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

Teze

Program: Filologie
Obor: Matematická lingvistika

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

Praha 2009

Cílem práce bylo 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. Tam, kde to je možné, používáme zavedených a odůvodněných lingvistických 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.

Úvodní definice

Na začátek uvedeme základní termíny, které používáme k popisu české morfologie, neboť jejich interpretace většinou není jednoznačná.

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

Slovní tvar je slovo, které má význam.[1]

Lemma je základní slovní tvar. Ve slovnících se používá jako slovníkové heslo.

Všechny slovní tvary, které lze vytvořit z jednoho lemmatu pomocí skloňování, časování nebo stupňování (obecně ohýbání), tvoří **paradigma**. Můžeme také říci, že paradigma je množina slovních tvarů, které náležejí danému lemmatu. Do paradigmatu zahrnujeme i nespisovné (nekodifikované) slovní tvary.

Morfologická kategorie je vlastnost slovních tvarů.

Morfologickou kategorii, která popisuje množinu slovních tvarů, budeme nazývat **relevantní morfologickou kategorií** této množiny.

Morfologická značka je řetězec znaků, který kóduje hodnoty všech relevantních morfologických kategorií pro nějaký slovní tvar nějakého lemmatu. Hodnoty morfologických kategorií popisují slovní tvary, proto můžeme i o morfologické značce říci totéž.

Lemmatizaci chápeme jako zobrazení λ z množiny slovních tvarů S do množiny lemmat L. 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. 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.

[1] Úvahy nad významem slova „význam“ jsou mimo rámec této práce.

Kromě homonymie jsou při lemmatizaci problematické ještě varianty. Vezměme si např. slovní tvary *diskuze* a *diskuse*. Máme je analyzovat jako dvě různá lemmata, nebo varianty lemmatu jednoho?

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 je dvouprvková množina {*diskuze*, *diskuse*}. Prvkům této množiny budeme říkat **variantní lemmata**.

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

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. Pro složeniny tedy definujeme vícenásobné lemma poněkud odlišně:

Vícenásobné lemma složeniny je množina lemmat jednotlivých složek složeniny.

Klíčovými termíny pro náš popis systému morfologie jsou: slovní tvar, lemma, morfologická značka. Z toho konkrétně vyplývá, že se nezabýváme víceslovnými výrazy.

Každý slovní tvar by měl být jednoznačně popsán svým lemmatem a morfologickou značkou. Tomuto požadavku říkáme **Zlaté pravidlo morfologie**. Důvod, proč v současných systémech není vždy splněno, spočívá zejména v nekonzistentním zpracování variant. Existují množiny slovních tvarů (většinou jde jen o dvojice), které jsou popsány stejným lemmatem a stejnou morfologickou značkou. Kvůli nejednoznačnému vymezení pojmů varianta a dubleta v literatuře zavádíme obecnější pojem mutace.

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.

S použitím termínu mutace můžeme přeformulovat Zlaté pravdilo morfologie a schematicky ho znázornit takto:

$$\boxed{\text{lemma} + \text{morfologická značka} + \text{mutace} = \text{jednoznačný slovní tvar}}$$

Mutace

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 (tak to udělali např. pro polský korpus IPI PAN), ale zavedení kategorie `Mutace` umožní vyřešit tento problém pro všechny podobné případy najednou.

Mutace mohou být dvojího typu, přičemž jeden nevylučuje druhý:

- **globální mutace** týkající se celého paradigmatu, tj. všech slovních tvarů,
- **flektivní mutace** týkající se jen některých tvarů daného paradigmatu.

Globální i flektivní mutace se mohou kombinovat. Globální mutace mají vícenásobné lemma a projevují se ve všech tvarech příslušného paradigmatu. Navíc mohou mít jednotlivé tvary ještě flektivní mutace. V příkladě {*diskuse, diskuze*} jsou to např. tvary pro 7pl: *diskusemi, diskuzemi, diskusema, diskuzema*, projevující se jako globální mutace *s—z* a flektivní mutace *emi—ema*.

Hodnoty kategorií `Globální mutace` a `Flektivní mutace` zavádíme nezávisle na jakémkoli hodnocení. Nevyjadřují tedy, jak je zvykem u variant, stylový ani žádný jiný příznak, neboť tato hodnocení jsou velmi subjektivní. 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. Jediným jejich smyslem je odlišit slovní tvary, které mají stejný morfologický popis.

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

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

Kromě všeobecných hodnot z tabulky existuje celá řada mutací typických pouze pro některé kombinace hodnot morfologických kategorií, nebo jen pro některá lemmata.

Kategorie `Flektivní mutace` ani `Globální mutace` nezahrnujeme do morfologické značky. Vyčleňujeme je jako další, speciální atributy popisu a souhrnně jim říkáme `Mutace`. V rámci této kategorie 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`. Takto lze zachytit obrovské 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:

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

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

Morfologické kategorie

Morfologické kategorie také dělíme na globální a flektivní:

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

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

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

- 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

Netradiční řešení některých kategorií a jejich hodnot

Slovní druh

Zavádíme 3 nové hodnoty:

- **cizí slovo** pro popis cizích slov objevujících se v českém textu,
- **prefixový segment** pro popis neúplných slov typu *tří až čtyřprocentní*,
- **složenina** pro popis slovních tvarů *oň, zač, koliks, skákalas* apod.

Funkce

Tuto kategorii jsme zavedli po vzoru brněnského systému pro vyjádření dvojí charakteristiky zájmen, číslovek a zájmenných příslovcí. Hodnoty kategorie `Funkce` se kombinují s hodnotami kategorie `Poddruh`, čímž umožňují přehledný popis těchto skupin lemmat. Hodnoty kategorie `Funkce` jsou:

- 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*)

Duál

Kategorie `Duál` se v současném pražském systému projevuje jako hodnota kategorie `Číslo`. Vyčleňujeme ji jako samostatnou kategorii proto, aby se při dalším zpracování lépe zachycovala shoda s tzv. duálními slovy (*oči, uši, nohy*, ...) a duálními koncovkami adjektiv v 7pl.

Osoba

Mezi tradiční hodnoty kategorie `Osoba` zařazujeme hodnotu novou, a to vykání, neboť její absence v současných systémech působí značné potíže na dalších rovinách popisu.

Stupeň

I v této kategorii přidáváme novou hodnotu pro popis tvarů typu *sebekrásnější, sebekrásněji*. Tato slova zařazujeme do paradigmatu lemmatu *krásný*.

Slovesný tvar

Tato kategorie se trochu podobá brněnské kategorie Mod pro popis slovesných tvarů. Její hodnoty jsou:

- 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*)

Nemusíme nyní zavádět kategorie času, způsobu ani slovesného rodu.

Na této kategorii je neobvyklé to, že se netýká výhradně sloves. Má-li nějaký slovní tvar hodnotu „trpný rod“, popisujeme ho jako jmenný tvar přídavného jména a jako takovému mu přiřazujeme slovní druh přídavné jméno. To, že se jedná o slovesné pasivum, je vyjádřeno právě hodnotou T kategorie `Slovesný tvar`.

Také hodnota K „kondicionál“ se netýká sloves. Žádný slovesný slovní tvar sám o sobě podmíněnost nevyjadřuje. Ta se dá vyjádřit pouze pomocí částice *by* nebo spojek *aby, kdyby*, včetně jejich časovaných tvarů. Proto i hodnotu „kondicionál“ mají pouze uvedená tři lemmata.

Jmenný tvar přídavných jmen

Tato kategorie popisuje jmenné tvary přídavných jmen, a to včetně slovesného pasiva. Má tedy kladnou hodnotu např. pro tvary *mlád, zavřena, ukryt*.

Stupeň intenzity slovesného děje

Mnoho nedokonavých 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 následující tabulka.

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

Dosadíme-li v tabulce místo X např. sloveso *mávat*, dostaneme sadu nových paradigmat. Jako lemma jim však přiřazujeme neprefigovaný tvar. Tedy např. tvaru *rozmávat* (*se*) přiřazujeme lemma *mávat*. Slovní tvary paradigmat s prefixem proto musíme odlišit od slovních tvarů bez prefixu. K tomu používáme právě kategorii `Stupeň intenzity slovesného děje`. Její název vyplývá z toho, že jednotlivá prefigovaná zvratná slovesa je možno s určitou tolerancí uspořádat podle intenzity děje do posloupnosti naznačené v tabulce.

Typ složeniny

Tato kategorie se týká pouze nového slovního druhu „složenina“. Její hodnoty jsou:

Hodnota	Popis	Příklady
n	zájmenná s. s lemmatem 2. složky *on*	*oň, proň*
c	zájmenná s. s lemmatem 2. složky *co*	*oč, zač*
t	jediná zájmenně-slovesná s.	*toť*
Z	zkratková s.	*apod, OPBH*
N	slovesná s. s 1. složkou substantivní	*oknos, latinys*
A	slovesná s. s 1. složkou adjektivní	*pěkného, zdravás*
P	slovesná s. s 1. složkou zájmennou	*komus, jehos*
C	slovesná s. s 1. složkou číslovkovou	*koliks, pěts*
V	slovesná s. s 1. složkou slovesnou	*hráls, běhalas*
D	slovesná s. s 1. složkou příslovečnou	*ještěs, včeras*
J	slovesná s. s 1. složkou spojka	*když, nebos*
T	slovesná s. s 1. složkou částice	*sis, ses*
S	slovesná s. s 1. složkou složenina	*očs, naňs*

Všem složeninám kromě zkratkových přiřazujeme vícenásobné lemma. V práci též ukazujeme, že jednotlivé složky složeniny i hodnoty jejich morfologických kategorií je možno zahrnout do korpusových dotazů, takže např. tvar *jemus* je možno nalézt i při dotazu na všechna zájmena v 3sg, ale i při vyhledávání sloves.

Morfologický slovník

Morfologický slovník obsahuje záznamy o lemmatech a jejich paradigmatech. Jednotlivé záznamy mohou být propojeny tzv. derivačními odkazy, které jsou obousměrné, nevyjadřují tedy jednosměrný vztah (co vzniklo z čeho), ale pouze vzájemnou příbuznost slov. Při praktickém využití odkazů např. pro automatický překlad jsou totiž třeba oba směry.

Informace o paradigmatech jsou ve slovníku uloženy v podobě vzorů.

Slovník nemusí zachycovat takové tvary, které lze pravidelně automaticky vytvořit, konkrétně jde o negaci a stupňování. Zde stačí jen do vzoru zaznamenat, zda se negativní či stupňované tvary mají tvořit, nebo ne.

K rozpoznání neznámých slovních tvarů, tedy takových, které nejsou popsány v morfogickém slovníku, se používá tzv. guesser. Popisujeme **prefixový guesser**, který na základě známých předpon umí rozpoznat neznámé tvary (např. *eurookny*) a přiřadit jim správné lemma a morfologické značky. **Postfixový guesser** k tomu využívá naopak slovní zakončení.

Vzory

K popisu pravidelných paradigmat v morfologickém slovníku používáme flektivní a derivační vzory. Formálně:

Flektivní vzor Ω je množina trojic $\langle$s, M, F$\rangle$, 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 $\langle$s, M, 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. Nazýváme ho kofix.

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**.

Flektivní vzory rozdělujeme 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. Lze je snadno kombinovat tak, aby popsaly různé možnosti flexe vyskytující se v českých paradigmatech. Parametry jsou zejména vhodné pro popis flektivních mutací slovních tvarů.

Stávající vzory, jak brněnské, tak i pražské, lze většinou pomocí vhodně zvolených parametrů převést na vzory nové.

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.