoho konkrétního lemmatu. I zde však existuje několik systematických typů variantních dvojic, které se mohou uplatnit u mnoha, někdy dokonce všech lemmat určitých vlastností. Asi nejznámější jsou mutace cizích slov přejatých do češtiny, kde se v původním jazyce (většinou latina) píše *s*, ale v češtině vyslovuje *z*. Podle doporučení posledního vydání Pravidel českého pravopisu (viz (Pravidla)) existuje několik pravidel a mnoho výjimek, jak taková slova správně psát, ale uživatelé jazyka si s tím často hlavu nelámou a píšou různě. Tvary se *z* se většinou považují za spisovné, mutace se *s* za knižní nebo zastaralé, ale v textech se setkáme s oběma. Všechny takové dvojice musíme považovat za mutace. Nejznámější mutace se týkají přípon *-ismus — -izmus*. Zde však připouštějí Pravidla možnosti obě a používá se více mutace *-ismus*[7], a to i tehdy, když se vyskytuje ve slově s více možnými *s/z*, např. *izomorfismus* (15) — *isomorfismus* (2). Čísla v závorce udávají frekvenci v korpusu SYN2005.

Další systematické mutace vznikají již zmíněným přidáním protetického *v-* před lemmata začínající na *o-*.

Důležitou vlastností globálních mutací je to, že obě (všechny) mohou být použity při vytváření odvozenin. Např. z mutací podstatných jmen *okno — vokno* lze utvořit mutace přídavného jména *okenní — vokenní*. Tyto mutace jsou opět globální.

Z toho, co bylo řečeno o globálních mutacích, je zřejmé, že základní tvary globálních mutací jsou vždy variantními lemmaty vícenásobného lemmatu (viz též kap. 2). Neplatí to však naopak — existují i taková vícenásobná lemmata, jejichž prvky jsou flektivní mutace (např. vícenásobné lemma {*myslit, myslet*}), nebo nejsou mutacemi vůbec, což je případ všech složenin (viz kap. 6).

Tabulka 4.4 ukazuje hlavní typy českých globálních mutací, bez ohledu na jejich klasifikaci, to znamená, že nedělá rozdíl mezi kodifikovanými a nekodifikovanými mutacemi. Poslední sloupec tabulky uvádí kódy pro hodnoty kategorie `Globální mutace`. Kódy v horní části tabulky vyjadřují hláskovou změnu v mutacích. Kód `d` zastupuje „dlouhé“ mutace, `k` „krátké“. Podobně `m` znamená „měkké“, `t` „tvrdé“. Mutace, které se vymykají běžným typům, se označují čísly, jak ukazuje poslední řádek tabulky.

[7] Nepoměr mezi oběma mutacemi nás velmi překvapil: Nejrozsáhlejší obecně dostupný český korpus SYN obsahuje pouze necelých 2100 výskytů slov s lemmatem zakončeným sufixem *-izmus* oproti téměř 133 tisícům s lemmatem na *-ismus*.

Typ	Příklad	Hodnoty GMU
o — vo	*okno — vokno*	0 — v
ý — ej	*mýdlo — mejdlo*	0 — j
z — s	*klauzule — klausule*	z — s
t — th	*tema — thema*	0 — h
é — í	*kolébka — kolíbka*	e — i
é — ý	*okénko — okýnko*	e — y
á — e	*originální — originelní*	a — e
á — a	*Abrahám — Abraham*	d — k
é — e	*acetylén — acetylen*	
ó — o	*salón — salon*	
ý — y	*apetýt — apetyt*	
í — i	*alexandrín — alexandrin*	
ů — u	*přezůvky — přezuvky*	
ú — u	*Plútarchos — Plutarchos*	
s — š	*student — študent*	t — m
t — ť	*vlaštovka — vlašťovka*	
n — ň	*šnůra — šňůra*	
d — ď	*dolík — ďolík*	
e — ě	*Bardejov — Bardějov*	
z — ž	*zbrzďování — zbržďování*	
jiné	*Afganistan — Afghanistan*	0 — 1

Tabulka 4.4: Přehled nejčastějších typů globálních mutací s příklady

Hromadění typů mutací v jednom lemmatu se vyjádří vícero hodnotami, viz tabulka 4.5. Jsou zde naznačeny možné kombinace tří typů globálních mutací: s — z a d — k, přičemž dlouhá mutace se v tomto případě rozpadá na další dvě možnosti, a to ú — ů. Tento poslední typ není zcela typický, proto jsme ho nezahrnuli do tabulky 4.4. Mutaci s *-ů-* jsme tedy nechali jako „dlouhou“ s hodnotou d, mutaci s *-ú-* jsme označili číslicí 1.

Lemma	Hodnota GMU
bluza	kz
blůza	dz
blúza	1z
blusa	ks
blůsa	ds

Tabulka 4.5: Příklad vícera hodnot kagegorie `Globální mutace`

Pro lemmata, která se vyskytují v mnoha mutacích (většinou jde o cizí vlastní jména) je nejvýhodnější označit mutace pomocí čísel, přestože by někdy bylo možné i v těchto jménech vystopovat uvedené typy. Může totiž dojít k tomu, že se v jednom lemmatu uplatní jeden typ vícekrát. V tom případě by označování mutací mohlo být krkolomné.

Příkladem takových mutací je množina různých zápisů země *Afghánistán*, kde se projevuje typ „jiný“ a dva typy „a—á“. V korpusu SYN se vyskytuje ve všech osmi možných mutacích: *Afghánistán, Afgánistán, Afganistán, Afgha-*

nistán, Afghanistan, Afganistan, Afghánistan, Afgánistan.

Kdybychom chtěli i v takových případech rozlišovat typy globálních mutací, bylo by třeba jejich hodnoty udávat i s místem v konkrétním lemmatu, kde k rozlišení typu dochází. Z uvedeného příkladu je ale zřejmé, že takto podrobný popis globálních mutací by byl pravděpodobně zbytečný. [8]

4.2 Flektivní morfologické kategorie

Flektivní morfologická kategorie je taková kategorie, jejíž hodnoty se pro jednotlivé slovní tvary jednoho paradigmatu liší.

Následuje seznam flektivních morfologických kategorií:

- Rod — GEN
- Číslo — NUM
- Duál — DUA
- Pád — CAS
- Osoba — PER
- Stupeň — DEG
- Negace — NEG
- Slovesný tvar — VRB
- Jmenný tvar přídavných jmen — NOM
- Stupeň intenzity slovesného děje — INT
- Typ složeniny — CMP
- Flektivní mutace — FMU

Probereme je nyní jednu po druhé. Některé jsou tradiční a mají i tradiční hodnoty, jiné jsme zavedli nově, někde jsme dokonce pozměnili tradiční hodnoty. Vše je podřízeno snadnější využitelnosti při automatickém zpracování češtiny při zachování veškeré lingvistické informace, kterou hodnoty kategorií nesou.

Ve výčtu hodnot následujících kategorií bude na posledním místě občas vystupovat hodnota nazvaná „sdružená hodnota“. Tím se myslí libovolná z hodnot předcházejícího seznamu. Podobná hodnota již v pražském systému existuje. Tam se však bere jako proměnná, která případně může být nahrazena jednou z konkrétních hodnot (např. rod cizího slova podle přívlastkového rozvití). Při některých experimentech se s ní takto dokonce pracovalo — morfologická značka obsahující tuto hodnotu se rozepsala, aby v ní byly jen konkrétní hodnoty, čímž se rozpadla do množství jednoznačných značek. My ale chápeme sdruženou hodnotu odlišně. V našem pojetí znamená to, že hodnota dané

[8] Dokonce by se v podobných případech (tedy u nejednotného zápisu cizích vlastních jmen) dalo uvažovat o tom, že nebude splněno Zlaté pravidlo morfologie.

kategorie se nedá a nikdy nepůjde rozlišit. Případy, kdy může daná kategorie nabývat více hodnot, řešíme možností přiřazení více morfologických značek danému slovu. Tím se nahradí částečně sdružené hodnoty u kategorií rod, číslo a pád, které obsahuje pražský systém.

4.2.1 Rod (GEN)

Kategorie je relevantní pro podstatná jména, přídavná jména, některá zájmena, některé číslovky, přechodníky a slovesa v příčestí činném.

- M: mužský životný
- I: mužský neživotný
- F: ženský
- N: střední
- X: sdružená hodnota

Kromě tradičních hodnot zachováváme i již zavedenou korpusovou praxi z pražského i brněnského systému a rozlišujeme dva mužské rody. Pravděpodobně přirozenější by bylo mít jen jeden mužský rod a životnost vyjádřit pomocí další kategorie. Nechceme však zavádět nové kategorie zbytečně, když se zdá, že současné pojetí nečiní problémy.

4.2.2 Číslo (NUM)

Kategorie je relevantní pro podstatná jména, přídavná jména, některá zájmena, některé číslovky, slovesa a kondicionálové částice a spojky.

- S: jednotné
- P: množné
- X: sdružená hodnota

Tradiční hodnota duál, která byla zachována i na Konkláve, se v našem návrhu stává samostatnou kategorií.

4.2.3 Duál (DUA)

Kategorie je relevantní pro duálová podstatná jména (viz dále) a slova s adjektivním skloňováním.

- +
- UNDEF

V mluvnicích českého jazyka se o duálu mluví jako o třetí hodnotě kategorie `Číslo`. Aniž bychom chtěli jakkoli napadat tento fakt, navrhneme na duál jiný pohled, který usnadní automatické zpracování českého jazyka. Zdůrazňujeme, že nám jde o zjednodušení automatických analýz, nikoli o zpochybňování lingvistických tradic.

Motivace

Hodnota duál kategorie `Číslo` má nepříjemnou vlastnost, způsobuje totiž v některých případech neplatnost jedné ze základních syntaktických vlastností češtiny, a to shody v čísle.
Např. ve větách (32) a (33) z korpusu (P v závorce znamená plurál, D duál):

zájem však vzbudil svými(P) *dvěma*(D) *knihami*(P) (32)

hleděla na ni upřenýma(D), *nevidoucíma*(D) *očima*(D), *které*(P) *ani nezamrkaly*(P) (33)

Číslovka *dvěma* z prvního příkladu je v duálu (D), přestože zájmeno *svými* i substantivum *knihami* jsou v čísle množném (P). Vzhledem k tomu, že číslovka *dva/dvě* ve skutečnosti žádné množné číslo nemá, stalo se zvykem považovat všechny duálové tvary této číslovky za tvary množného čísla. Pokud přijmeme tuto tezi, shoda tu samozřejmě je.

Ve druhém případě bychom očekávali shodu v čísle mezi podstatným jménem *očima* a vztažným zájmenem *které*, uvozujícím vedlejší větu, podobně, jako ve větě (34):

Podívala se na něho brýlemi(P), *které*(P) *v tu chvíli byly*(P) *velmi jasné.* (34)

Shody bychom mohli dosáhnout dvěma způsoby:

1. buď připustíme, že slovní tvar *které* může být kromě množného čísla i v duálu,
2. nebo prohlásíme všechny duálové tvary za množné číslo.

První varianta by ovšem znamenala, že bychom museli pro všechna vztažná zájmena ve všech pádech připustit dvojí hodnotu — množné číslo a duál. Další důsledek by byl ještě revolučnější — kvůli shodě podmětu s přísudkem bychom museli totéž udělat se všemi slovesy v množném čísle. Tato varianta se tedy zdá nepřijatelná.

Druhé řešení nám připadá rozumnější, musíme ho ale rozšířit, protože úplná ztráta hodnoty duál by nám přinesla zase jiné problémy. Především by se ztratil rozdíl mezi tvary přídavných jmen, zájmen, číslovek a některých podstatných jmen s koncovkami *-mi* a *-ma* v 7pl. Zavádíme tedy novou kategorii Duál s jedinou hodnotou +.

Tuto hodnotu mají

- podstatná jména, která tvoří duální tvary v 7pl, a to ve všech pádech množného čísla. Jde o lemmata (v závorce je uveden tvar pro 7pl): *oko (očima), ucho (ušima), ruka (rukama), noha (nohama), očičko (očičkama), očko (očkama), ouško (ouškama), ručička (ručičkama), ručka (ručkama), nožička (nožičkama), nožka (nožkama)*. Nadále jim budeme říkat duálová slova. Mezi duálová slova nezařazujeme *koleno*, *rameno* ani *prso*, přestože také tvoří duálové tvary, a to v 2pl a 6pl. Ty však nezpůsobují problémy se shodou. Viz též dále.

- přídavná jména, zájmena a číslovky s adjektivním skloňováním včetně základních číslovek *dva, oba, tři a čtyři*, ale jen v 7. pádě množného čísla. Tyto slovní tvary mají v 7pl buď hodnotu kategorie DUA=UNDEF (např. pro koncovku *-mi* (*krásnými*)), nebo DUA=+ (např. pro koncovku *-ma* (*krásnýma*)). Dvě možnosti duálové hodnoty, totiž + a UNDEF, mají i tvary vztažných zájmen *jimiž*, *jejichž* a *jehož* v množném čísle, přestože se slovní tvary duálu neliší od prostého množného čísla. Viz příklady dále.

  Kladnou hodnotu duálu může mít v 7. pádě množného čísla i číslovka *tři*, tvar *třema*, přestože ji žádná nám známá mluvnice jako možnost neuvádí, viz např. 2. díl Mluvnice češtiny ((Akademická mluvnice)) na str. 405). Podle této mluvnice by tedy bylo správné např. poněkud absurdní spojení *třemi nohama*.

4.2.3.1 Odbočka ke shodě

I když naše práce pojednává o morfologii, považujeme za vhodné se na tomto místě dotknout otázky shody, abychom podrobněji vysvětlili právě navržené řešení duálu. Shoda v čísle se nyní rozpadá na shodu v čísle a shodu v duálu.[9]

Shoda v čísle se naším návrhem výrazně zjednoduší. Např. ve větě

Oči (P) *se dívaly* (P) (35)

zůstává zachovaná shoda v čísle a nevadí, že *oči* mají hodnotu Duálu + (nadále budeme značit D+) a *dívaly* nemají hodnotu Duálu žádnou (v následujících několika příkladech budeme tento stav značit D-).

V našich příkladech dostáváme:

zájem však vzbudil svými(P D-) *dvěma*(P D+) *knihami*(P D-) (36)
hleděla na ni upřenýma(P D+), *nevidoucíma*(P D+) *očima*(P D+), (37)
které(P D-) *ani nezamrkaly*(P D-)

Shoda v čísle je všude jednoduše zachována.
Shoda v duálu se vyžaduje pouze u výše vyjmenovaných slov, a to jen v 7pl.

tmavé (P D-) *oči* (P D+), ale *tmavýma* (P D+) *očima* (P D+) (38)
očima (P D+), *kterýma* (P D+) (39)
oči (P D+), *kterýma* (P D+) (40)
oči (P D+), *kterými* (P D-).. špatně, není shoda v duálu, přestože je (41)
to 7.pád.
očima (P D+), *které* (P D-).. dobře, není 7.pád (42)
očima (P D+), *jimiž* (P D+) (43)
jejichž (P D+) *očima* (P D+) (44)

[9] Nezabýváme se shodou v rodě, která nečiní potíže, i když v našich příkladech vystupují *oči*, které jsou v jednotném čísle rodu středního a v množném rodu ženského. Podobnou vlastnost mají i *uši*.

V posledních dvou příkladech (43) a (44) mohou mít vztažná zájmena samozřejmě i hodnotu kategorie `Duál` nedefinovanou, ale v těchto konkrétních kontextech je jejich hodnota kladná.

V příkladě (36) je sice tvar *dvěma* v duálu, ale shoda se nevyžaduje, neboť podstatné jméno *kniha* nepatří mezi duálová slova.

Pro úplnost je třeba zde dodat, že duál se projevuje ještě v 2pl a 6pl následujících tří lemmat středního rodu: *rameno (ramenou), koleno (kolenou), prso (prsou)*. Vzhledem k tomu, že se vyskytují i nespisovné tvary *ramenech, kolenech* a alternativní *prsech* (zde ale dochází k nerozlišitelné homonymii s podstatným jménem *prs*), stanovujeme i v těchto případech pro rozlišení kladnou hodnotu kategorie `Duál`, která však nemá pro shodu praktický význam. Shodu v duálu stačí požadovat skutečně jen u duálových slov v 7pl.

Příklady:

na svých (P D-) *kolenou* (P D+)
se svými (P D-) *koleny* (P D-)

4.2.4 Pád (CAS)

Kategorie je relevantní pro podstatná jména, přídavná jména, zájmena, číslovky a předložky. U předložek je význam kategorie `Pád` poněkud odlišný — vyjadřuje rekci. V tomto případě to sice není morfologická kategorie, avšak její přiřazení předložkám je velmi užitečné.

- 1 až 7
- X: sdružená hodnota

Hodnoty pádu jsou tradiční, není třeba se jimi dále zabývat.

4.2.5 Osoba (PER)

Kategorie je relevantní pro zájmena, slovesa, kondicionálovou částici *by* a kondicionálové spojky *aby*, *kdyby*.

- 1 až 3
- v: vykání

Hodnota `v` není tradiční. Ukazuje se, že při analýze textů činí značný problém rozpor mezi množným číslem zájmena *vy* (často psáno jako *Vy*) a jednotným číslem minulého příčestí slovesa (*vy jste mluvil, vy byste mluvil*), pasiva (*vy jste chycen*), případně přídavného jména po sponě (*vy jste sám, vy jste hezká*). Rosen a Saloni v (Rosen – Saloni, 2006) navrhují zahrnout tato tzv. honorifika do slovesného paradigmatu, kde se z nejasných důvodů tradičně neuvádějí. Stejný názor je vyjádřen i v práci Panevové ((Panevová, 2008)).

Navrhujeme tedy zavést kromě tří klasických hodnot kategorie `Osoba` ještě hodnotu čtvrtou, tzv. vykání. Přiřazujeme jí kód v.

Kromě hodnoty PER=2 mají tedy určitá zájmena *vy* a *váš* také hodnotu v (PER=v). Jde zde o pravidelnou homonymii. Stejná homonymie nastává i v prézentu sloves (*vy rádi zpíváte — vy rád zpíváte*).

Panevová se ve své práci (Panevová, 2008) zabývá ještě otázkou, zda může být honorativ také v množném čísle, jestliže vykáme celé skupině osob, případně jen někomu ze skupiny. Nenachází však na ni jednoznačnou odpověď.

Zájmena *vy* a *váš* jako honorifika (PER=v), tedy mohou být jak v jednotném, tak v množném čísle. Jsou-li v množném čísle, homonymii nelze vyřešit jinak než ze sémantiky kontextu. V příkladě (45) mají tedy tvary *vy* a *jste* jednoznačné hodnoty kategorie PER, neboť NUM=S, zatímco v příkladě (46) jsou možné hodnoty kategorie PER dvě.

vy (PER=v NUM=S) *jste* (PER=v NUM=S) *mluvil* (NUM=S) (45)

vy (PER=v/2 NUM=P) *jste* (PER=v/2 NUM=P) *mluvili* (NUM=P) (46)

4.2.6 Stupeň (DEG)

Kategorie je relevantní pro přídavná jména a příslovce.

- 1: pozitiv
- 2: komparativ
- 3: superlativ
- s: typ sebe + komparativ

U stupňování se vedou spory, zda patří do morfologie, nebo spíše do slovotvorby, viz např. obšírné pojednání (Karlík – Hladká, 2004) s argumenty pro obě zařazení. Pro morfologickou analýzu i syntézu je důležité, aby bylo možno zachytit všechny slovní tvary a přiřadit jim rozumné hodnoty. Zařazujeme tedy kategorii `Stupeň` mezi morfologické, a to flektivní kategorie.

První tři hodnoty, vyjádřené čísly, nepotřebují komentář.

Poslední hodnota, tedy s, však není běžná. Nikde se do stupňování nepočítá. Týká se slov typu *sebekrásnější, sebekrásněji*. Tvoření přídavných jmen a příslovcí tímto způsobem je však velmi pravidelné a týká se všech stupňovatelných lemmat. Je tedy přirozené zachytit takto vytvořené tvary s lemmatem společným i pro stupňované tvary.

V dosavadní praxi se tyto typy přídavných jmen a příslovcí lemmatizují jako samostatné jednotky, např. $\lambda(sebemenší) = sebemenší$. Mnoho takových slov např. v pražském morfologickém slovníku není, a tak zůstávají nerozpoznána. Vzhledem k naprosté pravidelnosti jejich tvoření, vysoké produktivitě a zjevné příslušnosti k pozitivu přídavného jména či příslovce je přirozené je začlenit do paradigmatu příslušného pozitivu. Kategorie `Stupeň` nám pro takový popis připadá nejvhodnější. Pokračovat v číselné stupňovací řadě, jak by se na první pohled mohlo zdát přirozené, nám však nepřipadá vhodné, neboť slova typu *sebekrásnější* nezapadají logicky do řady 1., 2. a 3. stupně. Mají totiž odlišný význam. Proto jsme zvolili pro tento typ kód mimo číselnou řadu[10].

[10] Ale jak už bylo několikrát zdůrazněno, není to podstatné. Jde samozřejmě o pouhý kód.

4.2.7 Negace (NEG)

Kategorie je relevantní pro slovesa, přídavná jména, příslovce a (v omezené míře i) pro podstatná jména (*hlava nehlava*).

Má tyto hodnoty:

- N: pro záporné slovní tvary, které začínají záporkou *ne-*
- A: pro ostatní slovní tvary

V případě negace dochází u některých ajdektiv, substantiv a sloves ke sporům, zda existují pozitivní tvary či nikoliv. Souvisí to i s lemmatizací — má být lemmatem záporných tvarů základní tvar v pozitivním, nebo negativním tvaru? Tento problém se týká i lemmatizace, zabývali jsme se jím tedy i v kapitole 2.

Uveďme několik dalších příkladů:

Substantivum *nepřítel* je sice opakem slova *přítel*, ale většinou se již takto nechápe, proto by mělo být lemmatizováno jako *nepřítel*, s hodnotou NEG=A.

Nemoc není opakem *moci*, i zde jde o dvě lemmata, obě s hodnotou NEG=A.

Podobně na tom je přídavné jméno *nesmyslný*. I zde jsou dvě lemmata — *smyslný* i *nesmyslný*, obě s NEG=A.

Jsou ale i slova, která jsou sporná, např. adverbium *nekale* (viz též oddíl 2.2.2).

Nejjednodušší by bylo technické řešení, kdy by se prohlásilo, že všechna slova s předponou *ne-* mají hodnotu kategorie NEG=N a lemma se rovná základnímu slovnímu tvaru bez předpony *ne-*. Uvědomujeme si však, že takové řešení by se setkalo s velkou nevolí na straně většiny uživatelů korpusů, proto necháme rozhodnutí na správci konkrétního slovníku.

4.2.8 Slovesný tvar (VRB)

Kategorie je relevantní pro slovesa, přídavná jména (pasivum), částice a spojky (kondicionál).

- P: indikativ prézentu (*kolíbá*)
- B: budoucí čas (*ponese, bude*)
- F: infinitiv (*otevřít*)
- I: imperativ (*peč*)
- L: příčestí činné (*strouhal*)
- T: příčestí trpné (*zavřen*)
- K: kondicionál (*aby, kdyby, by*)
- p: přechodník přítomný (*starajíc*)
- m: přechodník minulý (*vstoupiv*)

Podobnou množinu hodnot má i brněnský systém pro kategorii `Mód`. I Konkláve se rozhodlo, že nebude kódovat tradiční kategorie sloves jako je Čas a Slovesný rod, protože existuje jen několik málo smysluplných kombinací hodnot těchto kategorií. V pražském systému to byly kombinace hodnot tří kategorií, a to Detailní určení slovního druhu (SUBPOS — pozice 2), Čas (TENSE — pozice 9) a Aktivum/pasivum (VOICE — pozice 12). Jediné kombinace, které se mohly vyskytnout, uvádí tabulka 4.6.

	Číslo pozice			Stručná vysvětlivka
	2	9	12	
⋆	B	P	A	přítomný čas
⋆	B	F	A	budoucí čas
⋆	f	-	-	infinitiv
⋆	i	-	-	imperativ
⋆	p	R	A	příčestí činné (včetně přidaného *-s*)
	s	H	P	pasivní příčestí se zakončením *-s*
⋆	s	X	P	pasivní příčestí
⋆	e	-	-	přechodník přítomný
⋆	m	-	-	přechodník minulý
	c	-	-	kondicionál slovesa *být*
	q	R	A	min. čas archaický (*vstalť*)
	t	F	A	archaický budoucí čas s *-ť* (*buduť*)
	t	P	A	archaický přítomný čas s *-ť* (*dávámť*, ale i *poradímť*)

Tabulka 4.6: Jediné možné kombinace 2., 9. a 12. pozice pražského systému.

Kombinace hodnot, u nichž je v tabulce 4.6 uvedena hvězdička (⋆), jsme sdružili do jedné hodnoty nové kategorie, kterou nazýváme `Slovesný tvar`.

Řádky tabulky bez hvězdičky probereme postupně:

Řádka s názvem „pasivní příčestí se zakončením *-s*“ zahrnuje slova typu *rytas, propuštěnas*. V korpusu SYN se 600 miliony slov jsme takových slov našli 121, avšak všechny jsou buď špatně označkovány, nebo jde o překlepy. Kdyby se někdy takový tvar vyskytl, bude se jednat o slovesnou složeninu (viz kapitolu o složeninách 6).

Řádka, nazvaná „kondicionál slovesa *být*“, označuje slovní tvary *by, bys, bych, bychom, byste*. Tyto slovní tvary pojednáváme ve zvláštní kapitole 5 spolu s dalšími dvěma kondicionálovými lemmaty *aby* a *kdyby*. Na tomto místě je pouze třeba upozornit, že nově zavedená hodnota Kondicionál (K) není relevantní pro slovesa, ale pro spojky (*aby, kdyby*) a částice (*by*).

Poslední tři řádky, kde je na druhé pozici q nebo t, jsou archaické, s *-ť* na konci slovního tvaru. Zde není třeba zvláštní značky. Rozdíl oproti slovnímu tvaru bez *-ť* zachycujeme pomocí kategorie Flektivní mutace FMU (viz kap. o mutacích 3).

Místo uvedených tří kategorií (pozic pražského systému) jsme se tedy rozhodli zavést jednu, jejíž hodnoty budou odpovídat všem jejich možným kombinacím. Kupodivu je jich jen 8 — ty, co jsou v tabulce 4.6 označeny hvězdičkou. Tyto kombinace se tedy staly hodnotami nové kategorie nazvané `Slovesný tvar`. K nim přidáváme devátou hodnotu Kondicionál, která je však odlišná od kondicionálu, uvedeného v tabulce 4.6 na 4. řádku zdola.

Na chvíli se zastavme u některých hodnot kategorie `Slovesný tvar`.

Příčestí trpné Tato hodnota kategorie `Slovesný tvar` se neurčuje u slovního druhu sloveso. V našem novém návrhu jsme totiž všechna příčestí trpná zařadili mezi jmenné tvary přídavných jmen. Formálně se tak chovají a často je velmi obtížné rozlišit, zda jde o příčestí trpné slovesa, nebo o jmenný tvar přídavného jména. Pro přídavná jména hovoří i fakt, že u příčestí trpného se může měnit, byť velmi omezeně, pád. Ve všech rodech i číslech může vystupovat v akuzativu, jak ukazují příklady (47) až (54):

Za hodinu jsme měli připravenu hromadu klestí (47)
... měl najatu restauraci (48)
Hranol má hranu podstavy rovnu a=24 cm (49)
... máme... hotovu dokumentaci (50)
Psychotesty již máme hotovy (51)
Základní návrh chceme mít hotov v březnu. (52)
Musíme mít připraven mírový plán (53)
Budeme mít připravena i vodní děla. (54)

Přesto, že příčestí trpné od slovesa vzniklo, tvoří protiklad k příčestí činnému a většinou vystupuje ve větě v její přísudkové části, formálně je možné ho vždy nahradit jmenným tvarem přídavného jména. Vzhledem k obtížné rozlišitelnosti je jednodušší, když ho vždy považujeme za přídavné jméno.

Kvůli zachování těsné vazby ke slovesu však vyplňujeme u těch jmenných tvarů přídavných jmen, která jsou zároveň příčestím trpným, i tuto hodnotu kategorie `Slovesný tvar`. Například tedy slovní tvar *otevřen* má tyto morfologické hodnoty:

POS=A, (Slovní druh: přídavné jméno)
SUB=S, (Poddruh: deverbativní)
GEN=M/I, (Rod: mužský životný nebo neživotný)
NUM=S, (Číslo: jednotné)
CAS=1, (Pád: 1)
VRB=T, (Slovesný tvar: příčestí trpné)
NEG=A, (Negace: pozitiv)
lemma=*otevřený*

Bude-li třeba vytvořit dotaz na vyhledání všech sloves, včetně příčestí trpných, lze to učinit takto: POS=V $\vee$ (POS=A $\wedge$ VRB=T)[11].

Složené slovesné tvary Složené slovesné tvary neurčujeme jako celek. Jsme si vědomi, že by bylo vítané, kdybychom rozpoznali ve větě složené slovesné tvary, ale tato úloha přesahuje rámec morfologie. Jak už jsme zdůraznili na začátku, zabýváme se jednotlivými slovy, nikoli jejich kombinacemi. Z toho důvodu také nemohou mít nedokonavá slovesa nikdy hodnotu kategorie VRB=B (budoucí čas).

[11] Vyhledají se všechna slovesa (POS=V) a ta přídavná jména (POS=A), která jsou současně slovesným trpným rodem (VRB=T).

Budoucí čas Hodnota budoucí čas se týká jen tvarů budoucího času slovesa *být* (tedy *budu, budeš, bude, budeme, budem, budete, budou*) a dokonavých sloves tvořících budoucí čas pomocí předpony *po-*.

Všechna ostatní dokonavá slovesa, byť sémanticky vyjadřující budoucí čas, mají hodnotu indikativ prézentu (VRB=P). Tedy: *nesu* (VRB=P), *ponesu* (VRB=B), ale *přinesu* (VRB=P).

Nedokonavá slovesa tvoří budoucí čas pouze ve složených slovesných tvarech, ta tedy nemají VRB=B nikdy (viz předchozí odstavec).

Kondicionál Tato hodnota popisuje tvary částice *by* a spojek *kdyby* a *aby*. Podrobně o kondicionálu pojedáváme v kapitole 5.

Přechodníky Přechodník přítomný určujeme pouze u nedokonavých sloves, přechodník minulý pouze u sloves dokonavých. U obouvidých sloves se mohou vyskytovat oba typy přechodníků. Podobně přechodníky pomocného slovesa *být* mohou být buď přítomné (*jsa, jsouc, jsouce* mají VRB=p), nebo minulé (*byv, byvši, byvše* mají VRB=m).

Oproti současnému pražskému systému vypouštíme možnost tvoření přechodníku přítomného pro dokonavá slovesa, jakožto zastaralý a už dávno neužívaný slovesný tvar. Při morfologické analýze jsme schopni takový tvar rozpoznat a správně určit pomocí guessru.

4.2.9 Jmenný tvar přídavných jmen (NOM)

Kategorie je relevantní pro přídavná jména.

Možné hodnoty:

- J: jmenný tvar
- UNDEF

Jmenné tvary přídavných jmen jsou v pražském systému popsány pomocí 2. pozice SUBPOS. Při zběžném pohledu by se mohlo zdát, že jmenný rod je poddruhem přídavného jména. Toto zařazení bylo přijato i na Konkláve. V tom případě bychom však jmenný rod nemohli zahrnout pod společné lemma dlouhého tvaru, neboť poddruh je kategorie globální. Jestliže chceme, aby např. $\lambda(sláb) = \lambda(slabý) = \{slabý\}$, musíme tuto kategorii vyčlenit zvlášť.

Mezi jmenné tvary přídavných jmen počítáme i tvary příčestí trpného sloves, neboť je často velmi obtížné je od sebe rozlišit. Slovní tvar *ukryt* mohl být odvozen jak z přídavného jména *ukrytý*, tak ze slovesa *ukrýt*. Podobně slovní tvar *spokojen* můžeme chápat jako odvozeninu od *spokojit* i od *spokojený*. Vzhledem k tomu, že se ve větě chovají tyto tvary stejně jako jmenné tvary ostatních přídavných jmen jmenných (např. *mlád*), mají dokonce omezenou flexi (4. pád všech rodů i čísel), zařazujeme je do této kategorie. Z toho speciálně vyplývá, že jejich slovní druh není sloveso, ale přídavné jméno.

Abychom však zachytili v popisu jejich slovesný charakter, je pro ně relevantní flektivní morfologická kategorie `Slovesný tvar`, a to s hodnotu T (příčestí trpné). Uvědomujeme si, že toto řešení není zcela v souladu s běžným chápáním slovesného trpného rodu, ale podařilo se nám tak jednoznačně

popsat sporné případy, kdy není jasné, zda jde o přídavné jméno či o sloveso, aniž by se tím ztratila jakákoliv informace. Viz též 4.2.8.

4.2.10 Stupeň intenzity slovesného děje (INT)

Kategorie je relevantní pro slovesa[12].

V oddíle 2.2.1.2 jsme upozornili na pravidelné tvoření zvratných podob nedokonavých sloves pomocí určitých předpon. Ukázali jsme, že s takto vytvořenými tvary by se nemělo zacházet jako se samostatnými slovesy, a zařadili jsme je pod lemma jejich neprefigovaného základního slovesa. Pomocí kategorie INT odlišíme tvary základního slovesa od z něho odvozených zvratných sloves prefigovaných.

Hodnoty jsou:

- r: pro předponu *roz-*
- p: pro předponu *po-*
- z: pro předponu *za-*
- n: pro předponu *na-*
- v: pro předponu *vy-*
- u: pro předponu *u-*

4.2.11 Typ složeniny (CMP)

Tato kategorie je relevantní jen pro složeniny a pojednáme o ní v kapitole 6 o složeninách.

4.2.12 Flektivní mutace (FMU)

Flektivní mutace se projevují především pomocí koncovek a jsou většinou systematické. Podle definice se nikdy netýkají celého paradigmatu, vždy jen některých kombinací hodnot gramatických kategorií. Systematické hodnoty kategorie FMU jsou z velké části zahrnuty přímo do ohýbacích vzorů, pojednáme o nich tedy v kapitolách o vzorech. Mezi flektivní mutace zahrnujeme i nekodifikované koncovky, které se však běžně používají, takže by se do systému paradigmat měly zahrnout jako varianty (mutace) koncovek spisovných. Na tento fakt poukázali už v roce 1992 Sgall a Hronek (viz (Sgall – Hronek, 1992)). Příkladem je koncovka *-ma* v 7pl všech skloňovaných slov, nebo používání *-ej-* u tvrdého adjektivního skloňování. V současném pražském systému nespisovné varianty většinou zahrnuty jsou a my je zachováváme. Lišíme se jen v jejich značení.

K mutacím dochází navíc u všech slov, jejichž skloňování kolísá mezi dvěma vzory stejného rodu, např. *stroj* a *hrad*, *muž* a *pán*, *kost* a *píseň*. Některé kombinace morfologických kategorií vytvoří podle obou vzorů stejný slovní tvar, u jiných je tvar odlišný. Právě tehdy je kategorie FMU relevantní.

[12] Dalo by se uvažovat i o slovesných odvozeninách — deverbativech.

Toto však není případ lemmat s kolísajícím rodem (např. *kredenc*), viz pojednání (Brabcová, 2004). Přestože se z lingvistického hlediska může jednat o varianty, není třeba zde tvary odlišovat pomocí kategorie FMU, neboť jsou rozlišeny hodnotou kategorie `Rod`.

Systematické případy variant (mutací) jsou uvedeny v kapitolách o vzorech. Nechceme tvrdit, že jsme na žádnou flektivní mutaci nezapomněli. Vyčerpávající přehled české morfologie ani nebyl cílem této práce. Nabízíme však způsob, jak konzistentně flektivní mutace zachytit. Není problém neuvedený typ flektivní mutace do seznamu zahrnout. Při kódování jejího typu stačí dodržet podmínku jednoznačné hodnoty v rámci konkrétní kombinace morfologických kategorií, jichž se mutace týkají.

Nesystematické mutace frekventovaných slov mohou mít své vlastní, specifické hodnoty, ostatní navrhujeme značit pomocí číslic. Existuje-li např. lemma X, u něhož se vyskytnou dva různé slovní tvary X_1 a X_2 se stejnými hodnotami všech relevantních morfologických kategorií, můžeme položit FMU=1 pro X_1 a FMU=2 pro X_2, jestliže typ mutace je neobvyklý a není pokryt standardním číselníkem.

Pro ilustraci se na tomto místě se ještě zmiňme o několika nepravidelných, leč četných mutacích.

Jde o tvary slovesa *jít*, které v přítomném a budoucím čase a v imperativu ztrácejí počáteční *j-*, např. *jdu — du*, *půjdu — pudu*, *jděte — děte*. Přiřazujeme jim FMU=g. V první osobě množného čísla se zde dokonce kombinují dva druhy mutace, oba flektivní. Jeden je pravidelný, tedy ztráta koncového *-e*, druhý nepravidelný, ztráta počátečního *j-*: *jdeme, deme, dem, jdem*. Tato kombinace hodnot gramatických kategorií lemmatu *jít* má tedy 4 různé flektivní mutace, jednu spisovnou, ostatní nespisovné.

Jiné nepravidelné flektivní mutace jsou např. *míň* a *méně* pro lemma *málo* a *líp*, *lépe* pro lemma *dobře*. Zde můžeme využít hodnot FMU=K pro kratší tvary *míň, líp* a FMU=D pro delší tvary *méně, lépe*.

Mezi flektivní mutace zařazujeme i mutace skloňování osobních zájmen. Jde o krátké a dlouhé tvary (*tebe — tě, mne — mě, mně — mi, jeho — jej*). Odlišujeme je opět hodnotami FMU=D a FMU=K.

Polský morfologický systém, vytvořený pro morfologickou anotaci korpusu IPI PAN (viz (Przepiórkowski, 2004)) zavádí pro zachycení této variability zvláštní morfologickou kategorii „Accentability“. My jsme se rozhodli využít kategorie `Flektivní mutace`. Důvodem je především určitá šetrnost — zavedení nové kategorie nám připadá zbytečné, když je možno využít v tomto případě jinak nevyužitou kategorii FMU, která je i tak velmi nesourodá.

Osobní zájmena pro 3. osobu jednotného i množného čísla, tedy *on, ona, ono, oni, ony* také mají mutace. Předchází-li těmto zájmenům předložka, mění se ve všech pádech počáteční *j-* na *n-*, případně *je-* na *ně-*. Dostáváme tedy dvojice *jeho — něho, jej — něj, ji — ni, jimi — nimi* atd. s hodnotami FMU=J a FMU=N.

I tyto mutace řeší v Polsku pomocí zvláštní kategorie, „Post-prepositionality“. Důvod, proč jsme zvolili řešení pomocí kategorie `Flektivní mutace` ,je stejný jako v předchozím případě.

Pro 4. pád jednotného čísla lemmatu *on* tak máme dokonce 5 různých slov-

ních tvarů: *ho, jeho, něho, jej, něj*[13], s hodnotami po řadě FMU=K, Dj, Dn, Kj, Kn).

Při označování hodnot kategorie `Flektivní mutace` většinou nepřiřazujeme hodnotu mutacím, které se užívají v psaném textu nejběžněji. Považujeme je za tzv. nulové mutace. Alternativní přístup by mohl všem takovým mutacím přiřadit FMU=0, což ale považujeme za zbytečné.

Námitka, že právě popsaný způsob označování mutací je složitý, je správná. Vzhledem k tomu, že jde o problém velmi rozsáhlý a mnohotvárný, domníváme se, že jednoduché řešení ani neexistuje. Naším cílem bylo navrhnout jednoznačné rozlišení mutací slovních tvarů a lemmat, aby vždy mohlo být splněno Zlaté pravidlo morfologie. Tohoto cíle jsme zřejmě dosáhli.

Pokud by si uživatelé slovníku, potažmo korpusů pomocí slovníku anotovaných, přáli mít hodnoty mutací jiné, je možno navržené hodnoty ohodnotit podle nějakého kritéria (např. podle kritizovaného stylového příznaku) a toto hodnocení vložit jako hodnoty do speciální nové kategorie. Tím by se různé hodnoty mutací sdružily do několika tříd, podle přání uživatelů.

Poznámka: Syntaktický slovní druh (SYN)

Na Konkláve se hovořilo též o kategorii `Syntaktický slovní druh`. Byla vymezena poměrně vágně:
`Syntaktický slovní druh` je kategorie, která vyjadřuje, jak se dané slovo obvykle chová v rovině povrchové syntaxe.
Tato kategorie měla usnadnit práci na pravidlové desambiguaci,

V novém návrhu od zavedení této kategorie v rámci morfologie upouštíme, neboť ve skutečnosti nejde o kategorii morfologickou. Neslouží totiž k popisu slovních tvarů, jedná se o kategorii syntaktickou. Je sice pravda, že některé kategorie, které jsme již zavedli, také nejsou zcela morfologické (např. kategorie `Poddruh`), ale pomáhají při popisu jednotlivých slovních tvarů. Kromě toho také rozlišují relevantnost dalších kategorií. Kategorie `Syntaktický slovní druh`, která měla být relevantní pro všechny slovní druhy, však žádné rozlišení na morfologické rovině nepřináší.

4.3 Morfologická značka

Morfologická značka je kód, pomocí něhož lze jednoznačně určit hodnoty všech relevantních morfologických kategorií pro daný slovní tvar.

V úvodní kapitole jsme prohlásili, že se konkrétní podobou morfologické značky zabývat nechceme. Kdybychom chtěli být důslední, tento oddíl o morfologické značce bychom vůbec do své práce nezařazovali. Morfologické kategorie a jejich hodnoty, jak jsme je zavedli v předchozích oddílech, je možné použít k vytvoření jakéhokoli kódu. Abychom si zjednodušili práci s vyjmenováváním kategorií a jejich hodnot v následujících kapitolách, přece jen morfologickou značku zavedeme.

Následující popis morfologické značky je tedy třeba brát jako příklad, jak také je možno morfologické kategorie kódovat. Pro konkrétní aplikace je možno použít jen některé morfologické kategorie, např. jen kategorie POS. Takovým

[13] Tvary typu *doň, zaň* sem nepočítáme, to jsou podle nového návrhu složeniny s jiným lemmatem.

kódům už nebudeme říkat morfologická značka, protože nekódují všechny relevantní morfologické kategorie, ale ve speciálních případech mohou být i takové kódy užitečné.

Morfologická značka musí být jednoznačná, aby bylo možno podle ní rozlišit rozdílné hodnoty morfologických kategorií. Teoreticky by bylo možno všechny kombinace hodnot morfologických kategorií nějakým způsobem očíslovat a jako kódu používat čísel. Takové kódování je samozřejmě nežádoucí, neboť není ani trošku intuitivní. Předpokládáme, že hodnoty morfologických kategorií se využijí k sestavení takového kódu, ze kterého půjdou jednoduše vyčíst.

Základní typy morfologických značek používaných pro češtinu jsou:

- poziční
- kompaktní
- hodnotový

Příkladem poziční značky je dosavadní morfologická značka pražská. Má jednotnou délku 15 pozic a každá její pozice kóduje jednu konkrétní kategorii (viz (Hajič, 2004)). Je zřejmé, že poziční značky „plýtvají místem“, protože neexistuje slovní tvar, pro který by bylo všech 15 kategorií relevantních. Na druhou stranu se s ním dobře pracuje, neboť každá kategorie má ve značce své neměnné místo.

Pražský systém používá i tzv. kompaktní značky (viz též (Hajič, 2004)), které jsou utvořeny tak, aby obsahovaly pouze hodnoty relevantních morfologických kategorií, a přitom zůstaly jednoznačné. Tento typ značek šetří místem, je však méně přehledný.

Příkladem hodnotového typu je značka brněnská (viz (Sedláček, 1999)). Její délka je proměnlivá, avšak vždy sudá, neboť obsahuje vždy dvojici ⟨název kategorie, její hodnota⟩ pro každou morfologickou kategorii relevantní pro daný slovní tvar.

V našem návrhu morfologické značky použijeme poziční typ a typ hodnotový. Poziční typ nám připadá přehlednější pro zachycení hodnot morfologických kategorií. Hodnotový typ má proti pozičnímu tu výhodu, že je snadné ho rozšířit tak, aby mohly mít kategorie více hodnot. Proto ho využijeme pro kódování hodnot mutací, a to jak globálních (kategorie GMU), tak flektivních (kategorie FMU).

Morfologickou značku rozdělíme na dvě části, globální a flektivní. Definovali jsme celkem 6 globálních morfologických kategorií a 12 flektivních. Z tohoto celkového počtu vyjmeme kategorie `Globální mutace` a `Flektivní mutace`, protože ty kódujeme jinak (viz kap. 3). Zbylých 5 globálních a 11 flektivních kategorií zakódujeme do poziční značky takto: 1. až 5. pozice tvoří globální část morfologické značky a obsahuje hodnoty těchto kategorií:

1. POS
2. SUB
3. FCE
4. ASP
5. ABR

Druhá, flektivní část značky, tedy pozice 6 až 16, obsahuje po řadě hodnoty těchto kategorií:

6. GEN
7. NUM
8. CAS
9. DUA
10. PER
11. DEG
12. NEG
13. VRB
14. NOM
15. CMP
16. INT

4.4 Relevantnost kategorií

V kapitole 4 jsme vyjmenovali všechny kategorie, které popisují v našem pojetí české slovní tvary. Ne všechny kategorie jsou však relevantní pro všechny slovní tvary. Relevantnost kategorie závisí především na kategorii `Slovní druh`, často ještě na kategorii `Poddruh` a na kategorii `Slovesný tvar`.

Množiny kategorií relevantních pro jednotlivé dvojice hodnot POS a SUB jsou přehledně zachyceny v tabulce 4.7. Číslování v první řádce se odkazuje k pozicím v návrhu morfologické značky. Relevantní kategorie jsou vyplněny znakem $\oplus$. Znak $\ominus$ v některých buňkách tabulky znamená, že tato kategorie je relevantní pouze někdy:

U zájmen záleží nejen na konkrétní kombinaci hodnot kategorií `Poddruh` a `Funkce`, ale i na konkrétním lemmatu, která kategorie je relevantní. Např. u substantivního určitého lemmatu *já* není zvykem určovat rod, zatímco u lemmatu *on*, které je rovněž substantivní určité, ano. Pouze kategorie `Pád` je relevantní pro všechna zájmena.

Podobné je to u číslovek. Číslovky 1 až 4 mohou vyjadřovat duál, ostatní substantivní určité číslovky ne. Číslovky 1 a 2 navíc vyjadřují i rod.

Číslovka řadová *první* má jako jediná relevantní kategorii `Stupeň` (viz příklady (55) a (56)[14]), což je naznačeno znakem $\ominus$ v příslušné buňce tabulky.

dle volebních preferencí čím dál prvnější (55)
jeho otec byl nejprvnějším dělníkem v přístavě (56)

Dále je to kategorie `Slovesný tvar` v případě spojek. Ta je relevantní pouze pro spojky *aby* a *kdyby*.

Pro slovní druh Složeniny jsme použili znak $\ominus$ u všech kategorií kromě `Typ složeniny`, která je naopak relevantní právě pouze pro složeniny. Relevantnost ostatních kategorií závisí právě na ní. Podrobná tabulka týkající se složenin je uvedena v kapitole o složeninách na str. 67.

[14] V pražském slovníku jsou tyto tvary klasifikovány jako přídavná jména.

1 POS	2 SUB	3 FCE	4 ASP	5 ABR	6 GEN	7 NUM	8 CAS	9 DUA	10 PER	11 DEG	12 NEG	13 VRB	14 NOM	15 CMP	16 INT
N	S		⊕	⊕	⊕	⊕	⊕	⊕			⊕				
	0			⊕	⊕	⊕	⊕	⊕			⊕				
A	U			⊕	⊕	⊕	⊕	⊕							
	GM		⊕	⊕	⊕	⊕	⊕	⊕			⊕				
	S			⊕	⊕	⊕	⊕	⊕		⊕	⊕	⊕	⊕		
	ostatní			⊕	⊕	⊕	⊕	⊕		⊕	⊕		⊕		
P		⊕		⊕	⊖	⊖	⊕	⊖	⊖						
C	1	U		⊕	⊖	⊕	⊕	⊖							
	1	ostatní		⊕			⊕								
	usd	⊕		⊕		⊕	⊕	⊕							
	r	⊕		⊕	⊕	⊕	⊕	⊕		⊖					
	p	⊕		⊕	⊕	⊕	⊕								
	nov	⊕		⊕											
V			⊕	⊕	⊕	⊕					⊕	Lpm			⊕
			⊕	⊕		⊕			⊕		⊕	PBI			⊕
			⊕								⊕	F			⊕
D	zájmenná	⊕		⊕											
	ostatní			⊕						⊕	⊕				
R				⊕			⊕								
J				⊕								⊖			
I															
T	c				⊕							K			
	7								2						
	ostatní														
S	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊕	⊖
F				⊕	⊕	⊕	⊕								
G				⊕											

Tabulka 4.7: Relevantní kategorie v závislosti na kategoriích `Slovní druh` a `Poddruh`

Příklady Několik příkladů morfologické značky je uvedeno v tabulce 4.8.

Slovní tvar	Morf. značka	Mutace	Lemma
okénko	NO---NS1---A----		{*okénko*,
vokýnko	NO---NS1---A----	Gvy	*okýnko*,
okénky	NO---NP7---A----		*vokýnko*}
vokýnkama	NO---NP7---A----	GvyFa	
pánové	NO---MP5---A----	Fv	*pán*
páni	NO---MP5---A----	Fi	
neutřen	AS-D-IS1---NTJ--		*utřený*
nes	VO-N--S--2-AI---		{*nést*,
pones	VO-N--S--2-AI---	Fp	*nésti*}
doběhla	VO-D--S--3-AL---		*doběhnout*
doběhnula	VO-D--S--3-AL---	Fn	
něj	PZU--MS2-3------	FKn	*on*
jeho	PZU--MS2-3------	FDj	
jeho	PUU--XXX-3------		*jeho*

Tabulka 4.8: Příklady morfologického popisu slovních tvarů

5 Kondicionál

Kondicionál se v češtině vyjadřuje pomocí „slůvek“ *by, aby* a *kdyby*. Už samo jejich zařazení mezi slovní druhy je problematické a vedou se o něm stále spory. Neděláme si zde nárok na jejich vyřešení, ale pro automatické zpracování češtiny je třeba tato slova nějak klasifikovat.

Mohlo by se zdát, že nemá smysl zavádět speciální kategorii pro pouhá tři lemmata, když je lze zadat výčtem. Ze strany uživatelů korpusů a ostatních lingvistů se však ukazuje potřeba mít jejich kondicionálovou povahu zvlášť vyznačenu.

V brněnské morfologii pro ně byl vytvořen nový slovní druh se značkou `kY`. Nemá speciální název.

Pražská morfologie řadí *by* (historicky) mezi slovesa, *kdyby* a *aby* (tradičně) mezi spojky. Od ostatních sloves se *by* odlišuje zvláštní hodnotou na 2. pozici (poddruh). Tvary *aby* a *kdyby* se odlišují od ostatních spojek pouze ve svých časovaných tvarech.

Ve slovenštině řeší podobný problém zavedením kategorie „kondicionálnosť“, která je relevantní pro slovní druh spojka a částice.

Slovenské řešení nám připadá nejvhodnější, ale rozhodli jsme se pro jeho odlišné zpracování.

Kondicionál se projevuje výhradně ve spojení se slovesem, pomocí něj se tvoří podmiňovací způsob. Proto ho zařazujeme jako novou hodnotu kategorie `Slovesný tvar`. Slovnědruhové zařazení však neměníme, takže se kategorie `Slovesný tvar` stává (trochu překvapivě) relevantní pro slovní druhy spojek a částic. Hodnota kondicionál se vlastně netýká žádného slovesa, ale právě jen spojek a částic, a to jen oněch tří vyjmenovaných v úvodu této kapitoly.

Zvláštní povaha těchto tří kondicionálových slov se projevuje už tím, že ač řazena mezi neohebné slovní druhy (částice a spojky), mají vyjádřenu osobu a číslo.

V tabulce 5.1 jsou přehledně zpracovány hodnoty všech relevantních morfologických kategorií pro paradigmata všech tří kondicionálových slov.

Dvojí hodnotu ve sloupci kategorie `Číslo` a `Osoba` u slovních tvarů *aby, kdyby, by* nelze libovolně kombinovat. Možné kombinace čísla a osoby jsou pouze S2 (2. osoba singuláru, viz příklady (57) až (59)), S3 (3. osoba singuláru, viz (60) až (62)) a P3 (3. osoba plurálu, příklady (63) až (65)). Kombinace P2 (2. osoba plurálu) možná není.

záleží na tom, aby sis udržel chladnou hlavu (57)
Chováš se, jako kdyby ses narodila v jiném století (58)
Myslím, že by ses měl připravit na menší šok (59)

Přemýšlí o muškách, které by si dala k snídani (60)
A kdyby se za něj provdala, (61)
Nechci, aby to někdo věděl (62)

Lemma	Slovní tvar	POS	SUB	VRB	NUM	PER	FMU
aby	*aby*	J	,	K	SP	23	—
	abych	J	,	K	S	1	—
	abys	J	,	K	S	2	—
	abychom	J	,	K	P	1	0
	abysme	J	,	K	P	1	1
	abyste	J	,	K	P	2	—
kdyby	*kdyby*	J	,	K	SP	23	—
	kdybych	J	,	K	S	1	—
	kdybys	J	,	K	S	2	—
	kdybychom	J	,	K	P	1	0
	kdybysme	J	,	K	P	1	1
	kdybyste	J	,	K	P	2	—
by	*by*	T	c	K	SP	23	—
	bych	T	c	K	S	1	—
	bys	T	c	K	S	2	—
	bychom	T	c	K	P	1	0
	bysme	T	c	K	P	1	1
	byste	T	c	K	P	2	—

Tabulka 5.1: Relevantní kategorie a jejich hodnoty kondicionálových slov

A to mě přivádí k myšlence, jaké by to bylo, kdyby v zoologických byly taky ukázky lidí. (63)

jako by tomu Jihoafričané nechtěli věřit (64)

pokřikovali na muzikanty, aby zase hráli (65)

6 Složeniny

Složenina popisuje slovní tvar, který zastupuje dva nebo více slovních tvarů (**složek** složeniny) a většinou mu není možné přiřadit jednoduše slovní druh. Příklady: *naň = na něj, byls = byl jsi*.

Složenina vzniká spojením těchto slovních tvarů, nejde však o prosté zřetězení. Většinou lze složeninu ve větě původními slovními tvary nahradit, aniž by se změnil smysl věty.

Složenině nelze přiřadit žádný z klasických slovních druhů, až na slovesnou složeninu typu V (viz dále). To byl důvod, proč jsme složeninu zavedli jako samostatný slovní druh. Toto řešení bylo přijato už na Konkláve, ač pod jiným názvem. Rozdělení do typů a jejich přesné vymezení provedla autorka této práce.

Složeninou v našem pojetí však nejsou slova vzniklá jedním z tradičních způsobů slovotvorby, která mají svůj vlastní význam a již se zařadila jednoznačně do některého slovního druhu. Složeninami například nejsou slova *černobílý, novotvar, spolupořádat*, přestože také vznikla složením různých slov.

6.1 Lemma složenin

Pro lemma složenin využijeme nově zavedeného konceptu vícenásobného lemmatu (viz kap. 2). Lemma složenin je tedy vícenásobné a jeho prvky jsou lemmata jednotlivých složek složeniny. To je výhodné pro vyhledávání jednotlivých lemmat v korpusech. Vícenásobné lemma zajistí, že složenina bude ve výsledku vyhledávání podle lemmatu libovolné své složky. To neplatí pro konkrétní slovní tvary, např. dotaz na slovní tvar *jemu* nezahrne slovesnou složeninu *jemus*, což je ale správné, protože se skutečně jedná o dva různé slovní tvary. Dotaz na lemma *on* s hodnotou kategorie CAS=3 již ale oba slovní tvary najde, neboť lemma *on* je součástí vícenásobného lemmatu slovního tvaru *jemus*.

6.2 Relevantní morfologické kategorie složenin

Podívejme se, jaké morfologické kategorie jsou pro složeniny relevantní a jakých mohou nabývat hodnot.

Možnost, že složenina bude mít tolik morfologických značek, kolik má složek, nepřichází v úvahu, neboť složenina sama o sobě je samostatným slovním tvarem a jako taková musí být popsána morfologickou značkou jako celek. Zavedení vícenásobné morfologické značky (podobně, jako jsme zavedli vícenásobné lemma) by bylo navíc neekonomické, protože u většiny složenin nedochází ke konfliktu mezi hodnotami kategorií relevantních pro jednotlivé složky. Složenina tedy může být popsána stejně jako každé jiné slovo jedinou morfologickou značkou.

Označme K_i množinu kategorií, která je relevantní pro i-tou složku složeniny.

Množina K kategorií relevantních pro složeninu vznikne sjednocením množin kategorií relevantních pro její jednotlivé složky:

$K = K_1 \cup \ldots \cup K_n$.

Označme dále průnik těchto množin:

$G = K_1 \cap \ldots \cap K_n$.

Hodnoty těch kategorií, které neleží v průniku těchto množin (tedy kategorie z množiny $K - G$), se přenášejí i na výslednou složeninu.

Hodnoty kategorií z množiny G však musíme vyřešit pro jednotlivé typy složenin zvlášť. V případě, že se jejich hodnoty shodují u všech složek složeniny, měla by i výsledná hodnota být stejná. Jestliže se však hodnoty kategorií z množiny G pro jednotlivé složky liší, je třeba tento konflikt vyřešit tak, aby výsledkem byla jednoznačná hodnota, samozřejmě bez ztráty informací o jednotlivých složkách složeniny.

Příklad:
Složenina *zač* má 1. složku *za*, s množinou relevantních kategorií

$$K_1 = \{\text{POS, CAS}\},$$

a 2. složku *co*, s množinou relevantních kategorií

$$K_2 = \{\text{POS, SUB, FCE, CAS, GEN, NUM}\}.$$

Tabulka 6.1 ukazuje jejich hodnoty:

	za	*co*	*zač*
POS	R	P	S
SUB	-	Z	Z
FCE	-	T	T
CAS	4	4	4
GEN	-	N	N
NUM	-	S	S

Tabulka 6.1: Hodnoty relevantních kategorií složek složeniny *zač*.

Množinu K tvoří kategorie v levém sloupci tabulky, v množině G jsou všechny kategorie, které jsou relevantní pro obě složky složeniny, tedy ty, které mají vyplněnou hodnotu ve 2. a 3. sloupci tabulky 6.1:

$$G = \{\text{POS, CAS}\}.$$

Hodnota kategorie POS z množiny G má pro výsledný slovní tvar *zač* hodnotu složenina (viz dále), hodnota kategorie CAS není u obou složek v konfliktu, výsledná hodnota může být tudíž stejná, tedy 4. pád. Ostatní kategorie (z množiny $K - G$) přebírá složenina od své složky *co*. Výsledné hodnoty všech kategorií jsou v posledním sloupci tabulky 6.1.

Slovní druh složenin je vždy „složenina“, nezávisle na tom, jaké hodnoty kategorie `Slovní druh` mají jednotlivé složky. Kategorie `Slovní druh` je totiž až na slovesné složeniny typu V (viz dále) vždy konfliktní, neboť POS $\in G$.

Slovní druhy, které tvoří složky složeniny, je možné odvodit z typu složeniny. Také je možné snadno upravit dotazy využívající hodnot kategorie `Slovní druh` tak, aby zahrnuly i složky složenin, jak ukážeme v oddíle 6.4 o vyhledávání složenin v morfologicky anotovaných korpusech.

Množina G se liší podle typu složeniny, proto teď tyto typy probereme jednotlivě. Podíváme se také na problematiku vyhledávání složenin. Uživatel hledající v korpusu zřejmě bude chtít, aby se složeniny zahrnuly mezi odpovědi na obecné dotazy podle příslušných morfologických kategorií. Proto bude třeba modifikovat některé jednoduché dotazy, aby se mezi výsledky dostaly i příslušné složeniny.

Relevantní kategorie pro jednotlivé typy složenin ukazuje tabulka 6.5 na straně 67.

6.3 Typy složenin

Složeniny rozdělíme do několika základních skupin a v jejich rámci potom na typy. `Typ složeniny` (CMP) je flektivní morfologickou kategorií, která je relevantní pouze pro složeniny a zastupuje vlastně kategorii poddruh, která je u složenin vyhrazena k popisu jedné ze složek (viz dále).

6.3.1 Typy zájmenné ... n, c

Zájmenné složeniny jsou slovní tvary vzniklé spojením předložky s akuzativní rekcí a tázacího zájmena *co* nebo substantivního určitého (osobního) zájmena *on*. Podle toho rozlišujeme dva typy, které jsou tvořeny uzavřenou množinou tvarů. Můžeme je tedy zadat výčtem.

Typ c

Výčet všech složenin:

1. *zač* = *za co*, lemma: {*za, co*}
2. *nač* = *na co*, lemma: {*na, co*}
3. *oč* = *o co*, lemma: {*o, co*}
4. *več* = *v co*, lemma: {*v, co*}
5. *začpak* = *za copak*, lemma: {*za, copak*}
6. *načpak* = *na copak*, lemma: {*na, copak*}
7. *očpak* = *o copak*, lemma: {*o, copak*}
8. ?*večpak* = *v copak*, lemma: {*v, copak*}
9. *Xzač* = *za Xco*, lemma: {*za, Xco*}
10. *Xnač* = *na Xco*, lemma: {*na, Xco*}
11. *Xoč* = *o Xco*, lemma: {*o, Xco*}
12. ?*Xveč* = *v Xco*, lemma: {*v, Xco*}

Za X v posledních řádcích tabulky je možné dosadit jednu z neurčitých předpon, které se podílejí na tvorbě neurčitých zájmen, číslovek a příslovcí, totiž *kdoví, bůhví*, a další. Jejich seznam je uveden v obr. 4.1 na str. 30.

Mezi složeniny nepatří *proč*, protože jde už o lexikalizovanou spřežku, kterou nelze zaměnit za *pro co*. Tvary označené otazníkem jsou možné zřejmě jen hypoteticky.

Typ n

Výčet všech složenin:

1. *zaň = za něho*, lemma: {*za, on*}
2. *naň = na něho*, lemma: {*na, on*}
3. *oň = o něho*, lemma: {*o, on*}
4. *proň = pro něho*, lemma: {*pro, on*}
5. *doň = do něho*, lemma: {*do, on*}
6. *?veň = v něho*, lemma: {*v, on*}

Neobvyklý archaický tvar *veň* jsme skutečně nalezli v Ottově slovníku naučném:

... na krátko byv zbaven svého úřadu v Irsku a znova veň uveden jako tajemník... (66)

6.3.2 Typ zájmenně-slovesný ... t

Tento typ je zastoupen pouze jediným tvarem: *toť*. Jak již bylo zmíněno v oddíle o poddruzích zájmen (4.1.2.3), je toto slovo homonymní s ukazovacím zájmenem a s částicí (viz tabulka 6.2 a příklady (67) až (70)). I jako složenina je homonymní. Může totiž zastupovat dvě různé dvojice složek, lišící se v kategorii `Číslo`, pokaždé však se stejným vícenásobným lemmatem.

Slovní tvar	Složky složeniny	Lemma	Slovní druh
toť	*to je*	{*to, být*}	složenina
toť	*to jsou*	{*to, být*}	složenina
toť	—	*toť*	zájmeno
toť	—	*toť*	částice

Tabulka 6.2: Slovní tvar *toť* a jeho možné interpretace

Příklady:

Kniha, toť kouzelná brána k dobrodružství... (složenina v jednotném čísle) (67)

... co opravdu rád slyším, toť hlasy ptactva a noční vítr. (složenina v množném čísle) (68)

Inu, inu, toť, toť,... (částice) (69)

Toť se ví. (ukazovací zájmeno) (70)

Další příklady ((27) až (29)) byly uvedeny na str. 27.

6.3.3 Typ zkratkový ... Z

Složeninami jsou i zkratky víceslovných frází, např. *atd, atp, např.* Jejich lemmata však nerozebíráme na jednotlivé složky, neboť by to často vedlo ke složitým vícenásobným lemmatům. Lemma zkratkového typu složeniny je konkrétní slovní tvar.

Ze stejného důvodu neurčujeme ani ostatní morfologické kategorie.

Jediné relevantní morfologické kategorie složenin typu Z tedy jsou POS=S a CMP=Z.

6.3.4 Typy slovesné ... N, A, P, C, V, D, T, J, S

Slovesné složeniny může tvořit slovo téměř libovolného slovního druhu s přidaným zakončením *-s*, které zastupuje slovní tvar *jsi* (*převrátils, jehos*). Typ slovesné složeniny vyjadřujeme znakem, který kóduje příslušný slovní druh.

Slovesné složeniny se netvoří z předložek, citoslovcí, prefixových segmentů a zřejmě ani z cizích slov.

Slovesné složeniny nemůžeme zadat výčtem, jak jsme učinili u předchozích, zájmenných typů. Jde o otevřenou třídu. Přesto můžeme slovesné složeniny rozdělit do několika typů, podle slovního druhu jejich první složky.

Přes vysokou produktivnost tvoření slovesných složenin však jejich výskyt není příliš častý. Do morfologického slovníku je nezařazujeme, rozpoznávají se pomocí guessru. Jejich rozpoznání je velmi snadné — odtržení koncového *-s* u neznámých slov ponechá rozpoznatelný slovní tvar. Pokud jsou splněny podmínky (nečetné) uvedené dále, jde o slovesnou složeninu.

Než vyjmenujeme a popíšeme jednotlivé typy, poznamenejme, že slovesné složeniny bez ohledu na typ obvykle nevznikají ze slov, u nichž by přidané *-s* činilo potíže s výslovností: **vlass, *pařezs*.

Pro rozpoznávání slovesných složenin v průběhu morfologické analýzy může být důležité ještě jedno zjištění, a to pořadí slovesné složeniny v klauzi. Slovní tvar *jsi*, který je implicitně ve slovesné složenině přítomen, má většinou funkci pomocného slovesa, a jako takový je příklonkou. Slovesná složenina tedy většinou stojí ve větě na jejím začátku, aby bylo splněno Wackernaglovo pravidlo.

Toto pravidlo však není absolutně spolehlivé, jak ukazují příklady (71) až (76) z korpusu SYN:

To není tak samozřejmé, jak říkals. (71)
někam jsem za Tebou prostě šla, kde docela určitě a úplně volně stáls. (72)
Seňko, Seňuško, ty jedinej mně zůstals. (73)
zpívej, jako jsi zpíval, ještě než zešedivěls. (74)
doznals, co nečinils (75)
Blesku, český Blesku, nevím, kolik lidí již svou září osvítils. (76)

Podle slovního druhu první složky složeniny rozeznáváme typy slovesných složenin. Tyto typy označujeme podle kódu pro příslušný slovní druh.

Typ N

První složku slovesné složeniny typu N tvoří podstatné jméno.

Příklady, které uvádíme pod čísly (77) až (81), jsou vymyšlené, neboť se tento typ příliš často nevyskytuje, a vzhledem k tomu, že dosud nebyl v korpusech značkován, není snadné ho cíleně vyhledat. Pouze příklad (82) je z korpusu SYN, ale byl nalezený víceméně náhodou:

Bez oknas/okens nemohl vidět ven. (77)
Oknus/Oknůms přidělal okenice. (78)
To oknos/Ta oknas rozbil ty. (79)
O okněs/oknechs nemluvil. (80)

Oknems/Oknys viděl dobře. (81)
Z latinys měl reparát loni. (82)

Typ A

První složku slovesné složeniny typu A tvoří přídavné jméno.

Ani tento typ složenin se příliš často nevyskytuje, podařilo se nám však nalézt v korpusu SYN příklad s první složkou v 1. pádě (příklad (83)), z čehož zejména vyplývá, že v tomto případě nevystupuje implicitní *jsi* ve funkci pomocného slovesa:

Věrnýs jak kůň, jak býk všaks vášnivý. (83)

Příklad (84) ukazuje slovesnou složeninu typu A utvořenou ze jmenného tvaru přídavného jména:

Salome, podobnas úponku (z písně Karla Kryla) (84)

Typ P

První složku slovesné složeniny typu P tvoří zájmeno.

Příklady na slovesnou složeninu typu P již tak řídké nejsou. Např. tvarů *[tT]ys* se v korpusu SYN2000 vyskytuje 1068, převážná většina pochopitelně v beletristické části. Ve slovesných složeninách typu P se však vyskytují i jiná zájmena, jak ukazují příklady (85) až (87).

Copaks to musel řešit zrovna takhle? (85)
Tos řekl ty, já ne. (86)
Všechno, o čems mluvil... (87)

Typ C

První složku slovesné složeniny typu C tvoří číslovka.

Nejčastější případy jsou zřejmě číslovky tázací, ale je možné jsou i jiné číslovky, jak ukazuje příklad (90).

Koliks jich koupila? (88)
Kolikráts to viděl? (Internet) (89)
Pěts jich nemohl porazit. (vymyšleno) (90)

Typ V

První složku slovesné složeniny typu V tvoří sloveso.

Slovesné složeniny se slovesy vyžadují sloveso v příčestí minulém činném (VRB=L)[1], jednotném čísle (NUM=S) a ve 2. osobě (PER=2). Nelze tedy např. **kupujis, *kupuješs, *kupujs* ani *kupovalis*.

[1] Složeniny s pasivem jsou řazeny mezi složeniny typu A.

Implicitně přítomné sloveso *jsi* ve složenině je v tomto případě vždy pomocné, protože není možné, aby v jedné klauzi byla dvě finitní slovesa.

Příklady:
... má milá ženo, bylas tak statečná... (91)
Koupils ho Ireně. (92)

Typ D

První složku slovesné složeniny typu D tvoří příslovce.

Vytváření slovesných složenin typu D nemá zdá se žádná omezení, jak je vidět z příkladů (93) až (98):

Včeras měl narozeniny. (93)
Posledněs říkala, že... (94)
A určitěs to ztratila? (95)
A ještěs mě nikdy neodměnil. (96)
Nikdys nechtěla vařit. (97)
Jaks k tomu došla? (98)

Typ T

První složku slovesné složeniny typu T tvoří zvratná částice.

Tento typ slovesných složenin lze zadat výčtem. Jsou to tyto slovní tvary: *sis, ses*.

Mezi složeniny typu T by se mohly počítat i tvary *bych, bys, bychom, bysme, byste*, protože i u nich jde o spojení částice (*by*) a tvaru slovesa *být*. My je však řadíme mezi částice s jednoduchým lemmatem *by*.

Typ J

První složku slovesné složeniny typu J tvoří spojka.

Některé spojky slovesné složeniny patrně netvoří. Jsou to zejména **as, *is, ?ales, ?čis*. Z většiny ostatních spojek slovesné složeniny utvořit lze, např. *nebos, protos, neboťs, protožes, zdas*. Jak je vidět, nezáleží to na jejich souřadnosti nebo podřadnosti.

Příklady (99) až (101) pocházejí z korpusu SYN:

... když̌s teda říkal,... (99)
... neměls už čas se zase stejnou cestou vrátit, nebos na to zapomněl. (100)
Nevím, jestlis ho vůbec znal. (101)

Mezi složeniny typu J nepočítáme spojky *abys, kdybys*, ani jejich ostatní tvary *abych, kdybych, abychom, kdybychom, abyste, kdybyste, abysme, kdybysme*, i když bychom je za složeniny považovat mohli. Podobně jako kondicionálová částice *by*, i v tomto případě volíme tradiční řešení a všechny uvedené tvary řadíme mezi spojky, s lemmaty *aby, kdyby*.

Typ S

První složku slovesné složeniny typu S tvoří složenina. Vícenásobné lemma těchto složenin má tři prvky: dvouprvkové vícenásobné lemma první složeniny a lemma *být* slovního tvaru *jsi*.

Jde o tvary vytvořené ze zájmenných složenin, tedy např. *oňs, začs*. Že nejde o pouhý teoretický případ, dokazuje úryvek z textu písně Hany Zagorové:

... o čem snil jsi ty, načs přísahal... (102)

Běžně se však tento typ složenin opravdu nevyskytuje.

6.4 Vyhledávání složenin v korpusech

V tomto oddíle se zamyslíme nad způsobem, jak zahrnout složky různých typů složenin do vyhledávacích dotazů.

Otázka souvisí s množinou relevantních kategorií pro jednotlivé typy složenin, viz 6.5 na str. 67.

Typy zájmenné a zájmenně-slovesné

U zájmenných typů jde o předložku s akuzativní rekcí a zájmeno v akuzativu. V množině G je kromě kategorie `Slovní druh` pouze kategorie `Pád`, který je ovšem pro obě složky shodný — akuzativ. Předložka další relevantní kategorie nemá, pro výslednou složeninu jsou tedy relevantní všechny kategorie, které jsou relevantní i pro příslušné zájmeno. Jsou to `Rod` (střední, mužský životný nebo neživotný pro typ n a střední pro typ c), `Číslo` (jednotné), `Osoba` (3), `Poddruh` (substantivní), `Funkce` (určitá pro typ n, tázací nebo vztažná pro typ c).

Složeniny tohoto typu by se měly, pokud si to uživatel přeje, zahrnout do výsledku na dotaz požadující všechny výskyty předložek. K tomu je ovšem třeba dosud jednoduchý dotaz POS=R modifikovat, a to tak, aby se nalezly nejen předložky (POS=R), ale i složeniny (POS=S) typu n a c (CMP=[nc]). Výsledný dotaz lze zapsat např. takto (před šipkou je jednoduchý dotaz, za šipkou dotaz modifikovaný):

$$\mathrm{POS} = \mathrm{R} \rightarrow (\mathrm{POS} = \mathrm{R}) \vee (\mathrm{POS} = \mathrm{S} \wedge \mathrm{CMP} = [\mathrm{nc}])$$

Podobně musíme modifikovat i dotaz na zájmena. Sem je třeba navíc zahrnout zájmenně-slovesnou složeninu typu t a slovesnou složeninu typu P (CMP=[nctP]):

$$\mathrm{POS} = \mathrm{P} \rightarrow (\mathrm{POS} = \mathrm{P}) \vee (\mathrm{POS} = \mathrm{S} \wedge \mathrm{CMP} = [\mathrm{nctP}])$$

Dotazy na ostatní morfologické kategorie mohou zůstat beze změny, neboť jejich hodnoty se stávají hodnotami příslušných kategorií zájmenné složeniny.

Složenina *toť* má 1. složku *to*, s množinou relevantních kategorií

$$K_1 = \{\mathrm{POS}, \mathrm{SUB}, \mathrm{GEN}, \mathrm{FCE}, \mathrm{CAS}, \mathrm{NUM}\}.$$

a 2. složku *je/jsou*, s množinou relevantních kategorií

$$K_2 = \{\text{POS, SUB, PER, NUM, VRB, NEG}\},$$

Tabulka 6.3 ukazuje jejich hodnoty.

	to	*je/jsou*	*toť*
POS	P	V	S
SUB	D	0	D
FCE	U	—	U
CAS	1	—	1
GEN	N	—	N
NUM	S/P	S/P	S/P
PER	—	3	3
VRB	—	P	P
NEG	—	A	A

Tabulka 6.3: Hodnoty relevantních kategorií složek složeniny *toť*.

Tabulka 6.4 uvádí přehled hodnot relevantních kategorií pro zájmenné složeniny.

Složenina	SUB	FCE	GEN	NUM	CAS	FMU	Příklad
typ c	Z	T/V	N	S	4	—	*oč*
typ n	Z	U	M/I	S	4	n	*oň*

Tabulka 6.4: Zájmenné složeniny a jejich relevantní kategorie

Typy slovesné

Přidaná koncovka *-s*, zastupující slovní tvar *jsi* slovesa *být*, nese tyto hodnoty relevantních morfologických kategorií:

PER = 2 SUB = b/0
NUM = S NEG = A
VRB = P

Všechny tyto kategorie mohou (ale nemusí) ležet v množině G.

V případě konfliktu hodnot kategorií PER, SUB, NUM, NEG bude hodnota výsledné složeniny rovna hodnotě náležející první složce složeniny, tedy např. pro tvar *židlemis* bude NUM=P, pro slovo *nevysokýs* bude NEG=N, pro slovo *měs* bude PER=1 a pro slovo *tomus* bude SUB=D.

Kategorie VRB leží v množině G pouze pro složeniny typu V a dále pak pro ty složeniny typu A, jejichž první složka je trpným rodem slovesa (např. *ukrytas, podobnas*), i když ta se vyskytují opravdu zřídka. V obou případech je hodnota této kategorie v konfliktu s hodnotou první složky, která je v prvním případě VRB=L, ve druhém VRB=T. Tyto hodnoty převádíme na hodnoty kategorie `Slovesný tvar` složeniny. Z typu složeniny je možné odvodit i hodnotu kategorie `Slovesný tvar` druhé složky (*jsi*) a vytvořit podle toho dotaz.

U ostatních slovesných složenin je VRB=UNDEF, i když tam není tato kategorie v konfliktu s kategoriemi první složky (VRB$\notin G$). Kdybychom položili

VRB=P podle hodnoty slovního tvaru *jsi*, nebylo by to konzistentní s předchozím rozhodnutím o hodnotě této kategorie u složenin typu V a A v pasivu. Slovesné složeniny zahrneme v případě potřeby do dotazu na všechna slovesa v prézentu jednoduše vyloučením neslovesných složenin (CMP ≠ [ncZ]):

$$\text{VRB} = \text{P} \rightarrow (\text{VRB} = \text{P}) \vee (\text{POS} = \text{S} \wedge \text{CMP} \neq [\text{ncZ}])$$

Podobně řešíme i kategorii `Osoba`, která je sice konfliktní pouze pro složeniny typu CMP=V a CMP=P, ale v zájmu konzistence ji u ostatních typů nepovažujeme za relevantní (PER=UNDEF). Následující dotaz vyhledá všechna slova s hodnotou kategorie PER=2 a slovesné složeniny kromě zájmenných (typ n a c), zájmenně slovesných (typ t) a zkratkových (typ Z).

$$\text{PER} = 2 \rightarrow (\text{PER} = 2) \vee (\text{POS} = \text{S} \wedge \text{CMP} \neq [\text{ncZt}])$$

Ve všech slovesných složeninách kromě typu t je vždy druhá složka v jednotném čísle. Z dotazu na jednotné číslo tak musíme vyloučit složeniny typu t, kde může být druhá složka v čísle množném, jednotné číslo je však vždy pokryto přímo hodnotou kategorie NUM=S:

$$\text{NUM} = \text{S} \rightarrow (\text{NUM} = \text{S}) \vee (\text{POS} = \text{S} \wedge \text{CMP} \neq [\text{Zt}])$$

Podobný dotaz lze vytvořit i pro negaci. Slovní tvar *jsi* je ve slovesných složeninách vždy s hodnotou NEG=A, dokonce i když je první složka negativní (*nebyls*). Je tedy otázkou, zda vůbec má dotaz na NEG=A v takovém případě smysl. Pro úplnost ho ale uvádíme:

$$\text{NEG} = \text{A} \rightarrow (\text{NEG} = \text{A}) \vee (\text{POS} = \text{S} \wedge \text{CMP} \neq [\text{ncZ}])$$

Dotaz na poddruh pomocné sloveso je komplikovanější. Ve většině případů je *jsi* ve slovesné složenině pomocné, ale není tomu tak vždy. Výjimkou mohou (ale nemusí) být složeniny se jménem v 1. pádě, viz příklady (103), (104) a také (83) a (84).

Drahoušku, ale tys moje žena! (103)
tys nejen blázen, ale ke všemu ještě pitomec! (104)

Ve větě *Tys blázen* má implicitní *jsi* SUB=0, zatímco ve větě *Tys byl blázen* má SUB=b. Hodnota celé složeniny *tys* je SUB=Z (substantivní zájmeno). V dotazu je tedy třeba vyloučit slovní tvary v 1. pádě. V případě, že v takové složenině je implicitní *jsi* ve funkci slovesa pomocného, je tato alternativa pokryta první možností v disjunkci nového dotazu. Vyhledají se jen ty případy, kdy jde skutečně o pomocné sloveso[2].

Kromě toho je třeba v dotazu vyloučit složeniny neslovesné, tj. zájmenné, zájmenně-slovesné a zkratkové.

Výsledný dotaz tedy bude vypadat takto:

$$\text{SUB} = \text{b} \rightarrow (\text{SUB} = \text{b}) \vee (\text{POS} = \text{S} \wedge \text{CMP} \neq [\text{ncZt}] \wedge \text{CAS} \neq 1)$$

[2]Za předpokladu, že je kategorie `Poddruh` ve složenině správně určena.

Ve složenině *toť* typu t se sice vyskytuje tvar slovesa *být*, ale zde nikdy nevystupuje v roli slovesa pomocného.

Je třeba ještě doplnit dotazy na slovní druhy jednotlivých složek složeniny:

$$\mathrm{POS} = \mathrm{x} \rightarrow (\mathrm{POS} = \mathrm{x}) \vee (\mathrm{POS} = \mathrm{S} \wedge \mathrm{CMP} = \mathrm{x}),$$

kde x zastupuje [NADCJT], tedy jeden ze slovních druhů, které mohou být první složkou slovesné složeniny. Dotaz je přímočarý — hledáme slovní druh x a složeniny typu x, což jsou právě ty, jejichž první složkou je tvar s hodnotou POS=x.

Vynechali jsme slovní druh sloveso a zájmeno. Dotaz na zájmeno musí zahrnout i složeniny zájmenné a zájmenně-slovesné:

$$\mathrm{POS} = \mathrm{P} \rightarrow (\mathrm{POS} = \mathrm{P}) \vee (\mathrm{POS} = \mathrm{S} \wedge \mathrm{CMP} = [\mathrm{PncS}])$$

Sloveso je přítomno ve všech slovesných složeninách a ve složenině zájmenně-slovesné, proto bude v dotazu jednodušší vyloučit složeniny ostatní (neslovesné):

$$\mathrm{POS} = \mathrm{V} \rightarrow (\mathrm{POS} = \mathrm{V}) \vee (\mathrm{POS} = \mathrm{S} \wedge \mathrm{CMP} \neq [\mathrm{ncZ}])$$

Na první pohled mohou modifikované dotazy vypadat složitě. Většina korpusových vyhledávačů (manažerů) však umožňuje, aby si uživatel definoval a pojmenoval určité dotazy, aby je mohl neustále využívat bez složitého vytváření. Toto jsou případy, kdy by se takové definice hodily.

Kromě toho je třeba poznamenat, že dosud takové vyhledávky nebyly možné vůbec. Uživatel, který nebude chtít složeniny do svých dotazů zahrnout, může i nadále používat dotazy, na které je zvyklý.

1 POS	2 SUB	3 FCE	4 ASP	5 ABR	6 GEN	7 NUM	8 CAS	9 DUA	10 PER	11 DEG	12 NEG	13 VRB	14 NOM	15 CMP	16 INT
S	Z	U			MI	S	4		3					n	
	Z	TV			N	S	4							c	
	D	U				SP	1							t	
				⊕										Z	
	⊕				⊕	⊕	⊕	⊕			⊕			N	
	⊕				⊕	⊕	⊕	⊕		⊕	⊕	(T)	⊕	A	
	⊕	⊕			⊕	⊕	⊕	⊕	⊕					P	
	⊕	⊕			⊕	⊕	⊕	⊕						C	
	⊕					S			2		⊕	L		V	⊕
	⊕	⊕				S				⊕	⊕			D	
	⊕					S								T	
	⊕					S								J	
	Z	UTV			MIN	S			3					S	

Tabulka 6.5: Přehled relevantních kategorií pro jednotlivé typy složeniny. Relevantní kategorie jsou pro jednotlivé typy vyplněny buď kódem (případně více možnými kódy) svých hodnot, nebo znakem ⊕, který zastupuje libovolnou hodnotu z množiny hodnot své kategorie.

7 Morfologický slovník

7.1 Vztah morfologického slovníku a morfologických nástrojů

Morfologický slovník v ideálním případě obsahuje „všechny“ slovní tvary českého jazyka spolu s jejich morfologickým popisem. Slůvko všechny je v uvozovkách z pochopitelných důvodů — nikdy totiž nebudeme schopni vytvořit takový slovník, který by obsahoval úplnou slovní zásobu nějakého živého jazyka.

Existující morfologické slovníky češtiny, o kterých jsme se již zmiňovali, totiž pražský a brněnský, jsou však dostatečně velké na to, aby se pomocí nich mohly rozpoznat běžně užívané slovní tvary jak standardní, tak i obecné češtiny. Oba slovníky jsou neustále obohacovány o nové slovní tvary, které tam dosud nebyly ať už proto, že se na ně „zapomnělo“, nebo proto, že nově vznikly (neologismy).

Přidávání nových a nových slovních tvarů do slovníku však nemusí být vždy jen pozitivní, jak by se na první pohled mohlo zdát. Nebezpečné je především přidávání cizích vlastních jmen, která mohou být silně homonymní. Jedno slovo tak může znamenat např. cizí příjmení, ale současně i název firmy. V každém z těchto případů má jiný rod, a pokud přebírá českou flexi, skloňuje se pokaždé podle jiného vzoru. Od českého slovníku samozřejmě nelze očekávat, že bude obsahovat všechna cizí slova. V okamžiku, kdy se v textu vyskytne takové slovo v jiném významu, než bylo již zachyceno ve slovníku, dochází k chybě. Podle slovníku se totiž dané slovo rozpozná, ale špatně. V takových případech by paradoxně bylo lepší, kdyby se nerozpoznalo. Jiné nástroje, využívající pravidel slovotvorby, by si s jeho rozpoznáním poradily lépe. Pro další zpracování, zejména desambiguaci, je totiž většinou výhodnější hodnoty morfologických kategorií raději podspecifikovat, tedy přiřadit jim sdruženou hodnotu, než je určit špatně.

Při návrhu struktury a obsahu morfologického slovníku je třeba mít na paměti, jakým způsobem se slovník bude využívat. Jistě by bylo hezké, kdyby morfologický slovník obsahoval veškeré morfologické informace o všech slovních tvarech daného jazyka, a to nezávisle na programových nástrojích, které ho využívají.

Na druhou stranu však je výhodné některé jevy nepopisovat pomocí slovníku, ale nechat to na nástrojích morfologické analýzy nebo syntézy. Jde zejména o tvary negace. Až na několik výjimek u slovesných infinitivů se tvoří zcela pravidelně pomocí předpony *ne-*. Toto pravidlo je velmi jednoduché, téměř univerzální, a zahrnuje navíc obrovské množství slovních tvarů.

Zdá se tedy rozumné nevkládat pravidelně negované slovní tvary do slovníku, ale nechat jejich zpracování na nástrojích. Slovníkové heslo by mělo pouze obsahovat informaci o tom, zda lze z daného slovního tvaru utvořit negaci, či nikoli, ale ne samotný negovaný tvar. Stejně lze zacházet se stupňováním pří-

davných jmen a příslovcí. Negaci, stejně jako stupňování, musí umět rozpoznat analýza a vytvořit syntéza.

Tato symetrie morfologické analýzy a syntézy však není nutná. Např. cizí slova, o kterých byla řeč výše, stačí umět jen rozpoznat, a na jejich rozpoznání není morfologický slovník třeba. V naprosté většině případů jde totiž o podstatná jména. Jako podstatná jména lze analyzovat i součásti víceslovných cizojazyčných celků, např. názvů písniček, měst, různé slogany, a to i tehdy, když ve svém původním jazyce patří k jiným slovním druhům. Podle našeho názoru tedy z cizích slov do morfologického slovníku patří jen ta nejběžnější. Je však třeba umět v textu rozpoznat i ta ostatní a přiřadit jim správné hodnoty morfologických kategorií, třeba podspecifikované, tzn. se sdruženou hodnotou některých kategorií. K rozpoznání neznámých slov se používá guesser.

7.1.1 Guesser

Morfologická analýza nerozpozná úplně všechna slova v neznámém textu. Z experimentů s Českým národním korpusem vyplývá, že 2 až 3 procenta slov zůstávají nerozpoznána (viz (Hlaváčová, 2001)). Mnoho slov, která nejsou obsažena ve slovníku, však lze rozpoznat pomocí tzv. guessru. Guesser využívá ortografických a morfologických pravidel k analýze neznámých slov. Jeho výsledkem je odhad relevantních kategorií neznámého slova a jejich hodnot a také rekonstrukce pravděpodobného lemmatu.

Guesser tedy přiřazuje hodnoty morfologických kategorií neznámým slovním tvarům. V ideálním případě přiřadí jedno (správné) lemma a jednu (správnou) morfologickou značku. Většinou je však možností více. Guesser by však měl být navržen tak, aby jich nebylo příliš, protože to potom ztěžuje následná zpracování, především desambiguaci. Pochopitelně by však mezi výsledky měla být i správná dvojice lemmatu a morfologické značky.

V českém jazyce lze využít k sestavení guessru jednak slovních zakončení, tedy přípon a koncovek, a jednak předpon.

7.1.1.1 Prefixový guesser

Využití předpon pro práci guessru je přímočaré. Mějme nerozpoznané slovo z, které lze rozdělit na dvě části: $z = p \cdot w$, kde p je známá předpona a w slovní tvar rozpoznatelný na základě morfologického slovníku. Je-li $\lambda(w) = \bar{w}$, potom $\lambda(z) = p \cdot \bar{w}$. Navíc obě slova, w i z, mají stejné relevantní morfologické kategorie se stejnými hodnotami, tedy:
Je-li $\mu(w) = \langle \bar{w}, m \rangle$, potom $\mu(z) = \langle p \cdot \bar{w}, m \rangle$, kde m je morfologická značka.

Vyjádřeno slovy — lemma neznámého slovního tvaru, který se skládá ze známé předpony a známého slovního tvaru, lze zrekonstruovat přidáním předpony k lemmatu známého slovního tvaru. Hodnoty morfologických kategorií jsou stejné (zde mohou nastat výjimky ve změně vidu u sloves).

Příklad:
Předpokládejme, že slovní tvar *z=eurooken* není obsažen v morfologickém slovníku. Můžeme ho však rozdělit na předponu *p=euro-* a slovní tvar *w=oken*, který ve slovníku je a my ho umíme rozpoznat. Budeme-li mít *euro* v seznamu

předpon, můžeme původní tvar *eurooken* lemmatizovat jako *eurookno* a přiřadit mu tyto hodnoty: POS=N, GEN=N, CAS=2, NUM=P, které jsou stejné jako hodnoty známého slovního tvaru *oken*.

Seznam předpon jsme vytvořili pomocí statistických metod ze slovních tvarů korpusu SYN2000. Popis metod i jejich výsledky lze nalézt v příspěvcích (Hlaváčová – Hrušecký, 2008) a (Urrea – Hlaváčová, 2005).

U každé předpony je možno uvést, s jakým slovním druhem se může kombinovat. Např. uvedená předpona *euro-* se nemůže připojit ke slovesu, ale jen k podstatnému, případně ještě přídavnému jménu. Většinu předpon, které se pojí se jmény, by možná někteří lingvisté nepovažovali za klasické předpony, ale spíše za složky slova vzniklého skládáním. Pro naše účely je důležité, že lze pomocí jejich odtržení velmi spolehlivě analyzovat neznámá slova.

Seznam předpon získaný automaticky pomocí nástrojů zmíněných výše jsme použili při nové implementaci morfologické analýzy, viz (Hlaváčová – Kolovratník, 2008), která pracuje s dosavadním pražským morfologickým slovníkem.

7.1.1.2 Postfixový guesser

Postfixový guesser má k dispozici seznam zakončení slovních tvarů s množinou možných morfologických značek a pravidel na vytvoření k nim příslušejících lemmat. Jestliže neznámé slovo končí jedním nebo více z těchto zakončení, guesser mu přiřadí příslušnou množinu morfologických značek a podle pravidel odvodí pravděpodobné tvary lemmat.

Některá slovní zakončení určují své morfologické kategorie jednoznačně, jiná nabízejí možností mnoho. Záleží zejména na délce zakončení, podle kterého se snažíme odhad provést. Při implementaci guessru je tedy třeba se rozhodnout, s jak dlouhými zakončeními chceme pracovat. Čím delší zakončení, tím méně možností guesser nabídne, ale tím více jich je zapotřebí. Krátkých zakončení není třeba tolik, jsou však většinou neúnosně mnohoznačná. Je třeba zvolit vhodný kompromis. Guesser použitý pro korpus SYN2000 (viz (Hlaváčová, 2001)) pracoval s postfixy délky 4, která vycházela z výsledků ještě starší práce na projektu MOZAIKA (viz (Kirschner, 1983))[1].

[1] Uvedený postfixový guesser by bylo možno rozšířit následujícím způsobem i na běžnější slovní tvary, ale v tom případě by bylo třeba zajistit i opačný postup jejich možného generování, tedy by to vlastně přestal být guesser.

Existuje např. řada známých zakončení, která nemusí splňovat požadavky na automatické vyhledávání postfixů, konkrétně na jejich délku. Je vhodné seznam známých postfixů přidat k seznamu vytvořenému automaticky. Máme teď na mysli zejména slova s číselnými předponami, které se sice rozpoznají pomocí prefixového guessru (viz 7.1.1.1), ale po jejich odtržení nevznikne analyzovatelný slovní tvar, např. *letý, nohý, hlavý, dílný, místný, ramenný* po odtržení předpon *dvou-, čtyř-, sedmi-*,... Není rozumné ani proveditelné zahrnout všechny možné kombinace těchto postfixů i prefixů do slovníku, zvláště u prefixů číslovkových. Seznam takových postfixů však tak rozsáhlý není a mohl by posloužit podobně jako už dnes slouží guesser prefixový. Navíc tyto ruční postfixy většinou „zaberou“ stoprocentně, tedy určují morfologické kategorie i lemma daného slovního tvaru jednoznačně.

Dále do tohoto seznamu můžeme přiřadit postfixy *-koli, -si, -pak* a další, pomocí kterých vznikají zájmena neurčitá odvozená od zájmen tázacích (*kdosi, jakýsi, čísi, kdokoli, jakýkoli, číkoli, kdopak, jakýpak, čípak,...*) a další. Jejich výčet je uveden v obrázku 4.1 a v tabulce 4.1.

Jednotlivé slovní tvary z paradigmatu těchto zájmen jsou zatím vyjmenovány jednotlivě v pražském morfologickém slovníku, ale jejich zahrnutí do takového postfixového guessru by bylo přehlednější a elegantnější. Zahrneme-li navíc požadavek na slovní druh, případně i poddruh řetězce vzniklého po odtržení postfixu, byly by odhady guessru i jednoznačné.

Např. postfix *-si* je příliš krátký na to, abychom mohli spoléhat na jednoznačnost morfolo-

Např. zakončení *-kyní* patří téměř na 100 % podstatnému jménu rodu ženského v 7sg nebo 2pl a s lemmatem *-kyně*. S touto znalostí můžeme úspěšně analyzovat slovní tvary, které neobsahuje slovník, viz příklady (105) až (108) z korpusu SYN. Příklad (108) obsahuje slovní tvar s překlepem, přesto je možno ho v textu správně rozpoznat.

byla ve výuce dospělých ostřílenou borkyní (105)

Falstaff zatroubil na ústup před tou čarodějnicí, před tou ďáblovou spřeženkyní (106)

s jedinou nadšenkyní, která se k nim přidala (107)

v rozhovoru s nejvyšší státní zastupkyní (108)

Zkrátíme-li postfix na *-yní*, stále ještě můžeme s velkou pravděpodobností hádat na stejný typ podstatného jména, ovšem vyloučíme-li velmi časté slovo *nyní*.

Zkrácení postfixu jen na *-ní* už nám příliš nepomůže, neboť slov s tímto zakončením je mnoho — nejčastější jsou podstatná jména rodu středního (vzor *stavení*) s velkou homonymií ve většině kombinací čísla a pádu, přídavná jména měkká (*jarní*) s homonymií ještě větší, přídavná jména tvrdá v 1pl rodu mužského životného (*úspěšní muži*), slovesa 4. třídy (*špiní*).

7.2 Struktura slovníku

Morfologický slovník by měl v ideálním případě obsahovat právě všechny slovní tvary daného jazyka, opatřené potřebnou morfologickou informací, tedy lemmatem a hodnotami všech relevantních morfologických kategorií vyjádřených pomocí morfologické značky a mutace.

Celý slovník si můžeme velmi zjednodušeně pro začátek představit jako tabulku s právě vyjmenovanými údaji, viz příklad v tabulce 7.1.

Slovní tvar	Lemma	Morf. značka	Mutace
okny	{*okno, vokno*}	`N-----NNP7-----A----`	-
oknama	{*okno, vokno*}	`N-----NNP7-----A----`	Fa
voknama	{*okno, vokno*}	`N-----NNP7-----A----`	FaGv

Tabulka 7.1: Příklad několika položek jednoduchého modelu morfologického slovníku

Značení hodnot kategorie `Mutace` je popsáno v oddíle 3.4.

Právě naznačený jednoduchý model slovníku jako tabulky však nezachycuje vztahy mezi jednotlivými lemmaty, které bychom rádi také ve slovníku měli. Jde nám především o vztahy derivační, které se ukazují být klíčovými pro praktické využití morfologického slovníku při automatických překladech. Jednotlivá paradigmata proto sdružujeme do záznamů.

Základní jednotkou morfologického slovníku je tedy záznam reprezentovaný lemmatem. Lemma může být vícenásobné. Každý záznam obsahuje informace

gických vlastností slovních tvarů s tímto zakončením. Přidáme-li však požadavek, že řetězec, vzniklý po odtržení tohoto postfixu, je nejen rozpoznatelný morfologickým analyzátorem, ale je to navíc zájmeno, je jeho analýza již jednoznačná — v tom smyslu, že má všechny kategorie až na FCE=N stejné jako základní zájmeno.

o jednom paradigmatu. Toto paradigma je tímto záznamem zcela a jednoznačně morfologicky popsáno.

Vztahy mezi lemmaty jsou vyjádřeny pomocí odkazů. My zde popíšeme odkazy vyjadřující derivační vztahy, ale obecně je možno definovat i jiné typy odkazů, např. synonymické, hypo- a hyperonymické, odkazy mezi vidovými dvojicemi. Teoreticky si může uživatel slovníku přidat libovolné typy odkazů, podle aplikace, ke které slovník využívá.

Reprezentantem paradigmatu je lemma, které slouží jako slovníkové heslo. Kdyby neexistovala v češtině homonymní lemmata, mohli bychom každé lemma považovat za klíč, podle kterého se ve slovníku vyhledává. Kvůli nejednoznačnosti to bohužel nejde. Homonymní lemmata rozlišujeme přidáním přirozených čísel. Klíčem záznamu v morfologickém slovníku je tedy buď lemma, jestliže k němu neexistuje homonymní lemma, nebo lemma s číselným sufixem, který rozlišuje homonymní lemmata (např. *žít-1 — žiji, žít-2 — žnu*).

Problém polysémie (víceznačnosti) v morfologickém slovníku neřešíme. Máme-li tedy dvě lemmata s odlišným významem, ale shodnými paradigmaty, budou zastoupena jediným záznamem (např. *kolej*).[2]

Rozšířené paradigma globálních mutací je reprezentováno vícenásobným lemmatem, které jsme definovali jako množinu. Množina nemůže být klíčem záznamu, proto ji vyjadřujeme jako jeden řetězec pomocí regulárních výrazů. Vícenásobné lemma {*okno, vokno*} z předchozího příkladu se pomocí regulárního výrazu zapíše jako *v?okno*.

Slovníkový záznam morfologického slovníku tedy obsahuje celé paradigma jednoho lemmatu, společně s morfologickými značkami jednotlivých slovních tvarů a označením typu jejich mutací. Tabulka 7.2 ukazuje příklad jednoho slovníkového záznamu.

Lemma	Slovní tvar	Morf. značka	Mutace
město	*město*	N-----NNS1-----A----	
		N-----NNS4-----A----	
		N-----NNS5-----A----	
	města	N-----NNS2-----A----	
		N-----NNP1-----A----	
		N-----NNP4-----A----	
		N-----NNP5-----A----	
	městu	N-----NNS3-----A----	
		N-----NNS6-----A----	Fu
	městě	N-----NNS6-----A----	Fe
	městem	N-----NNS7-----A----	
	měst	N-----NNP2-----A----	
	městům	N-----NNP3-----A----	
	městech	N-----NNP6-----A----	
	městy	N-----NNP7-----A----	
	městama	N-----NNP7-----A----	Fa

Tabulka 7.2: Příklad jednoho záznamu morfologického slovníku

Vzhledem k velké pravidelnosti českého gramatického systému je výhodné

[2] Morfologický slovník by se mohl v budoucnu rozšířit o další elementy popisu, včetně popisu sémantického.

do slovníku vkládat nikoli jednotlivé slovní tvary, ale využít pravidla, podle kterých se slovní tvary dají vytvořit, tedy vzory. Vzory výrazně zmenší velikost slovníku. Tabulka 7.3 ukazuje příklad z tabulky 7.2 zapsaný pomocí vzoru. Veškeré tvary, značky i kódy mutací jsou součástí vzoru.

Lemma	Kofix	Vzor
město	*měst*	mt

Tabulka 7.3: Příklad jednoho záznamu morfologického slovníku zapsaného pomocí vzoru

O vzorech obecně pojednáváme v kapitole 8, kapitoly 9 až 13 se věnují vzorům jednotlivých slovních druhů.

Kromě popisu paradigmatu obsahuje slovníkový záznam ještě derivační odkazy na jiné slovníkové záznamy, které popisují paradigmata příbuzných lemmat. Odkazy si můžeme představit jako šipky mezi jednotlivými záznamy. Zásadně používáme šipky oboustranné. To znamená, že jestliže vede odkaz od záznamu A k záznamu B, vede odkaz i od záznamu B k záznamu A. Z tohoto důvodu není označení derivační odkaz zcela přesné, protože nepopisuje vždy skutečné odvození jednoho lemmatu z druhého. Kvůli nejrůznějším aplikacím je však výhodné mít odkazy obousměrné. Vyhýbáme se tím občas i obtížnému rozhodování, které lemma je původní a které odvozené. Z hlediska aplikací je tato otázka podružná. Lemmata záznamů propojených derivačními odkazy ale raději nazýváme příbuzná lemmata.

Bylo třeba rozhodnout, jak derivační odkazy propojovat v případech, kdy je jich více. Zvolili jsme „hvězdicové uspořádání“. Znamená to, že jeden záznam je zvolen za základní a od něj vedou derivační odkazy k lemmatům příbuzným. Příklady jsou uvedeny na obrázcích 7.1 a 7.2. Na obrázcích jsou zachyceny jen ty derivace, které lze odvodit ze vzoru — pro zobrazené záznamy ze slovesných vzorů (viz kap. 12). Ostatní příbuzná slova, která do „hvězdy“ také patří (k obr. 7.1 by to bylo např. lemma *skok*), je třeba zadat do slovníku „ručně“. Rovněž ručně se musí propojit vidové dvojice, např. lemmata *skočit* a *skákat* z uvedených obrázků. Samozřejmě lze pomocí odkazů propojit i slova odvozená pomocí derivací prefixových.

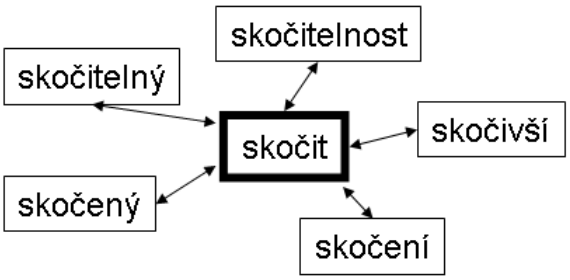
V zásadě nezáleží na tom, které lemma zvolíme za základní, důležitá je existence derivačních odkazů, které je propojují. Přesto jsme se pokusili o systematické řešení:

Základním lemmatem (středem hvězdy) je v případě slovesných derivací sloveso. V případě, že je mezi příbuznými lemmaty více sloves, je základním sloveso nejkratší. Tímto požadavkem upozaďujeme slovesa iterativní.

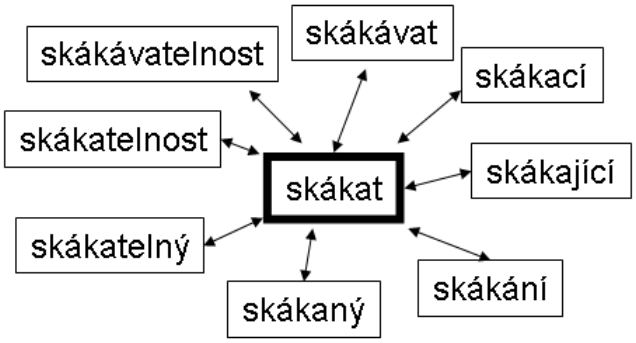
Není-li mezi příbuznými lemmaty sloveso, je základním lemmatem podstatné jméno. Není-li tam ani podstatné jméno, je základním lemmatem přídavné jméno.

Naskýtá se otázka, co všechno považovat za příbuzná lemmata. V prvním plánu jsme mezi příbuzná lemmata zahrnuli jen ta lemmata, která lze odvodit pravidelně pomocí přípon. Uvědomujeme si, že toto řešení je alibistické a má mnohá úskalí. Při implementaci konkrétního slovníku však nic nebrání propojení více lemmat. V rámci navrženého formátu je možné dokonce vytvářet i jiné typy odkazů. Důležitým typem odkazů bude např. propojení vidových

dvojic sloves.



Obrázek 7.1: Ukázka derivačních odkazů odvozených pravidelně ze základního slovesa *skočit*



Obrázek 7.2: Ukázka derivačních odkazů odvozených pravidelně ze základního slovesa *skákat*

Jelikož je většina derivačních vztahů velmi pravidelná, přidáváme pravidla pro vytvoření odvozených lemmat přímo do jednotlivých záznamů jako součást vzoru. Slovníku s tímto typem informace o derivacích budeme říkat kompaktní slovník. Pro uchování kompaktního slovníku použijeme formátu PML, který byl vytvořen jako jednotný datový formát pro ukládání lingvistických dat.[3]

Pro automatické použití kompaktního slovníku je třeba ho nejprve „rozgenerovat“ podle derivačních pravidel, čímž se vytvoří záznamy pro odvozená lemmata a oboustranné odkazy mezi nimi. Tyto odkazy se zakódují tak, aby je bylo možno využít v aplikacích, ke kterým se slovník používá. Teprve v další fázi je možno vytvářet jednotlivé slovní tvary.

[3] http://ufal.mff.cuni.cz/jazz/PML/doc/pml_doc.html

8 Vzory

V první části práce jsme se zabývali vymezením kategorií, které charakterizují české slovní tvary, a jejich hodnotám. Toto vymezení je postačující pro vytvoření morfologického slovníku, který obsahuje všechny slovní tvary spolu s jejich popisem pomocí uvedených kategorií. Vzhledem k pravidelnostem v českém gramatickém systému však je výhodné neukládat do slovníku jednotlivé slovní tvary každý zvlášť, ale zprostředkovaně, ve formě vzorů, které umožňují podstatně zmenšit objem slovníku. Vzory představují jakási pravidla, jak vytvářet a rozpoznávat slovní tvary jazyka. Potřebujeme je pouze kvůli popisu paradigmat.

Systém vzorů pro ohýbání českých slov se učí děti na základní škole. Pro automatické zpracování jazyka je však třeba systém školních vzorů zjemnit. Stejně jako máme v současné době dva hlavní systémy morfologických značek, tak i systém vzorů byl pro češtinu vytvořen dvakrát, v Praze a v Brně.

V následujícím oddíle stručně popíšeme oba zmíněné systémy vzorů a vysvětlíme, proč jsme se rozhodli pro vytvoření systému nového. V dalších oddílech potom podrobně popíšeme nový systém ohýbacích vzorů.

8.1 Stručné porovnání pražského a brněnského systému vzorů

Oba systémy, tedy pražský i brněnský, vycházejí ze základních vzorů, které se učí děti na základní škole. Vzhledem ke způsobu, jakým se se vzory pracuje, bylo třeba v obou systémech repertoár vzorů podstatně rozšířit.

Klára Osolsobě ve své disertační práci (Osolsobě, 1996), na jejímž základě je brněnský systém postaven, o vzorech píše: „Pod pojmem *vzor* rozumíme ... konkrétní slovo reprezentující množinu všech slov, která tvoří ohebné tvary pomocí identického inventáře koncovek, jejichž společným rysem dále je, že tvoří paradigmaticky odvoditelné tvary podle příslušného slovního druhu, a u kterých dochází ke stejným změnám finální skupiny kmene."

Pražský systém (viz (Hajič, 2004)) chápe vzor obdobně, až na to, že místo konkrétního slova jako reprezentanta množiny slov se stejnými koncovkami definuje abstraktní množiny „koncovek". Poslední slovo je v uvozovkách proto, že se ve skutečnosti nejedná o klasické koncovky, ale obecnější řetězce. O nich pohovoříme za chvíli. Pražský vzor tedy není konkrétní slovo, ale řetězec znaků, který do jisté míry, zejména pro podstatná jména, využívá zkratek klasických vzorů (např. `hd` pro `hrad`, `zn` pro `žena`, `kr` pro `kuře`). Jejich jemnější členění se ale vyjadřuje pomocí dalších číslic i písmen (`hd1`, `hd1x`, `zn15e` atd.). Vzory pro ohýbání dalších slovních druhů vykazují také prvky logického systému, avšak již nepřipomínají klasické školní vzory.

Odlišné je řešení výjimek. Brněnský systém pokrývá vzory celou slovní zásobu, tedy i výjimky, které jsou jedinečné. Podle takového vzoru se potom

ohýbá právě jedno lemma, které je v tom případě shodné s názvem vzoru. Pražský systém řeší výjimky jinak — doplněním výjimečných tvarů s jejich konkrétními značkami ke vzorům, nebo, v případě, že nelze použít vzor ani pro část paradigmatu, uvedením všech slovních tvarů s konkrétními morfologickými značkami.

V principu nezáleží na tom, jak se vzor pojmenuje, ani jakým způsobem se zachází s výjimkami. Důležité je, aby systém pokryl veškeré možnosti, které se v jazyce vyskytnou. Oba systémy tuto podmínku splňují, ne však beze zbytku. Za hlavní nevýhody považujeme:

- Brněnské názvy vzorů nemají systém (není to kritika vzorů, ale jejich názvů!). Ztěžuje to práci se slovníkem. Ten, kdo slovník obhospodařuje, se musí naučit všechny vzory nazpaměť.
- Pražské vzory systém mají, ale zdaleka ne dokonalý. I zde je obtížné udržovat přehled o názvech vzorů.
- Ani jeden ze systémů nepokrývá veškeré flektivní varianty (mutace).
- Ve vzorech se kódují dohromady globální i flektivní kategorie.
- Není dobře vyřešeno odvozování příbuzných lemmat. Pražský systém se derivacemi zabývá, ale silně přegenerovává, takže lze odvodit i slovní tvary, které v jazyce nejsou.

Proto jsme se rozhodli navrhnout nový systém vzorů. Netvrdíme, že je zcela dokonalý, ale uvedené nedostatky odstraní. Dalším důvodem je změna systému kategorií, včetně jejich hodnot, popsaná v předchozích kapitolách. Právě s nimi by morfologie měla pracovat.

8.2 Nové vzory

Při návrhu vzorů vycházíme z pražského systému. Ten není přímo založen na morfologické stavbě slova, ani na fonologických pravidlech. Jeho hlavní myšlenkou je práce s řetězci[1].

Hlavní rozdíl našeho pojetí vzorů oproti oběma zmiňovaným systémům je ten, že kóduje pouze flektivní kategorie. Globální kategorie jsou totiž stejné pro celé paradigma (tak byly definovány), není tedy třeba je kódovat pro každý slovní tvar zvlášť. Můžeme tak použít stejný vzor pro více slovních druhů, např. adjektivní vzory pro podstatná jména, číslovky a zájmena.

Další změna spočívá ve striktním oddělení flektivní části vzoru a části derivační. Toto oddělení se sice v pražském systému používá, je však pro každý vzor neměnné. Každý vzor má napevno přiřazenu množinu derivací, což na jednu stranu velmi zjednodušuje popis[2], na druhou stranu však vede k vytváření neexistujících lemmat (např. od slovesa *pospíchat* se pravidelnou derivací utvoří příslovce *pospíchaně*.)

[1] Lingvisticky přijatelnější popis nabízí např. dvouúrovňová morfologie, kterou navrhl Kimmo Koskenniemi v r. 1983 (Koskenniemi, 1983). Pro češtinu implementovala dvouúrovňový popis Hana Skoumalová (Skoumalová, 1997).

[2] Autor tohoto systému Jan Hajič měl dobré důvody k tomu, aby byl popis co nejjednodušší, protože ve své době potřeboval, aby zabíral v paměti počítače co nejméně místa. Jeho kritici si toto zhusta neuvědomují.

Vzory, které budeme popisovat, se tedy skládají ze dvou částí — flektivní a derivační. Přesto, že obě části zapisujeme do jednoho řetězce, budeme jim říkat flektivní vzor a derivační vzor. **Flektivní vzor** popisuje slovní tvary, **derivační vzor** kóduje pravidla, podle kterých se tvoří nová lemmata.

8.2.1 Flektivní vzory

Flektivní vzor je množina řetězců a k nim náležejících flektivních částí morfologických značek (nadále budeme psát zkráceně jen o morfologické značce) a flektivních mutací, jde tedy o množinu trojic ⟨řetězec, morfologická značka, flektivní mutace⟩. Připojí-li se řetězec z flektivního vzoru ke správnému začátku slova (jinému řetězci), vyjde platné české slovo (slovní tvar), jemuž náleží morfologická značka a flektivní mutace z druhých dvou členů příslušné trojice vzoru. Zmíněným řetězcům ze vzoru říkáme **zakončení**. Obecně nemá zakončení žádný konkrétní lingvistický význam. Může to být koncovka, ale nemusí.

Formálně:

Flektivní vzor Ω je množina trojic ⟨s, M, F⟩, kde s je řetězec, M platná morfologická značka a F flektivní mutace. Nulová flektivní mutace se může z trojice vypustit.

Řekneme, že slovní tvar w s morfologickou značkou M a flektivní mutací F byl vytvořen podle flektivního vzoru Ω, jestliže existuje trojice $\langle \mathrm{s}, \mathrm{M}, \mathrm{F} \rangle \in \Omega$ taková, že $w = p{\cdot}s$ pro nějaký řetězec p. Tečka $\cdot$ je zde i dále znak pro operaci konkatenace (zřetězení).

Jestliže existuje řetězec p a flektivní vzor $\Omega = \{\langle s_i, M_i, F_i \rangle;\ i = 1...n\}$, pro který množina $\{p \cdot s_i;\ i = 1...n\}$ tvoří celé paradigma nějakého lemmatu $\bar{w}$, potom říkáme, že lemma $\bar{w}$ se ohýbá (skloňuje, časuje nebo stupňuje) podle vzoru Ω. Flektivní vzor však nemusí popisovat celé paradigma, ale jen jeho podmnožinu.

Řetězec p nemusí mít žádný gramatický význam, i když ho často má. Může být kmenem, jeho částí, může obsahovat i celou nebo jen část přípony, a samozřejmě může obsahovat i předpony. Z toho důvodu pro něj nemáme žádné přesné lingvistické pojmenování. V dalším textu ho budeme nazývat kofix. Tento termín vymyslel student MFF UK David Kolovratník jako zkratkové slovo pro nejasné spojení řetězců tvořících kmen, prefix i sufix nebo jejich částí, které dohromady tvoří začátek slovního tvaru.

Kofix je nejdelší počáteční řetězec, který sdílejí všechny slovní tvary popsané jedním vzorem.

Stejně tak řetězce s_i, které vystupují v nějakém vzoru, nejsou (gramatické) koncovky. Právě proto jim říkáme vágně **zakončení**.

Pro skloňování stačí, až na několik výjimek, jeden flektivní vzor. Například paradigma lemmatu *hrádek* má kofix *hrád* a množinu zakončení {*ek, ku, kem, ky, ků, kům, kách, cích*}.

U slovesných flektivních vzorů je situace komplikovanější. Slovesné tvary jednoho lemmatu se často tvoří podle několika různých vzorů, neboť v nich častěji dochází k hláskovým změnám. V takovém případě má paradigma více kofixů i více množin zakončení. Kdybychom chtěli zachovat zásadu, že k jednomu lemmatu přísluší jediný flektivní vzor, museli bychom pracovat s velmi krátkými kofixy a počet vzorů by musel být velký. Proto k vytvoření vzorů využíváme

společných vlastností jednotlivých slovesných tvarů a hláskové změny řešíme pomocí vícera kofixů. Jedno slovesné paradigma se potom časuje podle několika flektivních vzorů s několika kofixy.

Flektivní vzory jsou navrženy tak, aby pokrývaly i systematické flektivní mutace. Ty jsou uvedeny jako třetí člen trojic ve vzoru.

V této práci navrhujeme nový systém flektivních vzorů, ale základní myšlenka pro vytváření slovních tvarů pomocí vzorů a kofixů zůstává stejná jako v pražském systému. Jen místo dvojic ⟨kofix, vzor⟩, které jsou přiřazeny lemmatům v současném systému, používáme trojice ⟨kofix, vzor, mutace⟩. Mutace z této trojice může být flektivní, globální i smíšená. Jde o hodnotu kategorie `Mutace`, která byla zavedena v oddíle 3.4.

Lemmata, která se ohýbají nepravidelně, nemají flektivní vzor. Místo toho záznam slovníku obsahuje všechny slovní tvary jejich paradigmatu s příslušnými morfologickými značkami. Alternativně by se daly nepravidelnosti řešit pomocí individuálních vzorů, které by platily vždy jen pro jediné lemma. Tento způsob je použit v systému brněnském. My se držíme praxe pražské, která výjimky popisuje pomocí jednotlivých slovních tvarů. Vyhneme se tak příliš vysokému počtu flektivních vzorů. Navíc je možno jednoduše přidávat nepravidelné tvary bez ohledu na vzory. Např. lemma *otevřít* má kromě pravidelného časování dva archaické tvary *otevru* a *otevrou*, které jsou flektivními mutacemi k tvarům *otevřu* a *otevřou*. Podobné mutace vykazují ještě lemmata *zavřít* a *uzavřít*. Tyto mutace nejsou systematické, nepopisujeme je tedy pomocí vzoru, nýbrž je coby výjimky přidáváme do paradigmatu zvlášť.

Paradigma popisujeme tedy dvojím způsobem:

1. pomocí kofixů a flektivního vzorů,
2. pomocí slovních tvarů a jejich morfologických značek.

První způsob lze chápat jako zkrácený zápis způsobu druhého.

Oba typy mohou být v záznamu přítomny současně. Je-li třeba, musí být navíc doplněny informací o mutaci, a to jak globální, tak flektivní, aby vždy platilo Zlaté pravidlo morfologie.

Součástí slovníkového záznamu jsou tedy dva typy trojic:

1. ⟨kofix, flektivní vzor, mutace⟩ nebo
2. ⟨slovní tvar, morfologická značka, mutace⟩.

V případě, že mutace je pro daný kofix nebo slovní tvar nulová, nemusí se uvádět. Množina takových trojic definuje celé paradigma lemmatu, které je klíčem daného záznamu. Slovníkový záznam může obsahovat jednu nebo více trojic jednoho nebo obou typů.

Všechny takové trojice přiřazené jednomu lemmatu tvoří předpis pro vytvoření celého paradigmatu tohoto lemmatu.

Flektivní vzory, které popíšeme v následujících oddílech, jsou rozděleny podle slovních druhů. Místo velkého množství vzorů, které popisují slovní zásobu, jsme zavedli pro každý slovní druh vzorů jen několik, zato parametrizovatelných. Potřebné množství vzorů nahrazujeme parametry, které jsou však pro každý vzor jiné. Jsme přesvědčeni, že tento systém flektivních vzorů je vhodnější než dosavadní vzory obou zmiňovaných systémů. Jeho hlavní předností je to, že lze parametry snadno kombinovat tak, aby popsaly různé možnosti

flexe vyskytující se v českých paradigmatech. Parametry jsou zejména vhodné pro popis flektivních mutací slovních tvarů, což je zřejmě největší slabinou současných systémů.

Stávající vzory, jak brněnské, tak i pražské, lze většinou pomocí vhodně zvolených parametrů převést na vzory nové. U pražského systému se tímto způsobem můžeme zbavit poměrně velkého množství výjimek, které doplňují paradigmata některých lemmat, protože stávající množina vzorů je nezahrnuje. Nelze převést některé brněnské vzory, a to ty, které popisují nepravidelné alternace v kmeni. Jak již bylo řečeno, brněnský systém se vypořádává s výjimkami zavedením vzorů i pro velmi malé množiny lemmat, v nejzazším případě i pro lemma jediné. Pražský systém výjimky popisuje pomocí přiřazení značek jednotlivým tvarům. Tento způsob jsme převzali i my.

Názvy vzorů vytváříme výhradně pomocí znaků anglické abecedy. Jednotlivé znaky tedy kódujeme bez háčků a bez čárek. Znak *ě* kódujeme systematicky pomocí znaku j.

8.2.1.1 Stupňování

Vzory pro skloňování přídavných jmen obsahují i předpis pro stupňování[3]. Pomocí tohoto předpisu se vytvářejí pouze tvary druhého stupně (komparativu). Třetí stupeň a nově zavedený stupeň `s` (*sebe-* + komparativ) se tvoří pravidelně ze stupně druhého, a to pomocí prefixů *nej-* a *sebe-*. Stejně pravidelné je i stupňování negovaných adjektiv (*nebezpečnější, nejnebezpečnější, sebenebezpečnější*).

Pražský systém vzorů obsahuje i předpis na tvoření superlativu, a to pomocí značky + u zakončení. Naše nové vzory to nedělají, neboť tvoření je naprosto pravidelné a může se tvořit od každého přídavného jména, které má druhý stupeň. Totéž se týká i stupňování příslovcí. Je-li tedy ve vzoru uveden předpis na vytvoření 2. stupně, automaticky to znamená, že se vytvoří i stupeň `3` a stupeň `s`. Starost o vytváření tvarů 3. stupně a stupně `s` u přídavných jmen a příslovcí necháváme tedy na nástrojích morfologické analýzy a syntézy.

8.2.1.2 Negace

Podobně zacházíme i s negací. Předpona *ne-*, která slouží k vytvoření negovaného tvaru přídavného jména, příslovce a slovesa, řídčeji i podstatného jména, nebývá obsažena ve slovníku, neboť je pravidelná. Proto ani u negace nepoužíváme praxe z pražského systému, který možnost negace vyznačuje speciálním znakem ve všech zakončeních uvedených ve vzoru.

Zde však existuje několik výjimek. Infinitiv některých sloves se totiž nevytváří prostým připojením *ne-* na začátek slova, ale dochází zde ke krácení v kmeni. Jako příklad uveďme infinitivy *brát — nebrat*, *spát — nespat*. Tyto tvary je třeba do slovníku zadat explicitně a současně zabránit, aby se tvořily nesprávné dlouhé tvary automaticky z afirmativního tvaru. Tomu zabráníme zavedením **parametru negace** přímo do názvu vzoru. Tento parametr vystupuje ve vzorech ve formě nepovinného prefixu (proto mu také říkáme **prefix**

[3] O stupňování se lingvisté přou, zda patří do morfologie nebo do slovotvorby, viz např. (Karlík – Hladká, 2004). Držíme se zavedené pražské i brněnské praxe a popisujeme stupňování v rámci morfologie.

negace), který může mít hodnotu A nebo N. Hodnota A znamená, že vzor lze použít pouze ke generování afirmativních tvarů, hodnota N připouští pouze tvary negativní. Absence prefixu neklade na negaci žádná omezení, tedy povoluje tvary jak afirmativní, tak negativní. Parametr negace může vystupovat jen u vzorů přídavných jmen, sloves, příslovcí a podstatných jmen. Vzhledem k velmi volným pravidlům při tvoření negace je lepší omezení na negaci ve vzorech nepoužívat, nemění-li negace kmen slovního tvaru. Zabránili bychom tak rozpoznání neobvyklých tvarů, jako v příkladu (109) z korpusu SYN a v příkladu (110) nalezeném na internetu.

Blbec neblbec, ale hlavně, že jsem zdravý (109)
V tradiční (nerychlé) restauraci však můžeme vidět také velké rozdíly (110)

Právě popsaný přístup k negaci a stupňování již byl částečně implementován, prozatím s původními pražskými značkami, viz (Hlaváčová – Kolovratník, 2008). S prefixem negace se zde zachází podobně jako s jinými prefixy s tím rozdílem, že se při rozpoznání změní hodnota kategorie NEG z NEG=A na NEG=N. Prefixy používané ke stupňování zase mění hodnotu kategorie DEG z DEG=2 na DEG=3 nebo DEG=s.

8.2.2 Derivační vzory

Velké množství českých slov vzniká odvozením z jiných slov. Na otázku, zda derivace patří do morfologie či nikoli, není jednoznačná odpověď. Vzhledem ke způsobu použití morfologického slovníku však derivační vztahy do slovníku zahrnujeme. Např. v automatickém překladu je občas nutné použít místo slovesa slovesné přídavné jméno (*Udělal jsem to. — Mám to udělané / uděláno.*). Jestliže budeme mít ve slovníku odkazy mezi slovesem a příslušnými deverbativy, můžeme tyto vztahy využít při nejrůznějších konstrukcích.

Jsou v zásadě dvě možnosti, jak to udělat:

1. pomocí pravidel,
2. pomocí odkazů mezi jednotlivými hesly.

Nejoperativnější je kombinovaný přístup. V případech, kdy se derivace tvoří naprosto pravidelně, postačí do slovníkového záznamu vložit pravidlo, jak z daného lemmatu vytvořit lemma odvozené.

Derivační vzor však nepřiřazujeme přímo k lemmatu, ale ke kofixu. Pomocí tohoto vzoru se odvodí z daného kofixu nové lemma se svým flektivním vzorem (včetně přesně nastavených hodnot všech parametrů). Z toho vyplývá, že derivační pravidla většinou nepoužíváme k odvozování nepravidelných lemmat, které nemají jeden pravidelný flektivní vzor. Derivační vzor se také využije pro utvoření derivačního odkazu mezi oběma lemmaty — tedy toho, v jehož záznamu je uvedeno, a toho, které se pomocí něho odvodí.

Na rozdíl od ohýbání tedy tvary vzniklé na základě derivačních vzorů nejsou součástí paradigmatu lemmatu, jehož tvary vzor popisuje. Derivační vzory vždy vytvářejí nová lemmata s vlastními paradigmaty.

Použití vzorů pro tvoření derivačních odkazů má několik výhod. Jednak se tím může výrazně zmenšit velikost slovníku. Víme-li např., jak z nedokonavých sloves páté třídy (vzor *dělá*) utvořit iterativní slovesa (*dělávat*), není třeba

je všechny zahrnovat do slovníku, vygenerují se samy pomocí jednoduchého pravidla.

Druhá výhoda spočívá v implicitní možnosti propojit v rámci slovníku příbuzná slova. Příbuzná slova lze propojit ručně, avšak automatické propojení pomocí derivačních vzorů je elegantnější a méně náchylné k chybám. V těch případech, kdy není možné odvozené slovo utvořit podle jednoduchého pravidla, je stále možnost v rámci slovníku pomocí odkazů příbuzná lemmata propojit ručně. Odvozovací pravidla však tento proces zjednodušují.

Další výhodou je snížení pravděpodobnosti chyby při údržbě slovníku. Jakýkoli ruční zásah do obsahu slovníku s sebou nese riziko chyby, proto je třeba se snažit o co největší využití pravidelných operací.

Některá odvozovací pravidla jsou velmi produktivní a pokrývají velké množství příbuzných dvojic. Existují však i derivace nepravidelné. Jistě je možno vytvořit pravidlo pro každé odvození. Ovšem derivační vzor, který by platil jen pro malý počet, případně dokonce jen jedno odvození, by neúměrně zvyšoval objem celého systému. Přijímáme stejnou zásadu jako v případě ohýbacích vzorů, které nevytváříme pro nepravidelná paradigmata. Ani zde nebudeme vytvářet vzory pro nepravidelná odvození.

Místo toho zařadíme odvozené slovo do slovníku jako plnohodnotné heslo a jeho souvislost s původním slovem zajistíme pomocí derivačního odkazu.

Vzhledem k tomu, že často není ani lingvistům jasné, které lemma je původní a které odvozené, nebudeme se snažit o zachycení směru tohoto vztahu a derivační odkazy umístíme do slovníku oboustranné bez jakýchkoli preferencí. Přesto budeme tyto odkazy i nadále nazývat derivační.

Derivační vztah v našem pojetí tedy propojuje dva záznamy bez ohledu na skutečné (nebo aspoň všeobecně přijímané) časové vztahy (co bylo dřív a co potom).

9 Vzory podstatných jmen

9.1 Obecné vlastnosti

Pro skloňování podstatných jmen používáme základní vzory, které se učí děti na základní škole. Zapisujeme je zkratkou složenou ze dvou písmen. Tato zkratka je většinou shodná se zkratkou vzorů pražského systému a mnemotechnicky vyjadřuje klasický základní vzor.

Zakončení z jednotlivých vzorů podstatných jmen jsou pravidelná, až na některé kombinace pádu a čísla, ve kterých dochází k více možnostem tvoření příslušného slovního tvaru. Je-li tato kombinace přípustná pro celou množinu lemmat ohýbaných podle daného vzoru, je hodnota kategorie `Flektivní mutace` přímo zahrnuta do příslušné trojice ⟨řetězec, morfologická značka, flektivní mutace⟩ vzoru. Příkladem může být dvojí tvar podstatných jmen v 7pl (*hrady* — *hradama*). Jestliže nějaká kombinace připouští více možností zakončení, ale ne pro všechna lemmata, říkáme jí **kritická kombinace**. U každého lemmatu musíme stanovit, které zakončení je pro něj správné, a přiřadit mu i hodnotu kategorie `Flektivní mutace` (např. 2sg *lesa* s FMU=a, ale *hradu* s FMU=u). U každého vzoru jsou kritické kombinace jiné, některé vzory kritické kombinace nemají, protože jsou zcela pravidelné.

Celý vzor podstatného jména se skládá z následujících částí:

1. z dvoupísmenné části kódující základní vzor,
2. z části kódující koncové znaky kofixu (většinou přímo slovního kmene),
3. z části kódující možná zakončení v kritických kombinacích,
4. z části definující derivace,
5. z nepovinné části kódující různá omezení na tvoření některých slovních tvarů.

Části 2 až 4 jsou závislé na základním vzoru, část 5 je nepovinná a může být přidána k libovolnému substantivnímu vzoru.

Část 2 říká, jak vypadá posledních několik znaků kmene, neboť podle nich se řídí množina zakončení. Často jsou to přímo koncové znaky, a stávají se tak součástí zakončení ze vzoru.

Možnosti zakončení v kritických kombinacích jsou zakódovány ve 3. části vzoru. Obecně budeme kritické kombinace zapisovat ve tvaru `Ksg` pro jednotné číslo nebo `Kpl` pro číslo množné, kde `K` je číslo pádu (hodnota kategorie `Pád`). V případě, že je ve vzoru uvedeno více hodnot pro některou kritickou kombinaci, vygenerují se podle nich příslušné hodnoty kategorie `Flektivní mutace`. Pořadí, ve kterém se mutace uvedou do vzoru, není významné, neboť mutace nejsou umístěny na žádnou škálu. Při práci na přiřazování vzorů se však snažíme zadávat jako první možnost tu mutaci, která je spisovná nebo, v případě obou spisovných nebo obou nespisovných, běžnější. Kdyby se totiž někdo rozhodl mutace číslovat, mohlo by mu takové řazení usnadnit práci.

Vzor může také obsahovat předpis na vytvoření derivací. V současné práci popisujeme pouze odvození přivlastňovacích přídavných jmen a u životných vzorů také odvozená feminina. Přivlastňovací přídavná jména je možno tvořit příponou *-ův* pro hodnoty GEN=M nebo příponou *-in* pro hodnotu GEN=F. Tvoření přivlastňovacího přídavného jména je pro jednotlivé vzory pravidelné, stačí tedy zadat, zda se tvoří, nebo netvoří. Příznak `V` tvoření přídavného jména přivlastňovacího umisťujeme do části 4, ale explicitně ho uvádíme jen u ženských vzorů, neboť u mužských životných vzorů lze přídavné jméno přivlastňovací utvořit zřejmě vždy.

O tvoření odvozených feminin pojednáme v oddíle 9.3 o životných vzorech.

Každý vzor navíc může obsahovat na konci znak P, znamenající, že se generují jen tvary plurálu, nebo S pro tvoření pouze singuláru. Není-li přítomen ani jeden z těchto znaků, generují se všechny tvary plurálu i singuláru.

Další vlastnost společná všem vzorům může být označení skutečnosti, že vzor se nemá použít pro vytvoření tvaru pro 2pl. Existuje totiž poměrně značné množství paradigmat, ve kterých právě v tomto pádě dochází k alternaci ve kmeni, přičemž všechny ostatní tvary jsou pravidelné, nebo aspoň pravidelné tvary tvoří jednu z více mutací (např. *bouda — bud, dílo — děl, kráva — krav, chvíle — chvil*). Jak je vidět z příkladů, je tato vlastnost společná více vzorům, i když zdaleka ne všem. Přesto je výhodnější připustit tuto možnost obecně pro substantivní vzory, než ji vypisovat u každého zvlášť. Fakt, že vzor platí pro všechny tvary kromě 2pl, se vyznačí uvedením znaku `2` na konec názvu vzoru.

Obě omezení, na číslo i na (ne)pravidelné tvoření 2pl, jsou obsahem části 5.

Substantivní vzory tedy mohou končit těmito řetězci, které omezují tvoření slovních tvarů:

- S: jen tvary jednotného čísla
- P: jen tvary množného čísla
- 2: všechny tvary kromě 2pl
- 2P nebo P2: všechny tvary množného čísla kromě 2pl (průnik dvou předchozích omezení)

Omezení negace se vyjádří pomocí parametru negace, který se však umisťuje systematicky u vzorů všech slovních druhu jako prefix před základním vzorem. Viz oddíl 8.2.1.2 v obecné kapitole o vzorech.

Všechny vzory podstatných jmen mají pravidelné flektivní mutace FMU=a se zakončením *-ma*, které jsou možné vždy v 7pl: *pány — pánama, stroji — strojema, ženami — ženama, kuřaty — kuřatama* atp. Tyto mutace jsou též součástí všech vzorů, kde o nich též pojednáme.

Nejběžnější mutace podstatných jmen jsme umístili do tabulky 9.1, která je uvedena na str. 84.

Následuje seznam vzorů, jejich popis a příklady.

V názvu vzoru je vždy tučně uveden dvoupísmenný kód názvu vzoru. Podtržené znaky jsou parametry, jejichž možné hodnoty jsou uvedeny ve výčtu, který vždy následuje. Hodnota ϵ znamená prázdný znak a ve skutečnosti se nezapisuje. Část vzoru 5 popisující možnost obecných omezení již u jednotlivých

POS	GEN	CAS	NUM	Mutace	Jeji kód	Mutace	Jeji kód	Mutace	Jeji kód
				Příklad		Příklad		Příklad	
N	I	2	S	*a*	**a**	*u*	**u**		
				kouta		*koutu*			
N	I	3	S	*u*	**u**	*i*	**i**		
				kořenu		*kořeni*			
N	I	6	S	*u*	**u**	*e/ě*	**e**	*i*	**i**
				obchodu/ lesu/ kořenu		*lese/obchodě*		*kořeni*	
N	IMN	6	P	*ích*	**i**	*ách*	**a**	*ech*	**e**
				domcích/ ptácích/ ramíncích/hotelích		*domkách/ ptákách/ ramínkách*		*hotelech/spisovatelech*	
N	M	36	P	*u/i*	**K**	*ovi*	**D**		
				pánu/muži/soudci		*pánovi/mužovi/soudcovi*			
N	M	15	P	*i*	**i**	*é*	**e**	*ové*	**v**
				invalidi		*invalidé*		*invalidové*	
N	MIN	3	P	*ům*	**d**	*um*	**k**		
				domům/pánům/městům		*domum/pánum/městum*			
N	F	1245	P	*i*	**i**	*e/ě*	**e**		
				lodi/noci		*noce/lodě*			
N	F	3	P	*ím*	**i**	*em/ěm*	**e**		
				lodím/nocím/Dejvicím		*nocem/loděm*			
				ům	**u**	*um*	**uk**		
				Dejvicům		*Dejvicum*			
N	FM	6	P	*ích*	**i**	*ech*	**e**		
				nocích/obyvatelích		*nocech/obyvatelech*			
N	F	7	P	*mi*	**K**	*emi/ěmi*	**0**	ema/ěma	**ea**
				loďmi/nocmi		*nocemi/loděmi*		*nocema/loděma*	
				ma	**ka**	*ima*	**ia**		
				loďma/nocma		*lodima*			
N	F	1	S	*e*	**e**	*a*	**a**		
				Marie		*Maria*			
N	F	36	S	*e*	**e**	*i*	**i**		
				Saše		*Saši*			
N	F	145	P	*e*	**e**	*i*	**i**		
				Saše		*Saši*			
N	N	2	S	*í*	**0**	*ího*	**h**		
				stavení		*staveního*			
N	N	6	S	*u*	**u**	*e/ě*	**e**		
				městu/mléku		*mléce/městě*			
N	F	2	P	*í*	**i**	-	**0**		
				jeskyní		*jeskyň*			
N	M	1-7	SP	vložené e	**e**	bez e	**E**		
				Bergerovi		*Bergrovi*			
N	N	2	P	vložené e	**e**	bez e	**E**		
				stanovisek		*stanovisk*			
N	M	7	P	*i*	**0**	*ama*	**a**	*ema*	**ea**
				obyvateli/muži		*obyvatelama*		*obyvatelema*	
N	MI	kromě 1S		*[oeus]s.**	**s**	*.**	**0**		
				Kolumbusovi		*Kolumbovi*			
NA	.	7	P	*mi*	**0**	*ma*	**a**		
				ženami/ drahými/ pány		*ženama/ drahýma/ pánama*			

Tabulka 9.1: Nejběžnější mutace podstatných jmen

vzorů nezmiňujeme, neboť je obecně použitelná u libovolného substantivního vzoru.

9.2 Neživotné vzory

9.2.1 HRAD

hdx-2sg-6sg-6pl

hdx-2sg-6sg**S**

hdx-2pl-6pl**P**

Vzor `hrad` má tři varianty. První varianta je nejčastější. Podle ní se vytvářejí všechny tvary jednotného i množného čísla. Kritické kombinace jsou, jak je vidět ze vzoru, 2sg, 6sg a 6pl.

Druhá varianta popisuje singularia tantum. V tomto případě jsou kritické kombinace pouze 2sg a 6sg a název celého vzoru musí být zakončen znakem `S`.

Třetí varianta popisuje pluralia tantum. Z kritických kombinací pochopitelně vypadnou 2sg a 6sg, naopak k 6pl přibyde 2pl, neboť existuje velká skupina substantiv, která má prázdné zakončení v 2pl (např. *Dukovany*, 2pl *Dukovan*, nikoli **Dukovanů*, jak by odpovídalo vzoru `hrad`). V tomto případě musí být na konci vzoru `P`.

Druhou a třetí variantu lze použít i pro popis takových paradigmat, jejichž množné a jednotné číslo mají různý kofix.

Význam jednotlivých částí:

x je zakončení kofixu. Většinou se jedná o poslední jedno až dvě písmena lemmatu.
Možné hodnoty parametru x:

- `r`
- `ch`
- `h`
- `g`
- `ek` (ve 2pl se vypouští -*e*-)
- `k`
- `en` (ve 2pl se vypouští -*e*-)
- `el` (ve 2pl se vypouští -*e*-)
- `et` (ve 2pl se vypouští -*e*-)
- `us`
- `ky`
- `y`
- ϵ — ostatní možná zakončení

2sg je zakončení 2. pádu jednotného čísla. Možné hodnoty parametru 2sg jsou současně skutečná zakončení. U kódů uvádíme též hodnotu kategorie `Flektivní mutace`.

- `a`, FMU=a
- `u`, FMU=u

Jestliže jsou možné obě hodnoty parametru 2sg, použijí se nejen k vytvoření slovních tvarů, ale i k přiřazení hodnoty `Flektivní mutace`.

6sg je zakončení 6. pádu jednotného čísla. Možné hodnoty parametru 6sg a kategorie `Flektivní mutace`:

- `u`, FMU=u
- `e`, FMU=e
- `j` pro zakončení -*ě*, FMU=e

Kategorie FMU je stejná u dvou zakončení proto, že tyto dvě alternativy se nikdy nemohou vyskytnout u stejné kombinace pádu a čísla jednoho lemmatu. To vyplývá z historického vývoje českého skloňování. Zakončení -*ě* nastává po hláskách b, f, m, p, v, d, t, n, zatímco -*e* po hláskách l, s, z. Se stejným typem mutace se setkáme ještě několikrát.

6pl je zakončení 6. pádu množného čísla. Možné hodnoty parametru 6pl:

- `a` pro zakončení -*ách*, FMU=a
- `e` pro zakončení -*ech*, FMU=e
- `i` pro zakončení -*ích* s automatickým změkčením předchozí souhlásky, FMU=i

2pl je zakončení 2. pádu množného čísla. Možné hodnoty parametru 2pl:

- `u` pro zakončení -*ů*, FMU=u
- `0` pro prázdné zakončení, FMU=0

Příklady různých lemmat se vzorem `hrad` jsou v tabulce 9.2. V posledním sloupci je uveden současný vzor z pražského systému.

9.2.2 STROJ

sjx

Vzor `stroj` je velmi pravidelný, nemá žádné kritické kombinace. Jediným parametrem je zakončení kofixu **x**.
Možné hodnoty parametru x jsou:

- `el`
- `ec`
- `en` pro zakončení -*eň*
- ϵ — ostatní možná zakončení

Příklady různých lemmat se vzorem `stroj` jsou v tabulce 9.3. V posledním sloupci je uveden současný vzor z pražského systému.

Lemma	Kofix	Vzor	Současný pražský vzor
problém	*problém*	`hd-u-u-e`	`hd1`
hotel	*hotel*	`hd-u-u-ei`	`hd1xx`
cirkus	*cirkus*	`hd-u-u-e`	`hd1`
virus	*vir*	`hdus-u-u-e`	`hdus`
kurs	*kurs*	`hd-u-eu-e`	`hd2`
oceán	*oceán*	`hd-u-ju-e`	`hd2x`
javor	*javor*	`hd-au-u-e`	`hd4`
Motol	*Motol*	`hd-a-euS`	`hd4`
Vlkov	*Vlkov*	`hd-au-ju-e`	`hd5x`
skřek	*skře*	`hdk-u-u-ia`	`hd1k`
hrádek	*hrád*	`hdek-u-u-ia`	`hd1ek`
srpen	*srp*	`hden-a-u-e`	`hd1ena`
svícen	*svíc*	`hden-u-u-e`	`hd1en`
tucet	*tuc*	`hdet-u-u-e`	`hd1et`
jazyk	*jazy*	`hdk-au-eu-ia`	`hdka`
gong	*gon*	`hdg-u-u-ia`	`hd1g`
běh	*bě*	`hdh-u-u-ia`	`hd1h`
poslech	*posle*	`hdch-u-u-ia`	`hd1ch`
Slapy	*Slap*	`hdy-u-eP`	`hdpy`
Dukovany	*Dukovan*	`hdy-0-eP`	`hdpy0`
Jeseníky	*Jesení*	`hdky-u-aiP`	`hdpky`

Tabulka 9.2: Příklady lemmat ohýbaných podle vzoru `hrad`

Lemma	Kofix	Vzor	Současný pražský vzor
jetel	*jetel*	`sj`	`sj1`
řetězec	*řetěz*	`sjec`	`sj1ec`
obratel	*obrat*	`sjel`	`sj1el`
oheň	*oh*	`sjen`	`--`

Tabulka 9.3: Příklady lemmat ohýbaných podle vzoru `stroj`

9.2.3 Kolísání mezi vzory HRAD a STROJ

hs

Název vzoru je utvořen z počátečních písmen obou vzorů. Podstatná jména, která kolísají mezi vzory `hrad` a `stroj`, jsou většinou pravidelná, proto v jejich vzoru nejsou žádné parametry. Pravděpodobně jedinými nepravidelnými výjimkami jsou lemmata *den* a *týden* (viz (Osolsobě, 1996) na str. 90).

V případě, že se tvary podle obou vzorů neshodují, jsou rozlišeny hodnotou kategorie `Flektivní mutace`. Kombinace kategorií jsou uvedeny v tabulce 9.4.

Pád	Zakončení (`hrad`)	FMU	Zakončení (`stroj`)	FMU	Příklad
2sg	*-u*	u	*-e*	e	*boubelu/boubele*
3sg/6sg	*-u*	u	*-i*	i	*boubelu/boubeli*
5sg	*-e*	e	*-i*	i	*boubele/boubeli*
1pl/4pl	*-y*	y	*-e*	e	*boubely/boubele*
6pl	*-ech*	e	*-ích*	i	*boubelech/boubelích*
7pl	*-y*	y	*-i*	i	*boubely/boubeli*

Tabulka 9.4: Tabulka mutací při kolísání mezi vzory `hrad` a `stroj`

9.2.4 ŽENA

znx

Význam jednotlivých parametrů:

U vzoru `žena` je parametr **x** povinný (není nikdy ϵ), a pokud není uvedeno jinak, znamená poslední souhláskovou skupinu před koncovým *a* v 1sg. Hodnoty jsou uvedeny v tabulce 9.5, ze které jsou též vidět klíčové kombinace pádu a čísla, kvůli nimž se jednotlivé hodnoty musí rozlišovat. V posledním sloupci tabulky je zakončení lemmatu přídavného jména přivlastňovacího, pokud se tvoří.

V posledních třech řádcích tabulky jsou klíčová zakončení uvedena v hranatých závorkách ([x], [y], [j]) proto, že zastupují více znaků. Konkrétně to jsou:

x : b, d, f, m, n, p, t, v
y : l, s, z
j : ž, š, č, ř, c, j... 1pl není *y*, ale *i*

Příklady uvádíme v tabulce 9.6. V posledním sloupci je uveden současný vzor z pražského systému.

Par.	1sg	[36]sg	2pl	V	Par.	1sg	[36]sg	2pl	V
k	*-ka*	*-ce*	*-k*	*-čin*	re	*-ra*	*-ře*	*-er*	*-řin*
r	*-ra*	*-ře*	*-r*	*-řin*	ke	*-ka*	*-ce*	*-ek*	*-čin*
g	*-ga*	*-ze*	*-g*	*-zin*	dj	*-ďa*	*-dě*	*-ď*	*-din*
h	*-ha*	*-ze*	*-h*	*-žin*	tj	*-ťa*	*-tě*	*-ť*	*-tin*
c	*-cha*	*-še*	*-ch*	*-šin*	nj	*-ňa*	*-ně*	*-ň*	*-nin*
e	*-ea*	*-ey/eje*	*-í*	—	nk	*-ňka*	*-ňce*	*-něk*	*-ňčin*
ve	*-va*	*-vě*	*-ev*	*-vin*	tk	*-ťka*	*-ťce*	*-těk*	*-ťčin*
ne	*-na*	*-ně*	*-en*	*-nin*	dk	*-ďka*	*-ďce*	*-děk*	*-ďčin*
le	*-la*	*-le*	*-el*	*-lin*	x	*-[x]a*	*-[x]ě*	*-[x]*	*-[x]in*
be	*-ba*	*-bě*	*-eb*	*-bin*	y	*-[y]a*	*-[y]e*	*-[y]*	*-[y]in*
te	*-ta*	*-tě*	*-et*	*-tin*	j	*-[j]a*	*-[j]e*	*-[j]*	*-[j]in*
me	*-ma*	*-mě*	*-em*	*-min*					

Tabulka 9.5: Možná zakončení vzoru `žena`

Lemma	Kofix	Vzor	Současný pražský vzor
půda	*půd*	`znx`	`zn1`
teta	*tet*	`znxV`	`zn1n`
bouda	*boud*	`znx2`	`zn1x`
tráva	*tráv*	`znx2`	`zn1x`
děl̆ba	*děl*	`znbe`	`zn2e`
tóga	*tó*	`zng`	`zn3`
duha	*du*	`znh`	`zn4`
archa	*ar*	`znc`	`zn5`
informatika	*informati*	`znk`	`zn6`
Maruška	*Maruš*	`znkeV`	`zn6en`
námitka	*námit*	`znke`	`zn6e`
koza	*koz*	`zny`	`zn7`
jehla	*jeh*	`znle`	`zn7e`
farma	*far*	`znme`	`zn8e`
pryčna	*pryč*	`znne`	`zn9e`
lodka	*lo*	`zndk`	`zn10e`
laňka	*la*	`znnk`	`zn12e`
mezera	*meze*	`znr`	`zn13`
hra	*h*	`znre`	`zn13e`
jachta	*jach*	`znte`	`zn14e`
jizva	*jiz*	`znve`	`zn15e`
Korea	*Kore*	`zne`	`zn19`
Saša	*Saš*	`znjV`	`zn20`
dožínky	*dožín*	`znkeP`	`znp6ge`
Karpaty	*Karpat*	`znxP`	`zn26`
Káťa	*Ká*	`zntjV`	`--`
Tatry	*Tat*	`znreP`	`--`

Tabulka 9.6: Příklady lemmat ohýbaných podle vzoru `žena`

9.2.5 PÍSEŇ

psx**e**

Význam jednotlivých částí:

x je kód posledního znaku lemmatu:
- `d` pro zakončení *-ď*
- `t` pro zakončení *-ť*
- `n` pro zakončení *-ň*
- `v` pro zakončení *-v*
- `c` pro zakončení *-č*
- ϵ pro ostatní zakončení

Parametr **e** může mít tyto hodnoty:
- `e` když se v 2sg vypouští *e*
- ϵ v ostatních případech

Vzory `psv` ani `psc` neexistují, protože v případech, kdy se před koncové *-v, -č* nevkládá ve 2pl *-e-*, se použije obecnější vzor `ps` (*kleč, křeč*, a další). Příklady jsou v tabulce 9.7.

Lemma	Kofix	Vzor	Současný pražský vzor
třešeň	*třeš*	`psne`	`ps1e`
Výtoň	*Výto*	`psn`	`ps1`
brukev	*bruk*	`psve`	`ps2`
Ohaveč	*Ohav*	`psceS`	`--`

Tabulka 9.7: Příklady lemmat ohýbaných podle vzoru `píseň`

9.2.6 KOST

kt

Vzor `kost` je zcela pravidelný, nemá tedy žádné parametry. Příklady jsou v tabulce 9.8.

Lemma	Kofix	Vzor	Současný pražský vzor
kost	*kost*	`kt`	`kt1`
dveře	*dveř*	`ktP`	`--`

Tabulka 9.8: Příklady lemmat ohýbaných podle vzoru `kost`

Existuje však množství lemmat, jejichž skloňování kolísá mezi vzory `kost` a `píseň`.

9.2.7 Kolísání mezi vzory KOST a PÍSEŇ

kpx-2sg-36pl-7pl

x je kód posledního znaku lemmatu:

- `d` pro zakončení -*ď*
- `t` pro zakončení -*ť*
- `n` pro zakončení -*ň*
- ϵ pro ostatní zakončení

Kritické kombinace můžeme rozdělit do tří skupin. Jsou to:

1. 2sg, 1pl, 4pl a 5pl, kódováno parametrem 2sg
2. 3pl a 6pl, kódováno parametrem 36pl
3. 7pl, kódováno parametrem 7pl

V ostatních kombinacích čísla a pádu jsou oba vzory totožné.

2sg je kód zakončení tvaru v 2sg, 1pl, 4pl a 5pl:

- `i` pro zakončení -*i* — FMU=i
- `e` pro zakončení -*e* — FMU=e
- `j` pro zakončení -*ě* — FMU=e

36pl je kód zakončení tvarů v 3pl a 6pl:

- `i` pro zakončení -*ím* v 3pl a -*ích* v 6pl — FMU=i
- `e` pro zakončení -*em* v 3pl a -*ech* v 6pl — FMU=e

7pl je kód zakončení tvaru v 7pl. Flektivní mutaci -*ma* ve výčtu pro 7pl neuvádíme, ta je pravidelná pro všechna podstatná jména. Může být k uvedeným mutacím přidána (např. tedy zakončení -*ema* má FMU=ea).

- `e` pro zakončení -*emi* — FMU=e
- `j` pro zakončení -*ěmi* — FMU=e
- `0` pro zakončení -*mi* — nulová flektivní mutace

Příklady jsme většinou převzali ze seznamu vzorů skupiny `kost` z disertační práce Kláry Osolsobě (Osolsobě, 1996). Jsou v tabulce 9.9.

Současné pražské vzory, u kterých je uveden znak $*$, nepopisují celé paradigma příslušného lemmatu. Tvary, které jsou v paradigmatu navíc, jsou uvedeny ve slovníku jako výjimky.

Lemma	Kofix	Vzor	Současný pražský vzor
mast	*mast*	`kp-ij-ie-0j`	`kt1`
moc	*moc*	`kp-ie-ie-e`	`ps18`
myš	*myš*	`kp-i-i-0`	`--`
noc	*noc*	`kp-i-i-e`	`--`
žluč	*žluč*	`kp-ie-i-e`	`ps5 *`
ocel	*ocel*	`kp-ie-i-e0`	`ps5 *`
oběť	*obě*	`kpt-i-ei-0`[1]	`--`
hruď	*hru*	`kpd-ij-i-j`	`ps5 *`
odpověď	*odpově*	`kpd-i-ei-j0`	`--`
choť	*cho*	`kpt-ij-i-j`	`ps7 *`
loď	*lo*	`kpd-ij-i-0j`	`ps6`

Tabulka 9.9: Příklady lemmat, která kolísají mezi vzory `píseň` a `kost`

9.2.8 NŮŠE

nsx-2pl-1sg

Kritickými kombinacemi vzoru `nůše` je 2pl, který může mít zakončení *-í* nebo prázdnou koncovku (např. *žákyně* — *žákyň* i *žákyní*).

U mutací typu *Marie* — *Maria* dochází k nejednoznačnosti i v 1sg.

Význam jednotlivých částí:

x je kód předposledního znaku lemmatu, před koncovým *-e/ě*:

- `n` pro zakončení *-ně*, je-li v 2pl zakončení *-ň*
- ϵ pro ostatní zakončení

2pl je kód zakončení 2pl. V případě, že zde dochází k alternaci ve kmeni, která znemožňuje pravidelné tvoření, může být tento parametr zcela vypuštěn. V tom případě je však povinné označení `2` na konci názvu vzoru. Parametr 2pl může mít tyto hodnoty:.

- `i` pro zakončení *-í*, FMU=D
- `0` pro prázdné zakončení, FMU=K

1sg je kód zakončení tvaru v 1sg:

- `e` pro zakončení *-e*, FMU=e
- `j` pro zakončení *-ě*, FMU=e
- `a` pro zakončení *-a*, FMU=a

Příklady uvádí tabulka 9.10.

[1] Mutaci *obětmi* (FMU=t) je třeba zachytit jako výjimku.

Lemma	Kofix	Vzor	Současný pražský vzor
nůše	*nůš*	`ns-i-e`	`ns1`
košile	*košil*	`ns-0i-e`	`ns1`
země	*zem*	`ns-i-j`	`ns2`
tradice	*tradic*	`ns-0i-e`	`ns10`
justice	*justic*	`ns-i0-e`	`ns3x`
učnice	*učnic*	`ns-0-e`	`ns10`
žákyně	*žáky*	`nsn-i0-jV`	`ns3n`
Natalie	*Natali*	`ns-i-eaV`	`ns1`
brýle	*brýl*	`ns-i-eP`	`ns7`
Bohunice	*Bohunic*	`ns-eS`	`ns07`
saně	*san*	`ns-i-jP`	`--`

Tabulka 9.10: Příklady lemmat ohýbaných podle vzoru `nůše`

9.2.9 MĚSTO

mtx-6sg-2pl

mtx-6sg**S**

mtx-2pl**P**

Vzor `město` má tři varianty. První varianta je nejběžnější. Podle ní se vytvářejí všechny tvary jednotného i množného čísla. Kritické kombinace jsou, jak je vidět ze vzoru, 6sg a 2pl.

Druhá varianta popisuje singularia tantum. V tomto případě je kritická kombinace pouze 6sg a název celého vzoru musí být zakončen znakem `S`.

Třetí varianta popisuje pluralia tantum. Z kritických kombinací pochopitelně vypadne 6sg, zůstane jen 2pl. V tomto případě musí být na konci vzoru `P`.

Mutace *-kách* a *-cích* (*políčko* — *políčkách* i *políčcích* v 6pl se týkají jen lemmat končících na *-ko* a jsou zcela pravidelné, není tedy třeba je zahrnovat mezi kritické.[2]

Pod vzor `město` spadají i neutra z latiny a řečtiny, končící na *-um* a *-on*.

[2] V práci (Osolsobě, 1996) jsou zmíněny i tvary *tanzích* (lemma *tango*), *tiších* (lemma *ticho*) a *vlazích* (lemma *vlaho*), které zde nepopisujeme, neboť, ač teoreticky přípustné, se nevyskytují. Na internetu jsme nalezli jediný výskyt tvaru *tiších*, a to v básni Otokara Březiny.

Význam jednotlivých částí:

x je většinou kód předposledního znaku lemmatu, před koncovým *-o*:

- `r` pro zakončení *-ro*
- `k` pro zakončení *-ko*
- `v` pro zakončení *-vo*, je-li v 2pl vložené *e*
- `l` pro zakončení *-lo*, je-li v 2pl vložené *e*
- `m` pro zakončení *-mo*, je-li v 2pl vložené *e*
- `n` pro zakončení *-no*, je-li v 2pl vložené *e*
- `j` pro zakončení s měkkou souhláskou *-čo* nebo *-jo*
- `kum` pro zakončení *-kum*, je-li v 2pl vložené *e*
- `rum` pro zakončení *-rum*, je-li v 2pl vložené *e*
- `um` pro zakončení *-um*, není-li v 2pl vložené *e*
- `on` pro zakončení *-on*
- ϵ pro ostatní zakončení

6sg je kód zakončení tvaru v 6sg:

- `u` pro zakončení *-u*, FMU=u
- `e` pro zakončení *-e*, FMU=e
- `j` pro zakončení *-ě*, FMU=e
- `i` pro zakončení *-i*, FMU=i

2pl je kód zakončení tvaru v 2pl:

- `0` pro prázdné zakončení
- `e` pro vložené *-e*, FMU=E
- `i` pro zakončení *-í* (*stadium — stadií*)[3]

Příklady jsou v tabulce 9.11.

9.2.10 MOŘE

mrx

Všechna slova vzoru `moře` mají zakončení *-e* nebo *-tě*.

Pod tento vzor zahrnujeme i převzatá slova se zakončením *-e*, která jsou částečně nesklonná, např. *finále, promile, faksimile*. Někdy se vyskytují s českými koncovkami, jindy se neskloňují, jak ukazují příklady (111) a (112) z korpusu SYN. Mají tak v kombinacích 3sg, 6sg, 7sg, 2pl, 3pl, 6pl a 7pl dvě hodnoty kategorie `Mutace` — MUT=0 pro klasické zakončení vzoru `moře`, MUT=e pro zakončení *-e*.

byl s finále dvouhry spokojen (111)
budu spokojena s finálem (112)

[3] Tento parametr určuje i tvar v 7sg: *-i* místo běžného *-y*. Jde vlastně o kolísání mezi vzory `město` a `moře`.

Lemma	Kofix	Vzor	Současný pražský vzor
slovo	*slov*	`mt-ju-0`	`mt1x`
maso	*mas*	`mt-eu-0`	`mt1e`
kolo	*kol*	`mt-e-0`	`mt1e`
teplo	*tep*	`mt-euS`	`mt8e`
Náchodsko	*Náchods*	`mtk-uS`	`mts`
kakao	*kaka*	`mt-u-i`	`--`
zastupitelstvo	*zastupitelst*	`mtv-uj-e`	`mt4`
lečo	*leč*	`mtx-eu-0`	`mt1i`
vojsko	*vojs*	`mtk-u-0`	`mt7`
ložisko	*ložis*	`mtk-u-e0`	`mt7y0`
sluníčko	*sluníč*	`mtk-u-e`	`mt7e`
patro	*pat*	`mtr-eu-e`	`mt3e`
divadlo	*divad*	`mtl-e-e`	`mt8e`
pásmo	*pás*	`mtm-uj-e`	`--`
specifikum	*specifi*	`mtkum-u-0`	`mt12`
neutrum	*neut*	`mtrum-u-e`	`mt12r`
stadium	*stadi*	`mtum-u-i`	`mt12i`
album	*alb*	`mtum-u-0`	`mt12`
epiteton	*epitet*	`mton-u-0`	`--`
sympozion	*sympozi*	`mton-u-i`	`--`
vrata	*vrat*	`mt-0P`	`mtp`
vrátka	*vrát*	`mtk-eP`	`mtp7`
kamna	*kam*	`mtn-eP`	`--`
stehno	*steh*	`mtn-ju-e`	`mt9e`

Tabulka 9.11: Příklady lemmat ohýbaných podle vzoru `město`

Parametr **x** vzoru `moře` může mít tyto hodnoty:

- `t` pro zakončení *-tě*
- `e` pro zakončení *-e* — typ *finále* FMU=e
- ϵ pro ostatní zakončení

Příklady jsou v tabulce 9.12.

Lemma	Kofix	Vzor	Současný pražský vzor
moře	*moř*	`mr`	`mr1`
staveniště	*staveniš*	`mrt`	`mr10`
finále	*finál*	`mre`	`--`

Tabulka 9.12: Příklady lemmat ohýbaných podle vzoru `moře`

9.2.11 KUŘE

krx

Do skupiny slov ohýbaných podle vzoru `kuře` zařazujeme, stejně jako v pražském systému, i skupinu přejatých slov se zakončením *-a* (*revma, klima, kóma*). **x** může mít tyto hodnoty:

- `t` pro zakončení *-tě*
- `d` pro zakončení *-dě*
- `n` pro zakončení *-ně*
- `e` pro zakončení *-e*
- `j` pro zakončení *-ě*
- `a` pro zakončení *-a*

Příklady jsou v tabulce 9.13.

Lemma	Kofix	Vzor	Současný pražský vzor
kuře	*kuř*	`kre`	`kr1`
nemluvně	*nemluv*	`krn`	`kr4`
koště	*koš*	`krt`	`--`
mládě	*mlá*	`krd`	`--`
poupě	*poup*	`krj`	`kr1x`
dítě	*dít*	`krjS`	`--`
revma	*revma*	`kra`	`kr5`

Tabulka 9.13: Příklady lemmat ohýbaných podle vzoru `kuře`

9.2.12 STAVENÍ

st

Všechna substantiva ohýbající se podle vzoru `stavení` mají jednotný systém zakončení, není třeba žádných dalších parametrů, viz tabulka 9.14.

Lemma	Kofix	Vzor	Současný pražský vzor
stavení	*stavení*	`st`	`st`
listí	*listí*	`stS`	`st`

Tabulka 9.14: Příklady lemmat ohýbaných podle vzoru `stavení`

9.3 Životné vzory

Součástí všech vzorů podstatných jmen rodu mužského životného je i informace o tom, jestli se z nich dá vytvořit femininum. Tato informace se kóduje znakem F v derivační části vzoru. Za ním musí následovat způsob, jakým se ženský rod tvoří. To s sebou nese i informaci o jeho skloňování a tvoření přídavného jména přivlastňovacího, viz tabulka 9.15.

Zakončení feminina	Hodnota parametru	Ženský vzor
-ka	`Fa`	`znkeV`
-kyně	`Fy`	`nsn-i0-jV`
-ice	`Fi`	`ns-0-eV`
-ová	`Fo`	`NFy`[4]

Tabulka 9.15: Hodnoty parametru popisujícího možnosti tvoření feminina od životných substantivních vzorů

Tyto hodnoty pokrývají jen nejběžnější případy. Z mnoha dalších maskulin se sice ženský rod tvoří, ale dochází k alternaci v kmeni, což z praktického hlediska znamená, že se mění kofixy. Pro tyto případy by bylo třeba navýšit počet vzorů, což nám nepřipadá rozumné. Jednodušší je zachovat pouze tato nejběžnější tvoření a ostatní případy vyřešit pomocí odkazů ve slovníku (viz kap. 7). Jako příklad můžeme uvést slovo *tulák* s ženským protějškem *tulačka*. Stejné řešení, tedy slovotvorný odkaz ve slovníku, budou mít i méně častá tvoření, např. *žid* — *židovka*.

Z podstatného jména rodu mužského životného se tvoří pravidelně přivlastňovací přídavné jméno. Příznak V pro jeho tvoření, jak byl zaveden na str. 83, tedy není třeba explicitně zadávat. Přídavné jméno přivlastňovací se tvoří vždy.

Všechny mužské životné vzory mají společné kritické kombinace 1pl a 5pl:

- `i` pro zakončení *-i*, FMU=i
- `e` pro zakončení *-é*, FMU=e
- `v` pro zakončení *-ové*, FMU=v

Často jsou možné dvě mutace, občas i všechny tři. V obou kombinacích přichází vždy v úvahu stejná množina zakončení, tzn. že není možné, aby zakončení 1pl bylo odlišné od 5pl. Tato zakončení 1pl a 5pl jsou stejná pro všechny životné mužské vzory. Otázkou kodifikovanosti se ani zde nezabýváme.

Vzory se vytvářejí stejně jako dosud, tedy ze 2 znaků připomínajících školní vzory.

9.3.1 PÁN

pnx

Hodnoty parametru **x** jsou uvedeny v tabulce 9.16, ze které jsou též vidět klíčové pády, kvůli nimž se jednotlivé hodnoty musí rozlišovat. Stejná hlásková

[3] Adjektivní vzor, viz kap. 10

změna, která je uvedena ve sloupci 1pl, se projevuje i v 6pl v mutaci `i`, pokud připadá u konkrétního lemmatu v úvahu, např. *archeolog — archeolozích, vlk — vlcích.*

Parametr	1sg	5sg	1pl	
`k`	*-k*	*-ku*	*-ci/kové*	
`ek`	*-ek*	*-ku*	*-ci/kové*	
`nk`	*-něk*	*-ňku*	*-ňci/ňkové*	
`ik`	*-ík*	*-íku*	*-íci/íkové*	odvozené fem. *-ice*
`r`	*-r*	*-re*	*-ři/rové*	
`er`	*-er*	*-re*	*-eři/erové/ři/rové*	
`t`	*-r*	*-ře*	*-ři/rové*	
`h`	*-h*	*-hu*	*-zi/hové*	
`g`	*-g*	*-gu*	*-zi/gové*	
`c`	*-ch*	*-chu*	*-ši/chové*	
`es`	*-es*	*-e*	*-i/ové*	
`us`	*-us*	*-e*	*-i/ové*	
ϵ	ostatní zakončení			

Tabulka 9.16: Možná zakončení vzoru `pán`

V současném pražském systému vzorů se několik alternativ vzoru `pán` liší pouze v pořadí mutací 1pl (včetně jejich hodnocení z hlediska spisovnosti), což však není patrné z názvu vzoru. Z našeho vzoru je pořadí vidět v názvu vzoru, i když, jak už bylo zmíněno výše, mutace nikterak nehodnotíme.

Příklady jsou v tabulce 9.17.

Lemma	Kofix	Vzor	Současný pražský vzor
páv	*páv*	`pnivFi`	`pn1`
žid	*žid*	`pnie`	`pn5`
machr	*mach*	`pntiv`	`pn6vv`
autor	*auto*	`pnrFa`	`pn6f`
Šuster	*Šust*	`pnerFo`	`pn6`
zloduch	*zlodu*	`pnciv`	`pn22`
odborník	*odborn*	`pnikiFi`	`pn16ik`
jézéďák	*jézéďá*	`pnki`	`pn16`
daněk	*da*	`pnnkiv`	`-`
sok	*so*	`pnkivFy`	`pn18`
oslík	*osl*	`pnikivFi`	`pn16ik`
sládek	*slád*	`pnekivFo`	`pn17`
chirurg	*chirur*	`pngv`	`pn19`

Tabulka 9.17: Příklady lemmat ohýbaných podle vzoru `pán`

9.3.2 MUŽ

mzx

x může mít tyto hodnoty:

- `l` pro zakončení *-el*
- `c` pro zakončení *-ec*, odvozené femininum: *-kyně*
- ϵ pro ostatní zakončení

Příklady jsou v tabulce 9.18.

Lemma	Kofix	Vzor	Současný pražský vzor
plynař	*plynař*	`mziFa`	`mz1z`
vědec	*věd*	`mzciFy`	`mz5`
obyvatel	*obyvat*	`mzleFa`	`mz1e`
tchoř	*tchoř*	`mzi`	`mz1`
vítěz	*vítěz*	`mzviFa`	`mz1z`

Tabulka 9.18: Příklady lemmat ohýbaných podle vzoru `muž`

9.3.3 Kolísání mezi vzory PÁN a MUŽ

pmx

Stejně jako u neživotných vzorů `hrad` a `stroj`, i mezi měkkým a tvrdým vzorem životným není ostrá hranice. Existuje několik lemmat, jejichž skloňování mezi těmito dvěma vzory kolísá. Aby bylo zachováno Zlaté pravidlo morfologie, je třeba pro tuto kategorii vymezit samostatný vzor, který se postará o správné přiřazení hodnot kategorii `Flektivní mutace`. Kombinace kategorií jsou uvedeny v tabulce 9.19.

Pád	Zakončení (`pán`)	FMU	Zakončení (`muž`)	FMU	Příklad
2sg/4sg	*-a*	a	*-e*	e	*markýza/markýze*
3sg/6sg	*-u*	u	*-i*	i	*markýzu/markýzi*
5sg	*-e*	e	*-i*	i	*markýze/markýzi*
4pl	*-y*	y	*-e*	e	*markýzy/markýze*
6pl	*-ech*	e	*-ích*	i	*markýzech/markýzích*
7pl	*-y*	y	*-i*	i	*markýzy/markýzi*

Tabulka 9.19: Tabulka mutací při kolísání mezi vzory `pán` a `muž`

Příklad ukazuje tabulka 9.20.

Lemma	Kofix	Vzor	Současný pražský vzor
vyvrhel	*vyvrhel*	`pmv`	`pn3j mz1`

Tabulka 9.20: Příklad lemmatu s kolísavým skloňováním podle vzorů `pán` a `muž`

9.3.4 PŘEDSEDA

pdx

x může mít tyto hodnoty:

- `k` pro zakončení *-ka*
- `j` pro měkká zakončení
- ϵ pro ostatní zakončení

Příklady jsou v tabulce 9.21.

Lemma	Kofix	Vzor	Současný pražský vzor
obejda	*obejd*	`pdv`	`pd1`
předseda	*předsed*	`pdvFy`	`pd1`
Sašenka	*Sašenk*	`pdkv`	`pd5`
turista	*turist*	`pdieFa`	`pd2f`
Míša	*Míš*	`pdjv`	`pd1`

Tabulka 9.21: Příklady lemmat ohýbaných podle vzoru `předseda`

9.3.5 SOUDCE

sc

Všechna lemmata skloňovaná podle vzoru `soudce` mají zakončení *-ce*. Dokonce jsou u tohoto vzoru vždy možná zakončení 1pl *-ci* a *-cové*, stejně tak lze vždy tvořit ženský vzor pomocí zakončení *-kyně*. Pro jednotnost s ostatními životnými vzory ale tyto alternativy vždy vypisujeme. Vzor `soudce` tedy vypadá takto: `scivFy`.

Lemma	Kofix	Vzor	Současný pražský vzor
soudce	*soud*	`scivFy`	`sc1o`

Tabulka 9.22: Příklady lemmat ohýbaných podle vzoru `soudce`

9.4 Adjektivní vzory

Některá podstatná jména mají adjektivní skloňování. Často jde o vlastní jména osob, ale mohou to být i jména obcí (*Vřesová*), nebo apelativa (*kapesné, vrátný*). V takovém případě pro skloňování přebíráme vzory přídavných jmen, ovšem s poznámkou, že jde o jména podstatná. To se vyznačí pomocí prefixu N před vlastním adjektivním vzorem.

Kromě toho je třeba ještě dodat informaci o rodu, protože adjektivní vzory obecně vytvářejí všechny rody. Rod se ve vzoru zachycuje hned za úvodním prefixem N, a to pomocí běžných kódů pro rod v rámci systému značek. Není třeba zaznamenávat možnost tvoření negace, ani stupňování, protože u podstatných jmen to není relevantní. Ze zpodstatnělých přídavných jmen se také netvoří žádné automatické derivace.

Příklady adjektivně skloňovaných substantiv jsou uvedeny v kapitole o adjektivních vzorech, v oddíle 10.4.

10 Vzory přídavných jmen

Podle adjektivních vzorů se ohýbají především přídavná jména, ale využívají se i k vytváření slovních tvarů některých zájmen (např. *který*), číslovek (*první*) a podstatných jmen (*kapesné*). Pražský systém vzorů má pro každý slovní druh zvláštní vzor, ale to není třeba. Je pouze nutno do vzoru zaznamenat, jakého slovního druhu se týká, aby se správně přiřadila hodnota kategorie `Slovní druh`.

Názvy adjektivních vzorů tedy mají povinný prefix, označující slovní druh, jehož paradigma je vzorem vytvářeno. Prvním znakem adjektivního vzoru je tedy kód slovního druhu: N pro podstatná jména, C pro číslovky, P pro zájmena a A pro přídavná jména.

Následuje nepovinný prefix negace (viz 8.2.1.2).

Dále se ajektivní vzor skládá z následujících částí:

1. z části kódující základní skloňování ve všech pádech, číslech a rodech,
2. z části kódující stupňování,
3. z části kódující možné derivace.

Jednotlivé části vzoru je možno téměř libovolně kombinovat, přičemž každá část generuje jen určitou podmnožinu slovních tvarů paradigmatu, třetí část potom podává předpis, jak vytvořit odvozené lemma s pevně daným vzorem pro ohýbání. V případě některých derivací, kde se liší kofix pro derivace od kofixu pro ohýbání, je možno první dvě části vzoru vynechat. Např. jmenný tvar *mlád* se ohýbá podle vzoru `AJ`, kde `J` patří až do třetí části vzoru — viz 10.2.1.

První část můžeme chápat jako vlastní název vzoru. Aby byl název co nejintuitivnější, je jím v případě tvrdého skloňování (zkrácená) poslední slabika lemmatu. Podle ní se totiž řídí skloňování i případné stupňování. V případě skloňování měkkého stačí jeden znak (volíme i).

Další části vzoru můžeme chápat jako parametry, které rozšiřují nebo upravují vlastnosti základní části vzoru.

10.1 Skloňování a stupňování

10.1.1 Základní část vzoru — skloňování

Běžné dělení adjektivních paradigmat na tvrdá a měkká pro automatické zpracování nestačí, neboť u tvrdých přídavných jmen dochází v 1pl a při stupňování ke změně v kmeni. Klasický tvrdý vzor je tedy třeba rozepsat.

Tabulka 10.1 vyjmenovává základní části adjektivních vzorů. Druhý sloupec tabulky stanoví zakončení lemmatu, jehož se vzor týká. Další sloupce potom ukazují změny v zakončení slovních tvarů pro zvláštní hodnoty morfologických kategorií. Z tabulky je vidět, čím se jednotlivé vzory od sebe liší a proč jsou od sebe odlišeny. V posledních dvou řádcích tabulky jsou uvedeny vzory pro

přídavná jména přivlastňovací.
Hvězdička ∗ v posledních dvou sloupcích znamená, že se tvar řídí podle dalších částí vzoru, pomlčka – znamená, že vzor příslušné tvary nevytváří.

Vzor	Zakončení lemmatu	1pl živ.	Stupň.	Odvozené D
sky	*-ský*	*-ští*	*-štější*	*-sky*
cky	*-cký*	*-čtí*	*-čtější*	*-cky*
hy	*-hý*	*-zí*	*-ž-*	** (-ze/-ho)*
chy	*-chý*	*-ší*	*-š-*	** (-še/-cho)*
ry	*-rý*	*-ří*	** (-rší/-řejší)*	** (-ře/-ro)*
ky	*-ký*[1]	*-cí*	** (-čí/-ší)*	** (-ce/-ko)*
ny	*-ný*[2]	*-ní*	*-n-*	** (-ně/-no)*
y	*-ý*[3]	*-í*	*	*
yy[4]	*-ý*	–	*	*
yi[5]	–	*-í*	*	*
i[6]	*-í*	*-í*	*	*
y23[7]	–	–	*-ší*	*
uv	*-ův*	*-ovi*	–	–
in	*-in*	*-ini*	–	–

Tabulka 10.1: Tabulka adjektivních vzorů

Vzor y23 se použije v tom případě, kdy ve 2. a 3. stupni dochází ke změně v kmeni, a je tedy nutné použít jiný kofix. Podle tohoto vzoru se samozřejmě skloňují jen stupňovaná přídavná jména.

Každým z těchto vzorů se definuje skloňování 1. stupně (pozitivu) pro všechny rody a obě čísla. Vzory sky a cky díky své pravidelnosti umožňují i jednoznačně určit zakončení 2. stupně, totiž *-štější, -čtější*. Vzor y23 naopak slouží pouze pro stupňování, a to vždy se zakončením *-ší*. Tomu je třeba přizpůsobit i kofix. Je-li uveden vzor y23, je třeba ke kofixu přidat kód K (krátký) nebo D (dlouhý) kategorie Flektivní mutace. U ostatních vzorů to není třeba, protože typ je dán kódem pro stupňování, který uvedeme v následující části vzoru — viz oddíl 10.1.2.

Stupňují se jen ta přídavná jména, která to nemají „zakázáno“ — viz alternativa 0 jako pokračování vzoru.

Přídavná jména se zakončením na *-ný* mohou mít buď vzor ny, nebo y, jestliže netvoří jmenný tvar v maskulinu, nebo jejichž jmenný tvar nemá epentetické e (*plný — pln*). Samozřejmě je třeba vzoru přizpůsobit kofix.

Vzory y23, sky a cky nemohou pokračovat žádným z následujících způsobů, které popisují stupňování, neboť jejich stupňování je jednoznačné a vždy stejné.

[1] ale ne na *-ský* ani *-cký*
[2] jmenný tvar v maskulinu s epentetickým *e* (*schopný — schopen*)
[3] ostatní tvrdá, všechny tvary
[4] ostatní tvrdá kromě 1pl masc. živ.
[5] tvrdá v 1pl masc. živ.
[6] všechna měkká
[7] skloňování pouze 2. a 3. stupně

10.1.1.1 Pravidelné mutace ve skloňování přídavných jmen

U všech přídavných jmen dochází k mutacím v 7pl (-[*ýí*]*mi*, -[*ýí*]*ma*, u tvrdých navíc i -*ejma*).

U přídavných jmen tvrdých je pravidelná mutace -*ej* spisovné mutace -*ý*, a to ve všech slovních tvarech, kde je -*ý* přítomno ve spisovné formě, tedy nejen *nový — novej*, ale i *nových — novejch*, *novými — novejma*, apod.

Další systematické mutace -*í* a -*ý* jsou v 1pl rodu mužského životného, podobně -*é* a -*ý* (*velké/velký domy*).

Přehled systematických mutací adjektivního skloňování je uveden v tabulce 10.2 i s kódy.

Kód	Vysvětlení	Příklad
`y`	*ý* místo *é* nebo *í*	*velký (domy, páni), velkýho*
`k`	krátká mutace	*velkym, otcovym*
`a`	zakončení -*ma* v 7pl	*velkýma*
`j`	-*ej*- místo -*ý*-	*velkej, velkejch*
`ja`	příklad kombinace	*velkejma*

Tabulka 10.2: Systematické flektivní mutace v adjektivních vzorech

Přídavná jména přivlastňovací mají několik specifických mutací. Mimo jiné se zde vyskytují (nekodifikované) mutace -*ovo* a -*ic* pro libovolný rod, číslo i pád, viz příklady (113) až (119) z korpusu SYN.

Zemřel v Havlíčkovo Brodě (113)
šerifovo pomocníku vůbec nenapadlo (114)
Na jevišti nastoupil Dismanovo soubor (115)
táhnul to později s kostelníkovo ženou (116)
sleduje sousedovic slepice (117)
se sousedovic ohařem (118)
od sousedovic trávníku (119)

Přehled těchto i dalších systematických flektivních mutací specifických pro přídavná jména přivlastňovací obsahuje tabulka 10.3.

Kód	Vysvětlení	Příklad
`o`	nesklonná mutace *-ovo*	*Novákovo (domy)*
`c`	nesklonná mutace *-ovic*	*Novákovic (domy)*
`e`	*-ině/-ově*	*matčině, otcově*
`u`	*-inu/-ovu*	*matčinu, otcovu*
`k`	krátká mutace	*otcovym, otcuv*
`E`	vložené *-e-* před r	*Bergerův*
`R`	bez vloženého *-e-* před r	*Bergrův*
`Eja`	příklad kombinace	*Bergerovejma*

Tabulka 10.3: Systematické flektivní mutace vzorů přivlastňovacích přídavných jmen

Tabulka 10.4 na str. 106 uvádí příklad vzoru `Aky0`. Liché sloupce obsahují zakončení, sudé sloupce potom značku, která se skládá z pěti pozic. První pozice kóduje kategorii `Slovní druh`, který je všude stejný, totiž přídavné jméno. Další pozice kódují kategorie `Rod`, `Číslo`, `Pád` a `Stupeň`, který je opět všude stejný. Značka není úplná, protože v ní chybí zakódování kategorie `Negace`. Je zřejmé, že příslušné slovní tvary mohou být jak negativní (NEG=N), tak pozitivní (NEG=A). Není to tedy morfologická značka podle přesné definice, která vyžaduje přítomnost všech relevantních morfologických kategorií. Kód za pomlčkou u některých zakončení kóduje flektivní mutaci.

V tabulce 10.5 na str. 107 uvádíme druhou část definice vzoru `Akys`, která obsahuje předpis na stupňování. První část vzoru, která popisuje skloňování pozitivu, je vlastně popis vzoru `Aky0`, tedy tabulka 10.4.

10.1.2 Stupňování

Druhá část vzoru, která popisuje stupňování, používá následujících kódů:

- `e` pro zakončení *-ejší*, FMU=D
- `j` pro zakončení *-ější*, FMU=D
- `s` pro zakončení *-ší*, FMU=K
- `c` pro zakončení *-čí*, FMU=K
- `0` pro nestupňovatelné

V prvních dvou případech se všem tvarům přiřadí dlouhá flektivní mutace (FMU=D), ve druhých dvou případech krátká flektivní mutace (FMU=K), a to i tehdy, když se u konkrétního lemmatu mutace netvoří.

Výjimky jsou, jak již bylo uvedeno, vzory `y23`, který slouží pouze pro stupňování, a vzory `sky` a `cky`, kde je zakončení stupňování pravidelné, takže není třeba ho zvlášť specifikovat. Jediné přípustné možnosti jsou `sky0` a `cky0` pro zákaz stupňování. Ostatní základní vzory nemají zakončení stupňování jednoznačně dáno (např. *divoký — divočejší*, ale *lehký — lehčí*), tudíž je třeba ho specifikovat touto druhou částí vzoru.

10.1.2.1 Mutace ve stupňování přídavných jmen

Několik přídavných jmen má krátkou a dlouhou mutaci pro stupňování. Dlouhá mutace končí na *-[eě]jší*, krátká je bez *-[eě]j-*. Jako příklady mohou sloužit dvojice *snadnější — snazší*, *hrubější — hrubší*. Není jich mnoho, jde o uzavřenou skupinu. Kódování ukazuje tabulka 10.6.

ká	ANP41	kejch	AFP61-Fj	ký	ANS41-Fy	kých	ANP61
ká	AFS51	kejch	AMP61-Fj	ký	AMP41-Fy	kých	AMP61
ká	ANP51	kejch	AIP21-Fj	ký	AFP41-Fy	kých	AIP61
ká	ANP11	kejch	ANP61-Fj	ký	ANP11-Fy	kých	AMP21
ká	AFS11	kejch	AFP21-Fj	ký	AIS51	kých	AFP21
ké	ANS11	kejch	ANP21-Fj	ký	ANS11-Fy	kých	AFP61
ké	AFS31	kejch	AIP61-Fj	ký	AFS61-Fy	kých	AIP21
ké	AIP11	kejch	AMP21-Fj	ký	AFP11-Fy	kých	ANP21
ké	AIP41	kejm	ANP31-Fj	ký	ANS51-Fy	kým	AIS61-Fky
ké	AFS21	kejm	AMP31-Fj	ký	AIS41	kým	ANS71-Fky
ké	ANS51	kejm	AIP31-Fj	ký	ANP41-Fy	kým	AMS61-Fky
ké	AMP41	kejm	AFP31-Fj	ký	AIS11	kým	AMS71-Fky
ké	ANS41	kejma	AFD71-Fj	ký	ANP51-Fy	kým	AIS71-Fky
ké	AFP11	kejma	ANP71-Fj	ký	AMP11-Fy	kým	ANS61-Fky
ké	AFS61	kejma	AIP71-Fj	ký	AMP51-Fy	kým	AMP31
ké	AIP51	kejma	AMP71-Fj	ký	AFP51-Fy	kým	AIS61-Fy
ké	AFP41	kejma	AFP71-Fj	ký	AFS21-Fy	kým	AMS71
ké	AFP51	kém	ANS61	ký	AMS51	kým	AMS61-Fy
kého	AMS21	kém	AIS61	ký	AIP41-Fy	kým	AFP31
kého	ANS21	kém	AMS61	ký	AMS11	kým	ANS61-Fy
kého	AIS21	kému	AMS31	ký	AIP51-Fy	kým	AIP31
kého	AMS41	kému	AIS31	ký	AFS31-Fy	kým	ANP31
cí	AMP51	kému	ANS31	ký	AIP11-Fy	kým	AIS71
cí	AMP11	kou	AFS41	kýho	AIS21-Fy	kým	ANS71
kej	AIS41-Fj	kou	AFS71	kýho	AMS41-Fy	kýma	AFD71
kej	AMS11-Fj	kými	AFP71	kýho	AMS21-Fy	kýma	ANP71-Fy
kej	AIS11-Fj	kými	AMP71	kýho	ANS21-Fy	kýma	AMP71-Fy
kej	AIS51-Fj	kými	AIP71	kýmu	AMS31-Fy	kýma	AFP71-Fy
kej	AMS51-Fj	kými	ANP71	kýmu	AIS31-Fy	kýma	AIP71-Fy
				kýmu	ANS31-Fy		

Tabulka 10.4: Příklad vzoru `AkyO`

ší	ANP1[23s]	šim	AIP3[23s]-Fk	šíma	AFD7[23s]
ší	AIS1[23s]	šim	AMS7[23s]-Fk	šíma	AMP7[23s]-Fa
ší	AMS5[23s]	šim	ANP3[23s]-Fk	šíma	AIP7[23s]-Fa
ší	AFS7[23s]	šim	ANS6[23s]-Fk	šíma	ANP7[23s]-Fa
ší	AFP5[23s]	šim	ANS7[23s]-Fk	šíma	AFP7[23s]-Fa
ší	AIP5[23s]	šim	AIS6[23s]-Fk	šími	AFP7[23s]
ší	AIP1[23s]	šim	AIS7[23s]-Fk	šími	ANP7[23s]
ší	AFS3[23s]	šim	AMP3[23s]-Fk	šími	AIP7[23s]
ší	AMP4[23s]	šim	AMS6[23s]-Fk	šími	AMP7[23s]
ší	AFS1[23s]	šim	AFP3[23s]-Fk	šímu	ANS3[23s]
ší	AFS2[23s]	ším	AMS6[23s]	šímu	AIS3[23s]
ší	AMS1[23s]	ším	AMS7[23s]	šímu	AMS3[23s]
ší	AMP1[23s]	ším	AIS6[23s]	šího	AIS2[23s]
ší	AMP5[23s]	ším	AIS7[23s]	šího	ANS2[23s]
ší	AFS6[23s]	ším	AMP3[23s]	šího	AMS2[23s]
ší	AIS4[23s]	ším	ANS6[23s]	šího	AMS4[23s]
ší	AFS5[23s]	ším	ANS7[23s]	ších	ANP6[23s]
ší	ANS1[23s]	ším	ANP3[23s]	ších	AMP6[23s]
ší	AFP1[23s]	ším	AFP3[23s]	ších	AFP2[23s]
ší	ANS5[23s]	ším	AIP3[23s]	ších	AIP2[23s]
ší	ANP4[23s]			ších	AMP2[23s]
ší	AIP4[23s]			ších	ANP2[23s]
ší	AFP4[23s]			ších	AIP6[23s]
ší	ANS4[23s]			ších	AFP6[23s]
ší	AIS5[23s]				
ší	AFS4[23s]				
ší	ANP5[23s]				

Tabulka 10.5: Příklad: část vzoru `Akys`, který kóduje stupňování se zakončením *-ší* (např. *těžší*). Značky jsou zapsány pomocí regulárního výrazu — ve všech případech tedy jde o zakončení stupně `2`, `3` nebo `s`.

Kód	Vysvětlení	Příklad
FD	dlouhé stupňování -[eě]jší	*hrubější, trpčejší*
FK	krátké stupňování -ší/-čí	*hrubší, trpčí*

Tabulka 10.6: Systematické flektivní mutace ve stupňování přídavných jmen

Tyto mutace se pochopitelně kombinují s mutacemi ve skloňování, takže např. pro komparativ 7pl lemmatu *hrubý* dostáváme 4 variantní tvary: *hrubějšími, hrubšími, hrubějšíma, hrubšíma*, které mají po řadě hodnoty FMU=D, FMU=K, FMU=Da. FMU=Ka.

Existují ještě dvě lemmata s mutacemi ve stupňování, které však nezapadají do paradigmatického vzorce z předchozího odstavce, protože mutace nejsou v zakončení, nýbrž ve kmeni. Jsou to lemmata *bílý* a *svatý*. U obou dochází při stupňování k alternaci v kmeni, čímž vznikne mutace. Tyto mutace se týkají všech slovních tvarů 2. i 3. stupně, i nově zavedeného stupně S (typ *sebekrásnější*).

Máme totiž *svatější* i *světější*, jakož i *bílejší* i *bělejší*, i když *bílejší* je výrazně méně časté. Jelikož se nejedná o systematický jev, používáme v tomto případě pro odlišení mutací obyčejné číslování. Na konkrétní hodnotě nezáleží, podstatné je mutace pouze odlišit.

10.2 Derivace

Uvedeme zde nejběžnější derivace z přídavných jmen. Výjimky je třeba zadat do slovníku zvlášť. Uvedené derivace popisují, jakým způsobem se z přídavných jmen tvoří odvozená lemmata. U každého typu derivace tedy ještě uvádíme, jakým způsobem se odvozená lemmata ohýbají.

Do derivační části umisťujeme i pravidlo na tvoření jmenného tvaru, přestože to derivace není. Se skutečnými derivacemi má však společné to, že se netvoří vždy, ale jen od některých lemmat. O derivaci v pravém slova smyslu však nejde, jmenný tvar patří do paradigmatu dlouhého tvaru přídavného jména, které je jeho lemmatem.

10.2.1 Tvoření jmenného tvaru

J následováno buď

- ničím (*jistý* — *jist*), nebo
- e pro vložení epentetického -*e*- u mužského rodu (*schopný* — *schopen*), nebo
- - pro vypuštění mužského singuláru (např. adjektivum *snadný* netvoří jmenný rod pro rod mužský, ale ostatní rody možné jsou).

Ohýbání jmenných tvarů spočívá ve vytvoření tvarů pro 1. a 4. pád všech rodů a čísel (v případě J- bez 1sg mužského rodu), i když 4. pád lze většinou považovat za archaismy (zejména u mužského rodu). Důležité je, že všechna tato tvoření jsou pravidelná, a proto pro ně není třeba specifikovat zvláštní vzory.

Výjimečné tvoření jmenného tvaru se vyskytuje mj. u těchto adjektiv: *mlád, sláb, stár, zdráv*. Zde jde o změnu délky v kmeni, takže vždy bude třeba použít dvojí kofix a vzor:

- zdrav `Ayj` pro skloňování a stupňování adjektiva *zdravý*,
- zdráv `AJ` pro odvození jmenných tvarů *zdráv, zdráva, zdrávo, zdrávi, zdrávy, zdrávu*.

Vzory `Ay[ejsc0]Je`, `Ayy[ejsc0]Je`, ani `Ayi[ejsc0]Je` nemají smysl, jsou tedy nepřípustné. Epentetické *-e-* se totiž vkládá vždy před poslední souhlásku před zakončením *-ý*, takže vzor tuto souhlásku musí kódovat.

Měkká adjektiva jmenné tvary netvoří nikdy, vzory `AiJ[e-]?` tedy také nemají smysl.

10.2.2 Tvoření příslovce

Používáme znak D následovaný:

- `e` pro zakončení *-e*
- `j` pro zakončení *-ě*
- `u` pro zakončení *-u*[8]
- `y` pro zakončení *-y*
- `o` pro zakončení *-o*

Poznámka:
Je-li u vzoru `ky` derivace `Dy`, adverbium má zakončení *-ky*, je-li derivace `De`, zakončení je *-ce*.

Odvozená příslovce se mohou stupňovat v závislosti na typu přídavného jména, ze kterého byly odvozeny (viz též tabulka 10.1).

`sky, cky` ... stupňuje se pravidelně (*-štěji, -čtěji*),

`ky` ... stupňuje se pravidelně (*-čeji*), ale je hodně výjimek (*hluboký — hluboce — hlouběji, úzký — úzce — úžeji, krátce — kratčeji, řídce — řidčeji, vysoký — vysoko, vysoce — výše*) — u těch je třeba derivaci uvést zvlášť,

`Do` ... nestupňuje se

`De` ... stupňuje se *-ji* (*milý – mile – mileji*), tedy pravidelně

`Du` ... stupňuje se *-eji* (*pomalý – pomalu – pomaleji*), tedy pravidelně

`Dj` ... stupňuje se *-ji* (*čistý – čistě – čistěji*), tedy pravidelně

Je-li původní přídavné jméno nestupňovatelné (má ve vzoru 0), nestupňuje se ani příslovce.

[8] Zdá se, že se zakončením *-u* existuje jen jediné příslovce tvořené z přídavného jména, a to *pomalu*.

10.2.3 Tvoření podstatného jména na *-ost*

Takto derivované podstatné jméno se vždy skloňuje podle vzoru `kost`, stačí tedy možnost této derivace zaznamenat jediným znakem, a to `N`. U vzoru `ky` se vytváří se zakončením *-kost*.

Tvoření jiných typů podstatných jmen není tak pravidelné, nezavádíme pro ně proto žádné pravidlo.

Jako příklad může sloužit kofix *strnul* se vzorem `AyeN`, který vytvoří všechny tvary přídavného jména *strnulý*, s pravidelným stupňováním *strnulejší* a odvozeným podstatným jménem *strnulost*.

10.3 Příklady

V tomto oddíle přinášíme příklady přídavných jmen spolu s kofixy a vzory, pomocí kterých se tvoří celé paradigma. Jako příklady jsme vybrali téměř všechny adjektivní vzory uvedené v disertační práci Kláry Osolsobě (viz (Osolsobě, 1996)). Jsou uvedeny v tabulce 10.7. Z příkladů je patrna parametrizovatelnost adjektivních vzorů. Pro názornost jsou v posledním sloupci uvedeny 4 tvary odvozené z příslušného kofixu a vzoru. Jsou to po řadě: komparativ, jmenný tvar, derivované podstatné jméno a derivované příslovce. V případě, že se některý tvar netvoří, je místo něj uvedena pomlčka (–). Pomlčka tedy v tomto sloupci neslouží jako oddělovač.

Poznámky pod čarou na této stránce se vztahují k odkazům v tabulce 10.7 na str. 111.

[8] Jmenný tvar v rodě mužském *mrtev* pomocí vzorů nemůžeme popsat. Do slovníku je zadán explicitně se svou morfologickou značkou. Vzor jsme nevytvořili proto, že jde o výjimku (ještě se vyskytuje *polomrtev*). Většina jmenných tvarů s vloženým *-e-* v rodě mužském končí na *-en*, což je ve vzorech zachyceno. Další výjimky jsou ještě *blízek*, *vesel* a *dalek*, které ale zavedené vzory popsat mohou.

[9] Příslovce *těžko/těžce* se zde odvodit nedají kvůli jinému kofixu při stupňování (*tíže/tížeji*).

Lemma	Kofix	Vzor	
čilý	*čil*	AyeNDe	*čilejší – čilost čile*
kovový	*kovov*	AyODj	*– – – kovově*
mrtvý	*mrtv*	AyOJ-Dj	*– mrtva – mrtvě* [9]
jiný	*jin*	AyO	*– – – –*
křepký	*křep*	AkycNDe	*křepčí – křepkost křepce*
blizoučký	*blizouč*	AkyODoe	*– – – blizoučko/ce*
sladký	*slad*	AkysNDoe	*sladší – sladkost sladko/ce*
těžký	*těž*	AkysN	*těžší – těžkost –* [10]
daleký	*dale*	AkyOJNDoe	*– dalek dalekost daleko/ece*
řídký	*řídk*	AkyONDoe	*– – řídkost řídko/ce*
	řid FMU=K	Ay23	*řidší – – –*
hluboký	*hlubo*	AkyONDoe	*– – hlubokost hluboko/ce*
	hlub FMU=K	Ay23	*hlubší – – –*
strohý	*stro*	AhyONDe	*– – strohost stroze*
vetchý	*vet*	AchyeNDe	*vetšejší – vetchost vetše*
moudrý	*moud*	AryeNDe	*moudřejší – moudrost moudře*
bílý	*bíl*	AyeDe	*bílejší – – bíle*
	bělej FMU=1	Ay23	*bělejší – – –*
nový	*nov*	AyeNDj	*novější – novost nově*
pustý	*pust*	AyseJDjo	*pustší/pustější pust – pustě/o*
chudý	*chud*	AysjJNDjo	*chudší/chudější chud chudost chudě/o*
společenský	*společen*	AskyD	*společenštější – – společensky*
pražský	*praž*	AskyOD	*– – – pražsky*
otrocký	*otroc*	AckyD	*otročtější – – otrocky*
řecký	*řec*	AckyOD	*– – – řecky*
starý	*star*	Ayys	*starší – – –*
	stař	AyiDe	*– – – staře*, též *staří*
	stár	AJ	*– stár – –*
drahý	*dra*	AhysNDe	*dražší – drahost draze*
snadný	*snad*	AnyjJ-NDoj	*snadnější snadna snadnost snadno/ně*
	snaz FMU=K	Ay23	*snazší – – –*
cizí	*ciz*	AieNDe	*cizejší – cizost cize*
jarní	*jarn*	AijDj	*jarnější – – jarně*
zadní	*zad*	Aij	*zadnější – – –*
	zaz FMU=K	Ay23	*zazší – – –*
ostatní	*ostatn*	AiODj	*– – – ostatně*
vlčí	*vlč*	AiO	*– – – –*

Tabulka 10.7: Příklady adjektivních vzorů pro přídavná jména

10.4 Adjektivní skloňování dalších slovních druhů

Podle uvedených vzorů se skloňují ještě některá lemmata jiných slovních druhů, a to zájmena, číslovky a podstatná jména. Pro správné přiřazení kategorie `Slovní druh` je třeba před vlastní název vzoru přiřadit kód slovního druhu, a to P pro zájmena, C pro číslovky a N pro podstatná jména. Přídavná jména mají označení A.

V těchto případech používáme totožné vzory pro skloňování, nikoli však už část pro stupňování. Je-li specifikace rodu jiná než A (přídavné jméno), automaticky se nevytváří stupňování. Není tedy třeba explicitně ve vzoru uvádět kód `0` v části pro stupňování. Taktéž se nepoužívá derivační část vzoru.

Znak A pro skloňování přídavných jmen na začátku názvu vzoru by se opticky mohl plést s možným označením prefixu negace, ale vzhledem k tomu, že označení slovního druhu je na začátku názvu adjektivního vzoru povinné, k záměně dojít nemůže. Prefix negace následuje vždy až za označením slovního druhu, tedy kdybychom např. chtěli mít NEG=A pro celé paradigma lemmatu *nekalý*, přiřadíme mu vzor `AAy`. Tím zabráníme tvoření slov s dvojitým prefixem *nene-* (**nenekalý*).

U podstatných jmen je navíc třeba specifikovat rod. To učiníme připojením kódu rodu na konec vzoru. Příklady uvádí tabulka 10.8.

Lemma	Kofix	Vzor	Současný pražský vzor
který	*kte*	`Pry`	`rypif`
takový	*takov*	`Py`	`-`
třetí	*třet*	`Ci`	`iccr`
druhý	*dru*	`Chy`	`-`
sterý	*ste*	`Cry`	`-`
šéfová	*šéfov*	`NFy`	`mdf`
Vřesová	*Vřesov*	`NFyS`	`mdn`
kapesné	*kapesn*	`NNy`	`mdn`
vrátný	*vrátn*	`NMy`	`mdm`
vrátná	*vrátn*	`NFy`	`mdf`
cestující	*cestujíc*	`NMi`	`jnm`
cestující	*cestujíc*	`NFi`	`jnm`
sudí	*sud*	`NMi`	`jnm`

Tabulka 10.8: Příklady podstatných jmen s adjektnivními vzory

11 Vzory pro příslovce

Přestože příslovce patří mezi neohebné slovní druhy, popisujeme v rámci morfologie jejich stupňování a možnost negace.

Příslovce, která se nestupňují, vzor nemají. Ve slovníku jsou zachyceny se svojí morfologickou značkou. V naprosté většině se tato příslovce ani nedají negovat. Pokud by se přece vyskytlo nějaké příslovce, ke kterému existuje negace, a přitom bylo nestupňovatelné, je třeba jeho negativní tvar zachytit ve slovníku samostatně.

Podobně jako u stupňování přídavných jmen, i v případě příslovcí nechváme tvoření superlativu a stupně `s` na morfologických nástrojích. I zde platí, že možnost tvoření 2. stupně implikuje automaticky i možnost pravidelného tvoření ostatních dvou stupňů.

Kromě toho máme, podobně jako v pražském systému, ještě dva sdružené vzory, které umožňují popsat jedním vzorem stupňování příslovcí se zakončením *-sky* a *-cky*. Tato příslovce vznikají výhradně z přídavných jmen, v kompaktně zapsaném slovníku by se tedy samostatně neměla vyskytnout. Možnost stupňování těchto typů příslovcí závisí na možnosti stupňování původních přídavných jmen.

Pro příslovce stačí jediný vzor, dali jsme mu název `adv`:

adv$\underline{\mathbf{x}}$

Parametr **x** může nabývat těchto hodnot:

`s` → stupňování příslovcí se zakončením *-sky*
`c` → stupňování příslovcí se zakončením *-cky*
ϵ → pravidelné stupňování pomocí zakončení *-ji*
`1` → pro daný kofix se tvoří jen pozitiv
`23` → pro daný kofix se tvoří všechny stupně kromě pozitivu

Tabulka 11.1 ukazuje přehled vzorů pro příslovce.

Pravidelné mutace stupňovaných příslovcí

V tabulce 11.1 jsou též zachyceny hodnoty kategorie `Flektivní mutace` pro různá zakončení stupňovaných tvarů. Kromě zde uvedených hodnot má mnoho příslovcí ve druhém a třetím stupni dva tvary — kratší a delší, který se z kratšího tvoří přidáním *-e* na konec slovního tvaru (*blíž — blíže*), s mutací FMU=e pro tvar se zakončením *-e*.

U některých příslovcí dochází ke kombinacím, např. pro lemma *snadně* máme 5 mutací druhého stupně (a stejně i stupně třetího a stupně s): *snáz, snáze, snadněji, snadněj* a *snadnějc* s mutacemi po řadě: FMU=K, FMU=Ke, FMU=D, FMU=Dj, FMU=Dc.

Vzor	Zakončení	Stupeň (DEG)	Flektivní mutace (FMU)
adv	0	1	0
	-ji	[23s]	0
	-j	[23s]	j
	-jc	[23s]	c
advc	*-cky*	1	0
	-čtěji	[23s]	0
	-čtěj	[23s]	j
	-čtějc	[23s]	c
advs	*-sky*	1	0
	-štěji	[23s]	0
	-štěj	[23s]	j
	-štějc	[23s]	c

Tabulka 11.1: Přehled vzorů příslovcí

K některým příslovcím se zakončením *-e*/*-ě* existují příslovce, většinou predikativní (SUB=R), se stejným kmenem a zakončením *-o*. Často se dají i stupňovat, čímž dochází k homonymii mezi stupňovanými tvary obou příslovcí.

Příbuznost takových dvojic lze vyjádřit pomocí derivačních odkazů ve slovníku.

Příklady

Lemma	Kofix	Vzor	Tvary 2. stupně
rychle	*rychle*	`adv`	*rychleji*
složitě	*složitě*	`adv`	*složitěji*
pomalu	*pomalu*	`adv1`	
	pomale	`adv23`	*pomaleji*
rusky	*rusky*	`adv1`	
divoce	*divoce*	`adv1`	
	divoče	`adv23`	*divočeji*
divoko	*divoko*	`adv1`	
	divoče	`adv23`	*divočeji*
domácky	*domácky*	`advc`	*domáčtěji*

Existuje poměrně velká množina příslovcí s nepravidelným stupňováním. Tato příslovce nepopisujeme pomocí vzorů. Každý slovní tvar z jejich paradigmatu má ve slovníku svůj záznam s přesným určením hodnot svých relevantních kategorií. Několik příkladů ukazují tabulky 11.2 a 11.3.

Lemma	Slovní tvar, příp. kofix	FMU	Kategorie, příp. vzor
draze	*draze*		DEG=1
	dráž		DEG=2
	dráže	FMU=e	DEG=2
	nejdráž		DEG=3
	nejdráže	FMU=e	DEG=3
	sebedráž		DEG=s
	sebedráže	FMU=e	DEG=s
draho	*draho*		DEG=1
	dráž		DEG=2
	dráže	FMU=e	DEG=2
	nejdráž		DEG=3
	nejdráže	FMU=e	DEG=3
	sebedráž		DEG=s
	sebedráže	FMU=e	DEG=s

Tabulka 11.2: Příklad rozepsání paradigmat příslovcí *draho* a *draze* pomocí jednotlivých slovních tvarů a jejich vlastností

Lemma	Slovní tvar, příp. kofix	FMU	Kategorie, příp. vzor
snadno	*snadno*		DEG=1
	snadně	FMU=D	`adv23`
	snáz	FMU=K	DEG=2
	snáze	FMU=Ke	DEG=2
	nejsnáz	FMU=K	DEG=3
	nejsnáze	FMU=Ke	DEG=3
	sebesnáz	FMU=K	DEG=s
	sebesnáze	FMU=Ke	DEG=s
snadně	*snadně*		DEG=1
	snadně	FMU=D	`adv23`
	snáz	FMU=K	DEG=2
	snáze	FMU=Ke	DEG=2
	nejsnáz	FMU=K	DEG=3
	nejsnáze	FMU=Ke	DEG=3
	sebesnáz	FMU=K	DEG=s
	sebesnáze	FMU=Ke	DEG=s

Tabulka 11.3: Příklad rozepsání paradigmat příslovcí *snadno* a *snadně* pomocí kombinace vzoru a jednotlivých slovních tvarů s vlastnostmi. Vzor `adv23` popisuje dlouhou mutaci stupňování příslovce, tedy s druhým stupněm *snadněji*.

Mutace, které jsou součástí vzoru `adv23`, se připojí k mutaci uvedené u kofixu.

12 Slovesné vzory

Česká slovesa jsou zřejmě nejsložitějším slovním druhem z hlediska popisu.

U sloves dochází k alternaci v kmenech mnohem častěji než u ostatních slovních druhů. Přitom systém koncovek není zas tak rozsáhlý, viz např. (Osolsobě, 1996).

Máme tedy na jedné straně velkou otevřenou množinu různých kofixů, na straně druhé potom uzavřenou a poměrně malou množinu možných zakončení. Pro vytváření vzorů je výhodné si možná zakončení rozdělit do několika základních podmnožin popisujících konkrétní slovesné tvary, a tyto podmnožiny potom přiřazovat kofixům, se kterými tvoří smysluplné slovní tvary.

Podobným způsobem je systém českých sloves popsán v (Osolsobě, 1996). My však budeme podmnožiny definovat přímo v názvech slovesných vzorů, podobně jako dosud u jmenných vzorů.

Slovesný vzor začíná vždy kódem pro slovní druh sloveso, tedy znakem `V`. Dále se skládá ze dvou částí — flektivní a derivační. V případě, že kofix lze použít jen pro jednu z obou částí, druhá může chybět.

Inspirací pro návrh flektivní části slovesných vzorů byla práce Simeona Romportla — viz (Romportl, 1970), především jeho formální způsob kódování slovesných tvarů. Romportl se však zabývá slovesnými tvary obecně, to znamená i vícečlennými, my popisujeme jen samostatné slovní tvary. Používáme také jinou množinu kategorií, takže jsme jeho způsob popisu museli poměrně značně přetvořit.

K popisu základního časování vymezíme 6 množin tvarů, jejichž kofixy se mohou lišit. Jsou to:

1. imperativ
2. prézens
3. préteritum
4. infinitiv
5. přechodník
6. pasívum

Podle nich se potom vytvářejí slovní tvary odpovídající ostatním hodnotám relevantních morfologických kategorií.

Prvních 5 kategorií je pro každé lemma povinných. Kvůli přehlednosti je zapisujeme ke každému kofixu jako pětici hodnot. V případě, že daný kofix se některé kategorie netýká, má v pětici hodnotu - (pomlčka).

Uvedenou pětici může předcházet prefix `B`, který znamená, že se vzor pro imperativ a pro prézens použije též pro vytvoření imperativu a budoucího času pomocí předpony *po-* (tzv. determinovaná slovesa).

Další část vzoru se týká tvoření trpného rodu, který je kódován jinak, neboť se nemusí tvořit vždy. Má tak tvar podobný vzoru derivačnímu.

Za pevnou pětici se uvádí typ derivace podle slovních druhů následovaný hodnotou, která přesně specifikuje, jakým způsobem se z daného kofixu odvozené slovo vytváří a jaký je její flektivní vzor. Odvozeniny s nepravidelným ohýbáním se automaticky nevytvářejí.

Popíšeme teď jednotlivé množiny slovesného vzoru. U každé uvedeme i nejběžnější pravidelné mutace. Jejich celkový přehled je uveden také v tabulce 12.1.

POS	VRB	PER	NUM	Mutace	Jeji kód	Mutace	Jeji kód	Mutace	Jeji kód
				Příklad		Příklad		Příklad	
V	F			*ti, ci*	**i**	*ct*	**t**	*[íý][sz]t*	**y**
				dělati, říci		*říct,* ne *dělat*		*vízt, nýst,* ne *číst*	
V	I			krátce	**K**	dlouze	**D**		
				plav, osvědč		*plavej, osvědči*			
				po-	**p**	*ž*	**z**		
				poběž		*budiž,řekněmež*			
V	ITP			měkce	**m**	tvrdě	**t**		
				mraž/mražen		*mraz/mrazen*			
V	PBL			*ť*	**t**				
				vímeť,budeť,vědělit					
V	P	1	P	me	**0**	m	**K**		
				neseme		*nesem*			
V	P	1	S	*i*	**i**	*u*	**u**	*ím/ám*	**d**
				maži, kupuji		*mažu, kupuju*		*sedím, vím, mazám*	
				im	**k**	měkce	**m**	tvrdě	**t**
				sedím, vím		*mažu*		*mazám*	
V	P	3	P	*í*	**i**	*ou*	**u**		
				maží, kupují, sázejí		*mažou, kupujou*			
				j	**j**	*i*	**k**		
				sázej, dělaj		vědi			
V	L			*l*[aoiy]?	**l**	*nul*[aoiy]?	**n**	*nul*[aoiy]	**n**
				tiskl		*tisknul*		*usnula,* ne *usl*	

Tabulka 12.1: Přehled nejběžnějších mutací v časování sloves

12.1 Flektivní vzor

12.1.1 1. pozice — Imperativ

V tabulce 12.2 jsou uvedeny kódy první pozice flektivní části slovesného vzoru a množiny zakončení, které jsou jimi popsány.

Hodnota	Množina zakončení	Příklady kofixu
0	0, *-me, -te*	*zruš, suď, ohlas, ohlaš*
j	*-ej, -ejme, -ejte*	*děl*
w	*-i, -ěme, -ěte*	*tiskn*
e	*-i, -eme, -ete*	*pošl*

Tabulka 12.2: Tabulka vzorů pro rozkazovací způsob

Pravidelné mutace

Ve slovesném imperativu se vyskytuje poměrně značné množství typů mutací.

Jednak jsou to krátké a dlouhé tvary (končící na *-ej*), např. *plav — plavej*, *maž — mazej*. Pomocí délky lze rozlišovat i rozkazovací způsoby, jejichž delší mutace končí ve 2. osobě singuláru na *-i*: *osvědč — osvědči*, *přivab — přivábi*. Kratší mutace mají FMU=K, delší FMU=D.

Někdy se mutace imperativu liší v tvrdosti (FMU=t) — *rozmraz, zaraz* a měkkosti — *rozmraž, zaraž* s FMU=m.

Dalším případem je mutace v imperativu determinovaných sloves s předponou *po-*, které přiřazujeme FMU=p. Jde např. o tyto dvojice: *nes — pones*, *jdi — pojď*, *jeď — pojeď*, *běž — poběž*, *slyš — poslyš*, ale už ne *posekej*, neboť existuje lemma *posekat*, ani *popojeď* od lemmatu *popojet*.

Mutace archaická s FMU=z připojuje k imperativu *-ž* (např. *vymyslemež, dejž, chraňtež*).

12.1.2 2. pozice — Prézens

Druhá pozice flektivní části slovesného vzoru je popsána v tabulce 12.3.

Hodnota	Množina zakončení	Příklady kofixu
o	*-u, -eš, -e, -eme/em, -ete, -ou*	*kop*
i	*-u/i, -eš, -e, -eme/em, -ete, -í/ou*	*kryj, maž*
a	*-ám, -áš, -á, -áme, -áte, -ají/aj*	*děl*
s	*-ím/im, -íš, -í, -íme, -íte, -í/ej/ejí*	*pros, sáz*
p	*-ím/im, -íš, -í, -íme, -íte, -í/ěj/ějí*	*trp, vypráv*

Tabulka 12.3: Tabulka vzorů a zakončení tvarů přítomného času

Označení vzorů pro přítomný čas je odvozeno od 3. osoby množného čísla, kromě vzorů s a p, které je nutné rozlišovat pouze kvůli tvarům *prosejí, prosej*, *trpějí, trpěj*.

Pravidelné mutace

Z tabulky 12.3 jsou vidět pravidelné mutace, ke kterým v prézentu dochází. Nejběžnější je zřejmě mutace 1. osoby plurálu *-eme/-em*. Spisovná koncovka je *-me* (*bereme*), ale u všech sloves první, druhé a třetí třídy, které tvoří první osobu množného čísla pomocí *-eme*, má navíc hovorový tvar *-m* (*berem, kynem, kryjem*). Spisovnou koncovku mutací neoznačujeme, koncovka *-m* má FMU=K (kratší).

V obecné češtině také často dochází ke krácení samohlásky. Nejčastěji jde o 1. osobu jednotného čísla (*musím — musim, platím — platim*, apod.), nebo 3. osobu čísla množného (např. *vědí — vědi*). Rozlišujeme zde hodnotu kategorie `Flektivní mutace` krátkou (FMU=k) a dlouhou (FMU=d).

Mutace 3. osoby plurálu mohou být dvojího typu: -[*aeě*]*jí*/-[*aeě*]*j* a -*í*/-*ou*. U prvního typu považujeme spisovnou mutaci s -*jí* za nulovou, obecně česká mutace -*j* má hodnotu FMU=j (*krájejí — krájej*). U druhého typu jde o slovesa 1. třídy (*mažou — maží, pečou — pečí*) a 3. třídy (*kupují — kupujou, kryjí — kryjou*) s FMU=i resp. FMU=u. Např. lemma *sázet* má v 3. osobě množného čísla tři mutace: *sázejí* (FMU=0), *sázej* (FMU=j) a *sází* (FMU=i). Podobné jsou i mutace 1. osoby singuláru -*i*/-*u* u sloves 1. a 3. třídy (*mažu — maži, kupuji — kupuju*), mají proto stejné hodnoty kategorie `Flektivní mutace`.

Časté je také kolísání časování přítomného času mezi pátou a první třídou (*kopu — kopám*). Mutace zde rozlišujeme podle krátké / dlouhé samohlásky. Zakončení z množiny `o` mají FMU=k, zakončení z množiny `a` FMU=d.

Zde se mohou navíc projevovat následující kmenové změny, které přidávají další hodnotu do kategorie `Flektivní mutace`:
š-s (*češu/češi — česám*)
ž-z (*mažu/maži — mazám*)
č-k (*skáču/skáči — skákám*).
Dohromady s mutací -*i*/-*u* sloves 1. třídy mají v 1. osobě jednotného čísla trojí hodnotu FMU: Příklady nalevo mají FMU=kmu/kmi (mutace i/u, krátká, měkká), napravo FMU=dt (mutace dlouhá, tvrdá).

Ještě zmíníme pravidelné archaické mutace s koncovým -*ť*, které se vyskytují i u minulého příčestí (např. *udělalť, udělámť*). V pražském systému se s nimi počítá, mají speciální hodnotu na 2. pozici. Přiřazujeme jim FMU=t.

12.1.3 3. pozice — Préteritum

Minulý čas je velmi pravidelný, má dvě možnosti tvoření. První je jednoduchá, druhá umožňuje stejnému kofixu přiřadit dvě různé množiny zakončení, jak je naznačeno v tabulce 12.4.

Hodnota	Množina zakončení	Příklady kofixu
`l`	*-l, -la, -lo, -li, -ly, -la*	*nes*
`n`	*-l, -la, -lo, -li, -ly, -la* *-nul, -nula, -nulo, -nuli, -nuly, -nula*	*usch*
`m`	*-la, -lo, -li, -ly, -la* *-nul, -nula, -nulo, -nuli, -nuly, -nula*	*us*

Tabulka 12.4: Množina zakončení tvarů minulého příčestí

Hodnota `m` se od `n` liší tím, že neobsahuje tvoření mužského singuláru.

Pravidelné mutace příčestí minulého činného

Jde o mutace se zakončením -*l* — -*nul* u sloves *trhl — trhnul*. U některých sloves tato mutace nepřipadá v úvahu u mužského rodu, u ostatních rodů však

ano (**usl* — *usnul*, ale *usla* — *usnula*). Tento rozdíl vyjadřuje dvojí hodnota parametru ve vzoru — viz hodnoty `m` a `n` v tabulce 12.4. Tvary s *-n-* mají FMU=n, stejně jako u podobného případu v infinitivu.

Další mutace se projevují u příčestí minulého zakončeného na Kl, kde K je souhláska. Zde se vytvářejí nespisovné mutace bez koncového *-l*, ovšem jen u mužského rodu (*vedl — ved, pletl — plet, nesl — nes, zábl — záb, kopl — kop, vezl — vez, klekl — klek, pomohl — pomoh, všiml — všim*). Mutace bez koncového *-l* značíme jako kratší — FMU=K.

12.1.4 4. pozice — Infinitiv

I infinitivy jsou velmi pravidelné. Množinu zakončení uvádí tabulka 12.5. Mutace infinitivu *odříci — odříknout* je třeba vyjadřovat jako dvě položky s různými kofixy a různým zakončením. Označení mutace se přidává ke kofixu.

Hodnota	Množina zakončení	Příklady kofixu
`t`	*-t, -ti*	*kopa*
`c`	*-ci, -ct*	*mo*

Tabulka 12.5: Zakončení tvarů infinitivu

Pravidelné mutace slovesného infinitivu

Všechny infinitivy mají dvě mutace: *-t* s FMU=0 a *-ti* s FMU=i (*být — býti*), nebo *-ci* s FMU=i a *-ct* s FMU=t (*moci — moct*). Tento typ mutace je tedy přítomen vždy i u následujících typů a kombinuje se s nimi.

Další mutace infinitivu se týkají jen několika málo sloves, např. *hanit — hanět, myslit — myslet, bydlit — bydlet*. U těchto sloves se projevuje stejná flektivní mutace ještě v příčestí minulém (*myslil — myslel*). Označujeme ji FMU=y resp. FMU=e.

Podobné jsou mutace *-ést — -íst*, *-ézt — -ízt* a *-éct — -[íý]ct*, např. u sloves *vést — vísť*, *vézt — vízt* a *péct — pícť*, případně *téct — týct*. Tyto mutace se však neprojevují v minulém čase. Mají stejné označení, tedy FMU=y resp. FMU=e. U těchto sloves zřejmě nedochází ke kombinované mutaci FMU=iy (**pícti*).

U mutací typu *říci — říknout, začít — začnout* mají tvary s *-n-* FMU=n. Tato mutace se také projevuje v minulém příčestí.

12.1.5 Přechodník `P`

Přechodníky, ač užívané dnes již zřídka, je třeba také umět rozpoznat, výjimečně i tvořit. Tabulka 12.6 uvádí množiny jejich zakončení.

První tři řádky tabulky popisují přechodník přítomný, poslední řádek pak přechodník minulý. Není třeba uvádět do vzoru, o jaký typ přechodníku se jedná. Jde o hodnotu kategorie `Slovesný tvar`, která se správně přiřadí příslušným vzorem, tedy VRB=p pro vzory `a, e, w`, VRB=m pro vzor `v`.

Přechodníky se pravidelně využívají k tvoření přídavného jména slovesného se zakončením *-cí* od přechodníku přítomného nebo *-vší* od přechodníku minulého. Oba tyto typy přídavných jmen se skloňují podle měkkého adjektivního

Hodnota	Množiny zakončení	Příklady kofixu
a	*-a, -ouc, -ouce*	*drhn*
e	*-e, -íc, -íce*	*vyprávěj*
w	*-ě, -íc, -íce*	*hromad*
v	*-v, -vši, -vše*	*nahromadi*

Tabulka 12.6: Vzory pro vytváření přechodníku

vzoru Ai0, tedy bez možnosti stupňování. Tvoření přechodníku tak automaticky implikuje tvoření těchto přídavných jmen.

12.1.6 Trpný rod T

Hodnoty kódů derivačních vzorů pro trpný rod ukazuje tabulka 12.7.

Hodnota	Množina zakončení	Příklady kofixu
Tn	*-n, -na, -no, -ny, -ni, -nu*	*nese*
Ta	*-án, -ána, -áno, -ány, -áni, -ánu*	*ps*
Tt	*-t, -ta, -to, -ty, -ti, -tu*	*táhnu*

Tabulka 12.7: Derivační vzory pro trpný rod

Hodnoty n a a jsou rozlišeny kvůli jednoduchému vytvoření lemmatu. Lemmatem přídavných jmen slovesných totiž není sloveso, ale přídavné jméno (viz kap. 10.2.1), konkrétně jeho dlouhý tvar. Takže např.
$\lambda(nesen) = nesený$,
$\lambda(táhnut) = táhnutý$,
$\lambda(psán) = psaný$ (změna v délce samohlásky v kmeni).

Mutace deverbativ

Některá slovesa tvoří dvojí trpný rod, bez měkčení kmenové souhlásky (*mrazen*) a s měkčením (*mražen*). Od obou lze odvozovat další deverbativa — podstatná jména (*mrazení* a *mražení*), přídavná jména (*mrazený*, *mražený*) a případně i příslovce. Tyto mutace se týkají i imperativu (*mraz* — *mraž*) (viz též 12.1.1). Přiřazujeme jim tyto hodnoty kategorie Flektivní mutace: FMU=m pro změkčené varianty, FMU=t pro nezměkčené.

12.2 Derivační vzory

Pražský systém slovesných vzorů, který vytvořil Jan Hajič, má velmi bohatý repertoár odvozování příbuzných deverbativ pomocí přípon. Náš systém by měl být schopen odvodit veškeré smysluplné derivace, které odvozuje současný pražský systém. Podívejme se tedy nejprve na typy lemmat, které jsou v pražském systému zachyceny. Lze je přehledně zpracovat do tabulky 12.8.

Tabulka má dva hlavní sloupce: první popisuje zakončení tvaru základního slovesa a jeho odvozenin, druhý zakončení tvaru slovesa iterativního a jeho odvozenin.

Základní		Iterativní	
-t		*-[íá]vat* `I`	
-cí `Ac` → Ai0		*-[íá]vací* → Ai0	
-cí/ší automaticky z přechodníku → Ai0		*-[íá]vající* → Ai0	
-ní/tí `N[nt]` → st		*-[íá]vání* → st	
-ný/tý/lý `A[ntl][j0]` → Ay[j0]	*-telný* `Ae` → Ayj	*-[ái]vaný* → Ay[j0]	*-[ái]vatelný* → Ayj
-ně/tě/le `D[ntl][j0]` → `adv1?`	*-telně* `De` → adv	*-[ái]vaně* → adv1?	*-[ái]vatelně* → adv
-nost/tost/lost `O[ntl]` → kt	*-telnost* `Oe` → kt	*-[ái]vanost* → kt	*-[ái]vatelnost* → kt

Tabulka 12.8: Pravidelné slovesné derivace popsané v pražském systému vzorů

Ve spodní části buněk jsou kódy, které mají následující význam. Kód před šipkou je součástí derivační části slovesného vzoru v případě, že se příslušné deverbativum tvoří. Budeme mu říkat derivační pravidlo. Proč není toto pravidlo uvedeno ve všech buňkách, vysvětlíme vzápětí. Kód za šipkou je kód již zavedeného vzoru, podle kterého se derivované lemma ohýbá.

Tak např. „`Ac` → `Ai0`“ znamená, že pomocí derivačního kódu (pravidla) `Ac` a příslušného kofixu se vytvoří odvozené lemma se zakončením *-cí* a bude se ohýbat podle adjektivního vzoru `Ai0` (měkké skloňování bez možnosti stupňování).

Třetí řádek obsahuje přídavná jména vytvořená pravidelně z přechodníků, proto zde žádné derivační pravidlo není třeba.

Alternativa `[ntl]` vyjadřuje trojí možné zakončení deverbativ, alternativa `[j0]` potom možnost nebo nemožnost stupňování odvozeného přídavného jména nebo příslovce. Pouze u přídavných jmen se zakončením lemmatu *-telný* volbu stupňování nepřipouštíme, neboť se domníváme, že zde je stupňování možné vždy.

Celý druhý sloupec se týká jen iterativních sloves, popisuje tedy jen lemmata typu *kupovávat, lehávat, končívat.*

Druhý sloupec obsahuje derivační pravidlo (`I`) pouze v první řádce. Vyjadřuje možnost tvoření iterativa od slovesa základního (z levého sloupce). Vzory, podle kterých se odvozené iterativní sloveso časuje, jsou uvedeny dále v oddíle 12.2.4.

Ostatní buňky pravého sloupce jsou pravidelně tvořené z tvaru iterativa, neobsahují tedy derivační pravidla (kódy před šipkou). Domníváme se totiž, že není třeba ve vzorech udržovat explicitně toto velké množství odvozenin od iterativních sloves a příslušná derivační pravidla zahrnujeme přímo do pravidla `I`.

Jestliže tedy existuje k slovesu příbuzné sloveso iterativní, pravidelně se z něj mohou tvořit všechny tvary naznačené ve druhém sloupci. Uvědomujeme si, že i zde může docházet ke generování velmi nepravděpodobných tvarů. Po několika sondách do korpusů i na internetu jsme však zjistili, že by nebylo rozumné derivační pravidla od iterativních sloves omezovat.

V následujících oddílech podrobně rozebereme jednotlivé buňky tabulky 12.8 a uvedeme příklady.

12.2.1 Přídavná jména slovesná `A`

Hodnoty kódů derivačních vzorů pro přídavná jména slovesná ukazuje tabulka 12.10.

Hodnota	Zakončení lemmatu	Vzor	Příklady kofixu
`An`	*-ný*	`AyO`	*unese*
`At`	*-tý*	`AyO`	*kopnu*
`Al`	*-lý*	`AyO`	*zemře*
`Ae`	*-telný*	`AyO`	*snesi*
`Ac`	*-cí*	`AiO`	*pozměňova*

Tabulka 12.9: Derivační vzory pro přídavná jména slovesná

Všechna přídavná jména slovesná odvozená podle uvedených vzorů mají automaticky přiřazenou hodnotu poddruhu SUB=S (deverbativní).

Kromě právě popsaného odvození přídavného jména slovesného se ještě tvoří přídavná jména od přechodníků. Pro ně není třeba zavádět žádné vzory, neboť se tvoří pravidelně, viz 12.1.5. Skloňují se podle vzoru `AiO` a mají poddruh SUB=G (od přechodníku přítomného), nebo SUB=M (od přechodníku minulého).

12.2.2 Deverbativní příslovce `D`

Hodnota	Zakončení lemmatu	Vzor	Příklady kofixu
`Dn`	*-ně*	`adv`	*unese*
`Dt`	*-tě*	`adv`	*rozvinu*
`Dl`	*-le*	`adv`	*ochraptě*
`De`	*-telně*	`adv`	*snesi*

Tabulka 12.10: Derivační vzory pro deverbativní příslovce

Příslovce lze tvořit od většiny přídavných jmen slovesných, ale ne vždy. Protipříkladem je přídavné jméno *kopnutý* s neexistujícím příslovcem **kopnutě*. Ani přídavná jména se zakončením *-ný* nemají automaticky příslovce — *zahájený*, ale **zahájeně*. Na druhou stranu ale, když příslovce existuje, lze ho odvodit pravidelně vždy záměnou koncové dlouhé samohlásky *-ý* za *-ě* v případě *-ný, -tý* a za *-e* v případě *-lý* (záviset — závislý — *závisle*). Ze zakončení *-cí* se příslovce nevytvářejí.

12.2.3 Podstatná jména slovesná N/O

Podstatná jména slovesná jsou dvojího druhu:

1. se zakončením *-í*, kód `N`, skloňování podle vzoru `st`,
2. se zakončením *-ost*, kód `O`, skloňování podle vzoru `kt`.

Podstatná jména slovesná se zakončením *-í* se odvozují z tvarů trpného rodu prostým připojením zakončení *-í* v případě hodnot `n` a `t` (*nesení, táhnutí*), nebo zakončení *-aní* pro hodnotu `a` (*psán — psaní*). Mají poddruh SUB=S (deverbativní). Poddruh druhého typu, se zakončením *-ost*, není deverbativní (SUB=0). Tabulka 12.11 rozepisuje jednotlivé možnosti podle kofixu.

Hodnota	Zakončení lemmatu	Vzor	Příklady kofixu
`Nn`	*-ní*	`st`	*unese*
`Nt`	*-tí*	`st`	*kopnu*
`On`	*-nost*	`kt`	*připojiště*
`Ot`	*-tost*	`kt`	*netknu*
`Ol`	*-lost*	`kt`	*pobloudi*
`Ne`	*-telnost*	`kt`	*nedotknu*

Tabulka 12.11: Derivační vzory pro podstatná jména slovesná

Derivace trpného rodu, slovesného podstatného jména, přídavného jména a příslovce se může zaznamenat do vzoru najednou, jestliže mají stejný kofix a jestliže současně následný parametr, tedy `n`, `t` nebo `l`, je shodný. Pro vytváření deverbativ jsou tedy teoreticky možné např. tyto kombinace (na pořadí nezáleží): `AN, AD, AT, ONT, DN, ADN, ADT, ADTN`, následované (alespoň) jedním z parametrů `n`, `t`, `l`.

12.2.4 Iterativní sloveso

Časování odvozeného iterativního slovesa, tedy derivační pravidlo `I` ukazuje tabulka 12.12.

Konec kofixu	Vzor	Příklad
-áv *-ív*	`ja---` (imperativ, prézens)	*kupovávej, kupovávám* *chodívej, chodívám*
-áva *-íva*	`--lt-` (préteritum, infinitiv)	*kupovával, kupovávat* *chodíval, chodívat*
-ávaj *-ívaj*	`----e` (přechodník přítomný)	*kupovávaje* *chodívaje*

Tabulka 12.12: Časování odvozeného iterativního slovesa

Pro odvozeniny z iterativních sloves se použijí stejná pravidla jako pro odvozeniny uvedené výše.

12.3 Sdružené slovesné vzory

Flektivní část vzoru je poměrně složitá proto, že v řadě sloves dochází ke změnám ve kmeni. Kromě toho však existuje velké množství sloves, která jsou

naopak velmi pravidelná a uvedený popis flexe, ač obecný, je pro ně zbytečně složitý. Ke stejnému závěru dospěl i Hajič, když pro ně ve své disertační práci (Hajič, 1994) zavedl sdružené vzory, které i úspěšně implementoval.

Jeho systém sdružených vzorů tak můžeme převzít, ovšem pouze jeho flektivní část. Pravidelné derivace, které jsou ke vzorům v pražském systému napevno připojeny, nevyužijeme. Místo toho použijeme pro derivace systém derivačních vzorů uvedených v oddíle 12.2, zejména pak v tabulce 12.8. Tato tabulka byla sestavena právě na základě derivačních částí sdružených vzorů. Pomocí nového systému si z ní však můžeme vybírat jen to, co se k danému kofixu hodí.

V tabulce 12.13 uvádíme seznam sdružených flektivních vzorů, které přebíráme z pražského Hajičova systému slovesných vzorů. U každého sdruženého vzoru jsou uvedeny kofixy a nově zavedené flektivní vzory, jejichž spojením je možno vygenerovat celé paradigma.

Do tabulky jsme nezahrnuli vidovou dvojici sdružených vzorů `itxd` a `itxn`, protože tyto vzory se liší od ostatních tím, že neobsahují předpis pro tvoření přechodníku přítomného, imperativu a trpného rodu. Umožňují tak jejich nepravidelné tvoření. V našem systému vzorů to nepotřebujeme.

V posledním sloupci tabulky je pro každý vid uveden příklad s kofixem odděleným pomlčkou od zakončení lemmatu.

Ve vzorech `noutd` a `noutn` je zahrnuta varianta ve tvoření minulého příčestí, a to uvedením hodnoty kategorie `Flektivní mutace` u kofixu.

Ke sdruženým vzorům je možno přidávat derivační vzory stejně jako ke vzorům jednoduchým, a tím vybrat pro konkrétní kofix jen ty derivace, které skutečně existují. Tím se můžeme vyhnout tvoření derivací jako např. **ženěně, *ženěnost, *ženívací, *ženívanost*, a dalších, které se derivují zcela pravidelně z pražského vzoru `nitn`.

Lemma	Nedokonavý vzor		Dokonavý vzor		Příklad
	Kofix	Vzor	Kofix	Vzor	
Xat	atn		atd		*děl-at / uděl-at*
	X	ja---	*X*	ja---	
	Xa	--lt-	*Xa*	--ltv	
	Xaj	----e	—	—	
Xovat	ovatn		ovatd		*děk-ovat / poděk-ovat*
	Xuj	0i--e	*Xuj*	0i---	
	Xova	--lt-	*Xova*	--ltv	
Xet	etn		etd		*kráj-et / nakráj-et*
	X	js---	*X*	js---	
	Xe	--lt-	*Xe*	--ltv	
	Xej	----e	—	—	
Xět	wtn		wtd		*reziv-ět / zreziv-ět*
	Xě	j-lt-	*Xě*	j-ltv	
	X	-p---	*X*	-p---	
	Xěj	----e	—	—	
Xdit	ditn		ditd		*chla-dit / ochla-dit*
	Xď	0----	*Xď*	0----	
	Xd	-p--w	*Xd*	-p---	
	Xdi	--lt-	*Xdi*	--ltv	
Xtit	titn		titd		*čtvr-tit / rozčtvr-tit*
	Xť	0----	*Xť*	0----	
	Xt	-p--w	*Xt*	-p---	
	Xti	--lt-	*Xti*	--ltv	
Xnit	nitn		nitd		*že-nit / ože-nit*
	Xň	0----	*Xň*	0----	
	Xn	-p--w	*Xn*	-p---	
	Xni	--lt-	*Xni*	--ltv	
Xit	iten		ited		*zuř-it / rozzuř-it*
	X	0s--e	*X*	0s---	
	Xi	--lt-	*Xi*	--ltv	
Xit	itin		itid		*barv-it / obarv-it*
	X	wp--w	*X*	wp---	
	Xi	--lt-	*Xi*	--ltv	
Xit	it0n		it0d		*křiv-it / zkřiv-it*
	X	0p--w	*X*	0p---	
	Xi	--lt-	*Xi*	--ltv	
Xnout	noutn		noutd		*sch-nout / usch-nout*
	Xn	wo--a	*Xn*	wo---	
	X	--n--	*X*	--n--	
	Xnou	---t-	*Xnou*	---t-	

Tabulka 12.13: Přepis pražských sdružených slovesných vzorů pomocí nových flektivních vzorů. Znak X v tabulce označuje kofix, ke kterému se sdružené vzory vztahují.

12.3.1 Příklady

Tabulka 12.14 ukazuje několik příkladů, jak se používají slovesné vzory pro popis flexe i derivací. Příklad lemmatu *zavřít* ukazuje, jak se kombinují vzory a konkrétní morfologické značky v záznamu jednoho lemmatu. Poslední příklad ukazuje řešení vícenásobného lemmatu.

Lemma	Kofix/Tvar	Vzor/Značka	Odvozená lemmata
zhasnout	*zhas*	`Vnoutd`	celé paradigma
	zhasnu	`VTANtAe`	*zhasnut, zhasnutý, zhasnutá, zhasnutelný*
hýbat	*hýb*	`Vatn`	celé paradigma
	hýba	`VAcAONe`	*hýbací, hýbatelný, hýbatelně, hýbatelnost*
	hýbá	`VNnI`	*hýbání, hýbávat*
zavřít	*zavř*	`Viu---`	imperativ, prézens
	zavře	`V--1-vTADNnNo`	*zavřel, zavřev, zavřen/ný/ně/ní/nost*
	zavří	`V---t-`	infinitiv
	zavřímo FMU=d	`AO-D-NS1----TJ--`	flektivní dlouhá mutace jmenného tvaru
	zavru FMU=t	`VO-D--S--1------`	flektivní tvrdá mutace 1.os. sg
	zavrou FMU=t	`VO-D--P--3------`	flektivní tvrdá mutace 3.os. pl
{ukrást, ukradnout}	*ukrad*	`Vnoutd`	celé paradigma
	ukrade	`VTANn`	*ukraden/ný/ná*
	ukrás FMU=d	`V---t-`	*ukrást* (mutace s dlouhou samohláskou)
	ukradnu FMU=n	`VTANt`	*ukradnut/tý/tá* (flekt. mutace s *n*)

Tabulka 12.14: Příklady slovesných vzorů

13 Vzory zájmen a číslovek

Číslovky a zájmena se většinou mohou skloňovat podle adjektivních vzorů, proto jsme se o nich zmínili již v kapitole 10, oddíle 10.4. Na tomto místě jen připomeneme, že se používá tvrdý i měkký adjektivní vzor bez možnosti stupňování. Před adjektivní vzor se uvádí kód slovního druhu, tedy C pro číslovky a P pro zájmena.

13.1 Číslovky

13.1.1 Číslovky základní

Skloňování číslovek základních je většinou nepravidelné. Zavádíme vzory jen pro lemmata se zakončením *-t* (*jedenáct* až *dvacet*, *třicet*, atd., a tvary typu *dvaapadesát*).

Vezmeme-li v úvahu i nekodifikované tvary typu *jedenácte, padesáte* a typu *dvacíti*, můsíme použít dva vzory. Nazveme je podle nejmenší číslovky, která se podle vzoru skloňuje, `C11` a `C20`.

Vzory uvádíme jako množinu trojic ⟨zakončení, morf. charakteristika, FMU⟩. Morfologická charakteristika zahrnuje kvůli přehlednosti místo celé morfologické značky jen hodnoty kategorií `Pád` a `Číslo`.

C11 = {⟨ 0, CAS=[145] NUM=S, 0⟩,
⟨ *-te*, CAS=[145] NUM=S, D⟩,
⟨ *-ti*, CAS=[2367] NUM=P, 0⟩}

C20 = {⟨ *-et*, CAS=[145] NUM=S, 0⟩,
⟨ *-eti*, CAS=[2367] NUM=S, 0⟩,
⟨ *-íti*, CAS=[2367] NUM=P, d⟩}

Ostatní základní číslovky se buď skloňují podle substantivních vzorů (např. *milion*), nebo mají nepravidelné skloňování, které je vyřešeno výjimkou (v Praze rozepsáním slovních tvarů, v Brně zvláštním vzorem).

Tabulka 13.1 uvádí příklad pro číslovky *patnáct* a *dvacet*.

Kofix	Zakončení	Pád CAS	Číslo NUM	Flektivní mutace FMU	Slovní tvar
patnáct	ϵ	[145]	S	0	*patnáct*
	-e	[145]	S	D	*patnácte*
	-i	[2367]	P	0	*patnácti*
dvac	*-et*	[145]	S	0	*dvacet*
	-eti	[2367]	S	D	*dvaceti*
	-íti	[2367]	P	0	*dvacíti*

Tabulka 13.1: Příklad tvarů číslovek *patnáct* a *dvacet*

Lemma	Kofix	Vzor	Poznámka
první	*prvn*	`Ci`	
	prvn	`Cy23`	*prvnější*
druhý	*dru*	`Chy`	*druhý, druzí*
pátý	*pát*	`Cy`	*pátý, pátí*
dvojí	*dvoj*	`Ci`	
paterý	*pater*	`Cyy`	**pateří* v 1pl živ. se neužívá

Tabulka 13.2: Příklady řadových a druhových číslovek se vzory. Číslovka *první* je zřejmě jediná, kterou lze stupňovat.

13.1.2 Číslovky řadové a druhové

Číslovky řadové a druhové se skloňují podle adjektivních vzorů. Příklady ukazuje tabulka 13.2.

13.1.3 Číslovky úhrnné a souborové

Mezi těmito číslovkami je poměrně nejasný rozdíl. Často se považují za jeden druh. Při jejich skloňování dochází k přechodům mezi oběma poddruhy, dokonce se může připlést ještě zakončení číslovek druhových, jak ukazuje tabulka 13.3 s příkladem skloňování *5 dveří* (záměrně nevypisujeme číslovku 5 slovy). Neurčujeme zde ani hodnotu kategorie `Číslo`.

Pád	Alternativní tvary		
1	*patery dveře*	*pateré dveře*	*patero dveří*
2	*patera dveří*	*paterých dveří*	*patero dveří*
3	*pateru dveří*	*paterým dveřím*	*patero dveřím*
4	*patery dveře*	*pateré dveře*	*patero dveří*
6	*pateru dveří*	*paterých dveřích*	*patero dveřích*
7	*paterem dveří*	*paterými dveřmi*	*patero dveřmi*

Tabulka 13.3: Příklad neostré hranice mezi číslovkami úhrnnými, souborovými a druhovými

Číslovky druhové jsme vydělili zvlášť kvůli možnosti jejich čistě adjektivního skloňování. Číslovky úhrnné a souborové zde pro jejich obtížné vymezení rozdělovat nebudeme. Vzory by se mohly jmenovat opět podle kombinace kódů kategorií `Slovní druh` a `Poddruh`, tedy `Cu` pro číslovky úhrnné a `Cs` pro číslovky souborové. Je třeba pouze rozhodnout, které tvary do jednotlivých poddruhů patří.

13.1.4 Číslovky násobné, opakovací a výčtové

Číslovky násobné, opakovací a výčtové mají charakter příslovce a jako takové se neskloňují, nemají tedy ani vzor.

13.1.5 Číslovky dílové

Číslovky dílové jsou jen tři, totiž *půl, čtvrt* a *třeť. Půl* je nesklonné[1], druhé dvě, ač mají substantivní skloňování, pojímáme jako výjimky a nepřiřazujeme jim speciální vzor.

13.2 Zájmena

13.2.1 Zájmena substantivní

Zájmena substantivní určitá nemají společný vzor.

Vytvoříme vzory `kdo`, `co` pro zájmena substantivní tázací a vztažná. Podle nich se skloňují i ta zájmena substantivní neurčitá a záporná, která vznikla z těchto dvou zájmen pomocí předpony nebo přípony, tedy např. *kdekdo, nikdo, všelico, kdokoliv, cosi*. Předponu zde opět pojímáme šíře, než je obvyklé, jako počáteční řetězec.

Zájmena s předponou jsou bezproblémová. V podstatě bychom s nimi mohli nakládat stejně, jako s předponovým guessrem (viz poznámka pod čarou na str. 70). To však neděláme, neboť se jedná o slova poměrně častá.

Se zájmeny vzniklými pomocí přípony je třeba zacházet odlišně, neboť zde dochází k flexi „uvnitř“ slovního tvaru (*cosi, čehosi, čemusi*, atd.).

Máme v zásadě dvě možnosti. Buď každé takové zájmeno pojmeme jako výjimku a umístíme do slovníku všechny jeho slovní tvary s morfologickými značkami. Tak je to v současném pražském slovníku. Druhá možnost přenechává práci morfologickým nástrojům.

První přístup je obecnější, neboť je veškerá informace uložena ve slovníku. Nástroje, které se vyvíjejí pro takový slovník, jsou potom obvykle snadno přenositelné na jiné jazyky.

Druhý přístup je zase elegantnější a přehlednější.

Rozhodli jsme se zůstat u současné praxe a speciální vzory pro skloňování těchto zájmen (zatím) nezavádět.

[1] Otázkou, zda slovní tvar *půli* náleží číslovce *půl*, nebo podstatnému jménu *půle*, se zde zabývat nebudeme.

13.2.2 Zájmena přivlastňovací

U některých zájmen přivlastňovacích můžeme využít adjektivního vzoru `Pi`. Jsou to tato zájmena:

- určité *její*,
- tázací a vztažné *čí*,
- záporné *ničí*,
- neurčitá *něčí, všeličí*, a další, viz tabulka 4.1.

Ostatní zájmena jsou výjimkami.

Neurčitá a záporná zájmena vzniklá z lemmatu *čí* pomocí přípon (*čísi, čípak* a další) se řeší stejně jako podobně tvořená zájmena substantivní, i když i zde by se dalo uvažovat o systémovém řešení (viz oddíl 13.2.1).

13.2.3 Zájmena ukazovací a vymezovací

Ta zájmena ukazovací a vymezovací, která mají adjektivní tvar, tedy *takový, onaký, taký, každý, samý, všeliký, veškerý*, se skloňují podle tvrdého adjektivního vzoru `Pky`, `Pry` a `Py`, podle zakončení svého lemmatu.

Pro ostatní se vzor nevytváří.

13.3 Ostatní zájmena

Zájmena zařazená do poddruhu ostatní (SUB=0) mají adjektivní skloňování podle vzorů `Pky` (např. *jaký, všelijaký*), `Pry` (např. *který, leckterý*) a `Py` (např. *žádný*). [2]

Problém se zájmeny odvozenými pomocí přípon (*jakýsi, kterýkoli*, ...) se řeší stejně jako u zájmen substantivních (viz oddíl 13.2.1).

[2] Ač jsme na několika místech tvrdili, že se nechceme zabývat výjimkami, upozorňujeme zde na neobvyklý slovní tvar *žáden*. Vyskytuje se jen v mužském rodě (u tvaru *žádna*, který se též vyskytuje, jde pravděpodobně vždy o překlep, stejně jako *žádno, žádni, žádnu* i *žádny*). Od tvaru *žádný* ho odlišujeme hodnotou FMU=e (vložené *e*)

14 Závěr

Předložená práce se zabývá systémem morfologického popisu češtiny přesto, že na toto téma bylo napsáno i řečeno již mnoho. Po více než deseti letech intenzivního užívání elektronických morfologicky anotovaných korpusů češtiny se totiž ukazuje, že leckteré detaily popisu potřebují revizi, doplnění, případně i zcela odlišný přístup. Všechny návrhy vznikly z připomínek, stížností i nápadů ze strany uživatelů korpusů řady SYN, i na základě vlastní práce s těmito korpusy a s pražským morfologickým slovníkem.

Cílem práce bylo navrhnout systém, který jednotně, konzistentně a co nejúplněji popíše všechny morfologické jevy, které jsou potřebné při práci s českým jazykovým korpusem.

V prvních kapitolách se zabýváme systémem kategorií, pomocí kterých se české slovní tvary popisují. Patří sem i pojednání o lemmatizaci. V druhé části potom navrhujeme nový způsob zápisu flektivních a derivačních vzorů.

V čem spočívá přínos předkládané práce:

Definujeme přesné vymezení jednotlivých morfologických kategorií a jejich hodnot. Přitom se snažíme být v souladu s tradičními lingvistickými popisy. V některých případech však navrhujeme vlastní, netradiční řešení, protože tradiční popis nevyhovuje požadavkům na použití při automatickém zpracování jazyka. Vždy však dbáme na to, aby veškerá lingvistická informace zůstala zachována.

Stanovujeme základní princip pro budování morfologického slovníku, a to Zlaté pravidlo morfologie. S tím souvisí i důsledné zpracování slovních variant. Předkládáme jejich formální popis, který není závislý na jejich neobjektivním stylovém hodnocení. Zavádíme termín mutace, který pojetí varianty rozšiřuje. Mutace potom dělíme na flektivní a globální, což nám usnadní zachytit jejich variabilitu a snadnou kombinovatelnost.

Pro globální mutace zavádíme tzv. vícenásobné lemma, které umožní zahrnout tyto mutace pod společné lemma, ale přesto je popisuje tak, aby zůstaly rozlišeny a neporušily tak Zlaté pravidlo morfologie.

Při popisu jednotlivých morfologických kategorií jsme narazili na několik případů, které dosavadní systémy pro automatické zpracování češtiny buď zcela ignorují, nebo popisují nekonzistentně, nepřehledně nebo ne příliš šikovně vzhledem k dalšímu zpracování morfologicky označených textů. Tyto problematické případy se snažíme řešit lépe. Toho dosahujeme většinou zavedením nových kategorií pro popis některých jevů, nebo zavedením nových hodnot kategorií tradičních.

Mezi nově zavedené kategorie patří kategorie `Duál`, která umožní snadnější práci při analýze češtiny na syntaktické rovině a rovinách vyšších.

Dalšími kategoriemi jsou `Flektivní mutace` a `Globální mutace`, které rozšiřují známé termíny varianta a dubleta. Pomocí těchto kategorií popisujeme ty slovní tvary, které se v hodnotách ostatních kategorií neliší.

Nová je i kategorie `Slovesný tvar`, která slučuje tradiční slovesné kategorie,

neboť jejich hodnoty se většinou nedají vzájemně kombinovat. Některé hodnoty této kategorie jsou relevantní pro jiný slovní druh než pro sloveso. Konkrétně jde o pasivum, které anotujeme jako jmenný tvar přídavného jména, a kondicionál, který je relevantní pouze pro tři lemmata: částici *by* a spojky *aby, kdyby*.

Nové hodnoty jsme přidali kategorii `Slovní druh`. Důležitá je hodnota `cizí slovo`, která umožní popsat cizojazyčná slova, především vlastní jména, u nichž nemá smysl se snažit o zařazení do českého morfologického systému. Cizí jména přinášejí v dalších rovinách zpracování mnoho problémů. Jejich vyjmutím z množiny tradičních slovních druhů s nimi můžeme zacházet podle potřeby odlišně.

Podáváme rozbor tzv. složenin, které se nedají jednoznačně zařadit do systému slovních druhů, neboť ve svém tvaru jich sdružují vícero. V popisu složenin použijeme též nově zavedený koncept vícenásobného lemmatu.

Mezi hodnoty kategorie `Osoba` zahrnujeme též zdvořilostní formu druhé osoby singuláru, tedy vykání, které sice do paradigmatu zájmen a sloves patří, ale nebylo do něj dosud formálně zařazeno.

Podobný je případ stupně `s` kategorie `Stupeň`, tedy tvarů typu *sebekrásnější, sebekrásněji*, které se také dosud neanalyzovaly jako součást paradigmatu příslušného pozitivu přídavného jména nebo příslovce, k němuž přirozeně náleží.

Dále se zabýváme tzv. „stupňováním sloves“, které je sice spíše okrajovým jevem, ale velmi pravidelným, takže jeho zařazení do paradigmatiky sloves je také přirozené.

Nově navrhujeme systém flektivních vzorů, a to tak, aby byly parametrizovatelné. Pomocí vhodného nastavení parametru jednotlivým vzorům dosáhneme lepšího pokrytí systémových slovních tvarů. Flektivní vzory jsou, podobně jako v dosavadním pražském systému vzorů, doplněny o vzory umožňující pravidelné tvoření slov odvozených. Na rozdíl od pražského systému však i zde zavádíme parametrizaci, která umožní volbu, které odvozeniny tvořit a které ne.

Je velmi pravděpodobné, že se najdou slova, která náš návrh úplně nepokryje. Jsme však přesvědčeni, že návrh je dostatečně obecný na to, aby umožnil i řešení složitých výjimek.

Návrh, který jsme zpracovali, je v současné době již částečně implementován:

- Vytvořili jsme nástroje pro převod vzorů ze současného pražského morfologického slovníku do nového systému.
- Nově zpracovaný nástroj pro morfologickou analýzu pracuje s prefixovým guessrem navrženým na základě zmiňovaného výzkumu předpon.
- Postfixový guesser jsme použili při morfologické anotaci jedné z verzí korpusu SYN2000.
- Začali jsme pracovat na návrhu konkrétního schématu pro uchování morfologického slovníku v systému PML (Prague Markup Language) pro ukládání jazykových dat. V novém formátu budou především snadno zaznamenatelné derivační vztahy mezi jednotlivými lemmaty.

Chtěli bychom implementaci co nejdříve dokončit, aby mohly morfologické nástroje začít pracovat již s novými kategoriemi a jejich hodnotami. S tím sou-

visí i prapůvodní impuls k započetí této práce, totiž sjednocení pražského a brněnského pohledu na morfologické anotace, zejména vytvoření jednoznačného převodu mezi morfologickými značkami pražskými a brněnskými. Výrazně se tak zjednoduší vzájemná spolupráce.

Literatura

Akademická mluvnice. *Mluvnice češtiny 2.* Praha, ACADEMIA, 1986.

Brabcová, R. Kolísání rodu substantiv. In *Korpus jako zdroj dat o češtině*, s. 47–50. Brno: Masarykova univerzita, 2004.

Cruse, D. A. *Lexical Semantics.* Cambridge, UK, Cambridge University Press, 1986.

Čermák, F. Povaha a úzus interjekcí: případ češtiny. In *Computer Treatment of Slavic and East European Languages*, s. 299–307. Slovak Academy of Sciences, 2007.

Hajič, J. *Disambiguation of Rich Inflexion.* Praha, Karolinum, 2004.

Hajič, J. *Unification Morphology Grammar.* PhD thesis, Matematicko-fyzikální fakulta Univerzity Karlovy v Praze, 1994.

Hanks, P. – Pustejovsky, J. Common Sense About Word Meaning: Sense in Context. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference, TSD 2004*, Berlin Heidelberg, 2004. Springer-Verlag.

Havránek, B. – Jedlička, A. *Česká mluvnice.* Praha, Státní pedagogické nakladatelství, 1981.

Hlaváčová, J. Pravopisné varianty a morfologická anotace korpusů. In *Grammar & Corpora / Gramatika a korpus*, s. 161–168. Praha: Academia, 2008.

Hlaváčová, J. – Hrušecký, M. Affisix — Tool for Prefix Recognition. In *Lecture Notes in Artificial Intelligence, Proceedings of the 11th International Conference, TSD 2008*, s. 85–92, Berlin Heidelberg, 2008. Springer-Verlag.

Hlaváčová, J. – Kolovratník, D. Morfologie češtiny znovu a lépe. In *Informačné Technológie – Aplikácie a Teória. Zborník príspevkov, ITAT 2008*, s. 43–47, 2008.

Hlaváčová, J. – Lopatková, M. Variants and Homographs: Eternal Problem of Dictionary Makers. In *Proceedings of the 11th International Conference, TSD 2008*, s. 93–100, Berlin Heidelberg, 2008. Springer-Verlag.

Hlaváčová, J. Morphological Guesser of Czech Words. In *Proceedings of TSD 2001*, s. 70–75. Springer-Verlag Berlin Heidelberg, 2001.

Jiranová, P. Morfologická a syntaktická charakteristika českých číslovek vyjadřujících počet entit, jejich souborů a druhů. Diplomová práce, Filosofická fakulta UK v Praze, 2008.

Karlík, P. – Hladká, Z. Kam s ním? (Problém stupňování adjektiv). In *Život s morfémy. Sborník studií na počest Zdenky Rusínové*, s. 73–93. Brno: Masarykova univerzita v Brně, 2004.

Kirschner, Z. MOSAIC — A Method of Automatic Extraction of Significant Terms from Texts. Technical report, Faculty of Mathematics and Physics, Charles University, Prague, 1983.

Kopečný, F. *Základy české skladby*. Praha, Státní pedagogické nakladatelství, 1962a.

Kopečný, F. *Slovesný vid v češtině*. Praha, Nakladatelství ČSAV, 1962b.

Koskenniemi, K. Two-level morphology: a general computational model for word-form recognition and production. Technical Report Publication No. 11, Helsinki: University of Helsinki Department of General Linguistics, 1983.

Křístek, M. Způsoby vymezování stylové příznakovosti v lexiku (na materiálu současné češtiny). In *Varia IX: zborník materiálov z IX. kolokvia mladých jazykovedcov*, s. 102–112. Slovenská jazykovedná spoločnosť pri SAV, 2002.

Kučera, K. Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora. In *Computer Treatment of Slavic and East European Languages*, s. 121–125. Slovak Academy of Sciences, 2007.

Lopatková, M. – Žabokrtský, Z. – Benešová, V. Valency Lexicon of Czech Verbs VALLEX 2.0. Technical Report 34, UFAL MFF UK, 2006.

Osolsobě, K. O rozdílech mezi pražským a brněnským značkováním. Nepublikováno.

Osolsobě, K. *Algoritmický popis české formální morfologie a strojový slovník češtiny*. PhD thesis, Filosofická fakulta Masarykovy univerzity v Brně, 1996.

Panevová, J. Honorifika v češtině (České vykání - teorie a korpusová data). In *Vybrané kapitoly z české gramatiky*. Praha: Academia, 2008. v tisku.

Pravidla. *Pravidla českého pravopisu*. Ústav pro jazyk český, Praha, Pansofia, 1993.

Przepiórkowski, A. *The IPI PAN Corpus: Preliminary version*. Warszawa, IPI PAN, 2004.

Romportl, S. *Struktura gramatické složky slovesných tvarů určitých v češtině*. Praha, ACADEMIA, 1970.

Rosen, A. – Saloni, Z. Kategorie honorativu v českých konjugačních paradigmatech. *Slovo a slovesnost*. 2006, , 1, s. 36–45.

Sedláček, R. Morfologický analyzátor češtiny. Diplomová práce, Fakulta informatiky Masarykovy univerzity v Brně, 1999. http://nlp.fi.muni.cz/projekty/ajka.

Sgall, P. – Hronek, J. *Čeština bez příkras*. Praha, H&H, 1992.

Skoumalová, H. Czech lexicon by two-level morphology. In *Proceedings of the Second European Seminar of TELRI – Language Applications for a Multilingual Europe*, s. 123–145. IDS/VDU, Mannheim/Kaunas, 1997.

Slovenská morfologie. *Morfológia slovenského jazyka*. Bratislava, Vydavatelstvo SAV, 1966.

Ševčíková, M. Pronouns Introducing Content Clauses. In *Grammar & Corpora / Gramatika a korpus*, s. 277–284. Praha: Academia, 2008.

Šimandl, J. Kvantifikátory v korpusech ÚČNK a možnosti jejich značkování. Nepublikováno, 2007.

Šimková, M. O lexikálnom význame častíc. *Slovenská reč*. 2001, , 66, s. 37–51.

Šmilauer, V. *Novočeské tvoření slov*. Praha, Státní pedagogické nakladatelství, 1971.

Tušková, J. M. *Variantní a dubletní tvary v současné deklinaci apelativních feminin*. Spisy Pedagogické fakulty MU, sv. č. 98. Brno, Masarykova univerzita, 2006.

Urrea, A. M. – Hlaváčová, J. Automatic Recognition of Czech Derivational Prefixes. In *Proc. CICLING 2005*, s. 189–197. Springer-Verlag Berlin Heidelberg, 2005.

A Přehled kategorií a jejich hodnot

Uvádíme souhrnný přehled kategorií a jejich hodnot z kapitoly 4. Číslování odpovídá návrhu morfologické značky z oddílu 4.3, souhrn tedy neobsahuje hodnoty kategorií `Flektivní mutace` a `Globální mutace`.

U hodnot kategorie `Slovní druh` uvádíme rovnou všechny poddruhy, které jsou pro daný slovní druh relevantní. Souhrnný přehled všech poddruhů je uveden dále pod číslem 2, kde je v závorce místo příkladu kód slovních druhů, pro které je daná hodnota relevantní.

1. Slovní druh (POS)
   - N: podstatné jméno
     - S: deverbativní typu *věznění, pokrytí,...*
     - 0: ostatní
   - A: přídavné jméno
     - U: přivlastňovací (*matčin, otcův,...*)
     - G: od přechodníku přít. (*mající, sedící, beroucí,...*)
     - M: od přechodníku min. (*ušedší, nakupovavší,...*)
     - S: ostatní deverbativní (*namazaný, zemřelý, nakousnutý, namazán, nakousnut,...*)
     - 0: ostatní (*jarní, starý,...*)
   - P: zájmeno
     - Z: substantivní (*já, kdo, nikdo, oni,...*)
     - U: přivlastňovací (*můj, čí,...*)
     - D: ukazovací (*ten, takový,...*)
     - V: vymezovací (*každý, všechen, týž, sám*)
     - 0: ostatní
   - C: číslovka
     - 1: základní (*jedna, sto,...*)
     - r: řadové (*druhý, pátý,...*)
     - u: úhrnné (*dvé, patero,...*)
     - s: souborové (*dvoje, patery,...*)
     - d: druhové (*dvojí, paterý,...*)
     - n: násobné (*dvakrát, pětkrát,...*)
     - o: opakovací (*podruhé, popáté,...*)
     - v: výčtové (*zadruhé, zapáté,...*, ale i *druhé* z dvojice *za druhé,...*)
     - p: dílové (*půl, čtvrt, třeť*)
   - V: sloveso
     - m: modální (*moci/moct, mít, mívat, muset, musívat, smět, chtít, hodlat, dát se, dávat se, dovést, umět*)
     - f: fázová (*začít, začínat, přestat, přestávat, zahájit, skončit,...*)
     - b: pomocná (*být, bývat, mít, dostat*)
     - 0: ostatní (*navštívit, koupat se,...*)

- D: příslovce
  - P: místní (*kudy, tudy, odkud, nikudy, nikam; daleko, nedaleko,...*)
  - T: časová (*kdy, nikdy; včera, odpoledne,...*)
  - D: způsobová (*jak, všelijak; krásně, velmi, široce,...*)
  - R: predikativní (*jasno, možno, teplo, volno,...*)
- R: předložka
- J: spojka
  - ^ (stříška): souřadicí (*a, ale, nebo,...*)
  - , (čárka): podřadicí (*protože, když, až, -li,...*)
  - * (hvězdička): matematické operace (*plus, minus/mínus, krát, děleno* — neplést s *děleno* jako jmenný tvar přídavného jména *dělený*)
- I: citoslovce
- T: částice
  - 7: zvratné (*se, si*)
  - c: kondicionálová (pouze *by*)
  - 0: ostatní (*ba, ano, bože, nechť, ať,...*)
- F: cizí slovo (K)
- G: prefixový segment (K)
- S: složenina (K)
- X: neznámé slovo

2. Slovní poddruh (SUB)
   - ^ (stříška): spojka souřadicí (J)
   - , (čárka): spojka podřadicí (J)
   - * (hvězdička): matematické operace (J)
   - b: pomocné sloveso (V)
   - c: kondicionálová částice (T)
   - d: druhová číslovka (C)
   - D: ukazovací zájmeno (P)
   - f: fázové sloveso (V)
   - G: adjektivum od přechodníku přít. (A)
   - J: způsobové příslovce (D)
   - m: modální sloveso (V)
   - M: adjektivum od přechodníku min. (A)
   - n: násobná číslovka (C)
   - o: opakovací číslovka (C)
   - p: dílová číslovka (C)
   - P: místní příslovce (D)
   - R: predikativní příslovce (D)
   - r: řadová číslovka (C)
   - S: deverbativní adjektivum/substantivum (AN)
   - s: souborová číslovka (C)
   - T: časové příslovce (D)
   - U: přivlastňovací adjektivum/zájmeno (AP)
   - u: úhrnná číslovka (C)
   - v: výčtová číslovka (C)
   - V: vymezovací zájmeno (P)
   - Z: substantivní zájmeno (P)
   - 1: základní číslovka (C)
   - 7: zvratná částice (T)
   - 0: ostatní (vše)

3. Funkce (FCE)
   - U: určitá (všechna osobní zájmena, určité číslovky, *tady, teď*,...)
   - N: neurčitá (*někdo, čísi, několik, někdy,...*)
   - Z: záporná (*nikdo, ničí, nijak,...*)
   - T: tázací (*kdo, čí, kolik, kde,...*)
   - V: vztažná (*kdo, čí, jenž, kdy,...*)
   - S: zvratná (*se, si, sobě, sebe, sebou*)

4. Slovesný vid (ASP)
   - D: dokonavý (*koupit, napsat, doručit, narodit se,...*)
   - N: nedokonavý (*kupovat, psát, doručovat, chodívat,...*)
   - O: obouvidý (*referovat, absolvovat, izolovat,...*)

5. Zkratka (ABR)
   - + : ano

6. Rod (GEN)
   - M: mužský životný
   - I: mužský neživotný
   - F: ženský
   - N: střední
   - X: sdružená hodnota
7. Číslo (NUM)
   - S: jednotné
   - P: množné
   - X: sdružená hodnota
8. Duál (DUA)
   - +
9. Pád (CAS)
   - 1 až 7
   - X: sdružená hodnota
10. Osoba (PER)
    - 1 až 3
    - v: vykání
11. Stupeň (DEG)
    - 1: pozitiv
    - 2: komparativ
    - 3: superlativ
    - s: typ sebe + komparativ
12. Negace (NEG)
    - N: pro záporné slovní tvary, které začínají záporkou *ne-*
    - A: pro ostatní slovní tvary
13. Slovesný tvar (VRB)
    - P: indikativ prézentu (*kolíbá*)
    - B: budoucí čas (*ponese, bude*)
    - F: infinitiv (*otevřít*)
    - I: imperativ (*peč*)
    - L: příčestí činné (*strouhal*)
    - T: příčestí trpné (*zavřen*)
    - K: kondicionál (*aby, kdyby, by*)
    - p: přechodník přítomný (*starajíc*)
    - m: přechodník minulý (*vstoupiv*)
14. Jmenný tvar přídavných jmen (NOM)
    - J: jmenný tvar
    - 0: ostatní přídavná jména

15. Stupeň intenzity slovesného děje (INT)
    - r: pro předponu *roz-*
    - p: pro předponu *po-*
    - z: pro předponu *za-*
    - n: pro předponu *na-*
    - v: pro předponu *vy-*
    - u: pro předponu *u-*
16. Typ složeniny (CMP)
    - n: zájmenný (*proň, zaň*)
    - c: zájmenný (*oč, zač, začpak*)
    - t: zájmenně-slovesný (*toť*)
    - Z: zkratkový (*atd., apod.*)
    - A: slovesný, 1. složka je A (*krásnýs*)
    - N: slovesný, 1. složka je N (*latinys*)
    - P: slovesný, 1. složka je P (*jemus*)
    - C: slovesný, 1. složka je C (*koliks*)
    - V: slovesný, 1. složka je V (*zavřelas*)
    - D: slovesný, 1. složka je D (*včeras*)
    - T: slovesný, 1. složka je T (*sis, ses*)
    - J: slovesný, 1. složka je J (*protožes*)
    - S: slovesný, 1. složka je S (*načs*)

B Kopie účastnického slibu z Konkláve

Přísahám na svou čest,

že během celého konkláve budu jednat podle svého nejlepšího
morfologického svědomí,

zejména že

zapomenu na veškeré aplikace, na kterých jsem dosud pracoval(a),

a obrátím svou mysl výhradně do budoucnosti,
aby výsledkem našeho jednání byla skutečně čistá a průzračná
morfologie českého jazyka,
nezatížená jakýmikoliv partikulárními zájmy.

Karel Pala

Vladimír Petkevič

Karel Oliva

Jan Hajič

Klára Osolsobě

Jaroslava Hlaváčová

Cikháj, 21.října 2005

Rejstřík