Problém variantních tvarů slov při automatickém zpracování jazyka

Jaroslava Hlaváčová

ÚFAL MFF UK v Praze
hlavacova@ufal.mff.cuni.cz

Abstrakt *Zápis slov v mnoha jazycích není jednoznačný, existují různé varianty. Někdy se jedná o varianty rovnocenné, jindy jsou některé nářeční, nespisovné či jinak příznakové. Při automatickém zpracování jazyka však chceme umět rozpoznat všechny, a současně jim přiřadit stejný základní tvar, tzv. lemma. Na druhou stranu ale potřebujeme všechny varianty od sebe nějakým způsobem odlišit, abychom např. mohli při automatické syntéze zvolit tu správnou. Příspěvek se zabývá možným řešením tohoto problému, a to zavedením tzv. vícenásobného lemmatu. Uvedeme možnosti jeho využití při konkrétních aplikacích, zejména v korpusové lingvistice.*

1 Úvod

Existují slova, která lze psát (a často i vyslovovat) více způsoby. Příkladem je dvojice slov *citron — citrón*. Jak se mají jazykové slovníky vypořádat s takovými dvojicemi, obecně n-ticemi?

Nejjednodušší by bylo prohlásit každou variantu za samostatné slovo, ale to odporuje obecnému chápání konceptu slova. Navíc to s sebou přináší spoustu technických problémů při automatickém zpracování textů, ať už z hlediska analýzy nebo syntézy. Vyhledává-li např. uživatel — lingvista nějaké takové slovo v jazykovém korpusu, je většinou rozumné, aby výsledkem byly všechny varianty, pokud si speciálně nechce některé odfiltrovat. Naopak při syntéze, tedy automatickém tvoření jazykových dat, je problém, jakou variantu pro výstup vybrat. Dosavadní slovníky, včetně těch, které se používají pro automatické zpracování jazyka, zacházejí s takovými slovy nejednotně, často i v rámci jediného slovníku.

2 Vícenásobné lemma

V předchozích odstavcích jsme pod pojmem **slovo** vlastně rozuměli slovníkové heslo, tedy základní slovní tvar, většinou nazývaný **lemma**. Funkce, která každému slovnímu tvaru přiřadí jeho základní tvar, lemma, se nazývá **lemmatizace**. Je to jeden ze základních kroků automatického zpracování jazyka. Jaké lemma mají mít varianty z Úvodu? Má lemma ortograficky odpovídat slovnímu tvaru (*citronu* → *citron*, *citrónu* → *citrón*), nebo se má zvolit jeden zástupce pro obě varianty? Který?

Obě řešení mají svá pro i proti. První případ je jednodušší a naprosto jednoznačný. Je to však čistě technické řešení, které nepostihuje důležitý fakt, že nejde o dvě různá slova, pouze o rozdílné zápisy. Kdyby např. chtěl uživatel korpusu vyhledat všechny varianty daného slova, musel by si o ně explicitně říci, tj. vytvořit dotaz tak, aby je všechny zahrnul. To může být problematické u takových slov, která mají variant hodně — uživatel ani nemusí všechny znát. Příkladem může být název státu *Afghánistán*, který má v korpusus SYN [2] varianty *Afghánistán, Afghánistan, Afgánistán, Afghanistán, Afganistán, Afganistan, Afgánistan, Afghanistan.* Jednodušší by samozřejmě bylo, kdyby všechny varianty byly obsaženy už ve slovníku. Zavádíme tedy koncept vícenásobného lemmatu:

Vícenásobné lemma je množina lemmat se stejným významem lišících se pouze zápisem (ortografické varianty).[1]

Vícenásobným lemmatem z našich příkladů jsou tedy množiny {*Afghánistán, Afghánistan, Afgánistán, Afghanistán, Afganistán, Afganistan, Afgánistan, Afghanistan*} a {*citron*, *citrón*}. Prvkům této množiny budeme říkat **variantní lemmata**.

2.1 Vícenásobné lemma v diachronní lingvistice

Vícenásobnými lemmaty se zabýval také Karel Kučera [5], 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. Nic ale nebrání tomu, využít vícenásobné lemma i pro diachronní korpusy. Dokonce si myslíme, že toto řešení je obecnější, neboť synchronní lemma k diachronnímu nemusí existovat, nebo jich může být více, a stojíme zase před problémem, které vybrat. Je tedy možné zařadit historický zápis lemmatu do

[1] Významem slova *význam* se v tomto příspěvku nezabýváme. Chápeme ho intuitivně.

množiny variantních lemmat, a tím z něj vytvořit prvek vícenásobného lemmatu.

3 Varianty a mutace

Termín varianta není v lingvistice definován jednotně, navíc se používá ještě termín dubleta, s podobným významem. Přehled různých chápání obou termínů je stručně podán v [6]. Zde si autorka vzala za východisko Mluvnici češtiny [1], 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. 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. Navíc to skutečně nejsou informace morfologické.

Nějakou kategorii, která rozliší slovní tvary se stejnými hodnotami ostatních morfologických kategorií, však potřebujeme, abychom mohli jednotlivé způsoby zápisu od sebe popisem odlišit. V práci [3] navrhujeme pro naše účely termín jiný, nezatížený množstvím nejednoznačných významů, a to mutace. Jeho vymezení je čistě technické:

Mutace jsou takové dvojice slovních tvarů, které mají stejné (vícenásobné) 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.

Rozlišujeme **mutace flektivní**, které se liší v zakončení (např. *hradu — hradě*), a **mutace globální**, které se projevují v celém paradigmatu, tzn. ve všech tvarech (např. již uvedený *citron — citrón*). Jednotlivá variantní lemmata jednoho vícenásobného lemmatu jsou globální mutace.

Jak bylo naznačeno hned v úvodu, budeme se zabývat pouze mutacemi globálními. Někdy jde o pouhé ortografické varianty (*atomismus — atomizmus*), někdy o různou výslovnost (*citron — citrón*), případně ovlivněnou obecnou češtinou (*mýdlo — mejdlo*). Vždy ovšem platí, že pokud je slovo možno nějakým způsobem ohýbat (časovat, skloňovat nebo stupňovat), mají mutace stejný ohýbací vzor, i kdyby byl nepravidelný.

Typ	Příklad	Značka
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*	

Tabulka 1. Přehled nejčastějších typů globálních mutací s příklady

3.1 Hodnoty kategorie Mutace

Jak jsme uvedli výše, upouštíme od dnes běžného označování mutací na základě subjektivních hodnocení. Kvůli jednoznačnému popisu každého slovního tvaru (i lemmatu) je však třeba mutacím nějaké hodnoty přiřadit. Hodnoty mohou být v zásadě libovolné, např. číslování, neboť jejich hlavní motivací je rozlišení. Když už ale mutace rozlišujeme, můžeme k tomu využít nějaké jejich konkrétní vlastnosti, jejichž hodnoty budou vycházet ze samotného zápisu mutací, nikoli z vnějších zdrojů, které se mohou časově i místně lišit (spisovnost vs. nespisovnost vs. nářečnost apod.).

Tabulka 1 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, které lze použít pro jednoznačné odlišení slovních tvarů, které se morfologicky jinak neliší (viz [3]). Kód `d` zastupuje „dlouhé“ mutace, `k` „krátké“. Podobně `m` znamená „měkké“, `t` „tvrdé“. [2] Výhoda takového značení je vidět z příkladu, ve kterém si uživatel přeje vypsat z korpusu všechna slova, v jejichž lemmatu došlo

[2] Kdybychom chtěli mutace popsat přesně, museli bychom do značky zahrnout i polohu rozdílů v zápisu jednotlivých mutací. Jak už bylo uvedeno výše, cílem není přesný popis, ale pouze rozlišení jednotlivých mutací. Z toho důvodu jsou i typy navrženy co nejobecněji a není možné podle nich zjistit přesný tvar zápisu.

k měkčení souhlásky. Schematicky můžeme takový dotaz zapsat takto:

$$\text{glob. mutace} = m$$

Hromadění typů mutací v jednom lemmatu se vyjádří vícero hodnotami, viz tabulka 2, kde jsou naznačeny možné kombinace globálních mutací d-k (dokonce dvakrát) a 0-h (viz tabulka 1).

Lemma	Značka
Afghánistán	hdd
Afghánistan	hdk
Afghanistán	hkd
Afghanistan	hkk
Afgánistán	0dd
Afganistán	0kd
Afgánistan	0dk
Afganistan	0kk

Tabulka 2. Příklad vícera hodnot kategorie globální mutace

4 Vícenásobné lemma v morfologickém slovníku

Morfologické slovníky, popisující (všechny) slovní tvary jazyka, většinou tyto tvary neobsahují přímo. Místo toho využívají ohýbacích vzorů. Každému slovnímu tvaru je tak ve slovníku přiřazeno lemma a vzor, podle kterého je možné vygenerovat všechny tvary daného slova. Tento princip může být zachován i v případě, použijeme-li místo jednoduchého lemmatu lemma vícenásobné. Každému slovnímu tvaru libovolného variantnímu lemmatu je tedy přiřazena množina všech těchto lemmat, tedy celé vícenásobné lemma. Vzhledem k tomu, že globální mutace se obvykle neliší v příponách, vzor bývá pro všechny prvky vícenásobného lemmatu stejný, stačí tedy připojit jediný vzor k celému vícenásobnému lemmatu.

Pro praktické použití slovníků je však třeba, aby množiny vícenásobných lemmat byly reprezentovány jednoznačným identifikátorem. Na tvaru identifikátoru vlastně nezáleží, mohou to být třeba čísla, ale kvůli přehlednosti a čitelnosti jsme zvolili reprezentaci slovní. Většina lemmat totiž i nadále bude jednoprvková, a tak není třeba jejich identifikátor, kterým dosud bylo lemma samotné, měnit.

Jak zvolit jednoznačný identifikátor množiny vícenásobného lemmatu? Existuje řada možností. Lingvisté by rádi měli jako identifikátor ten prvek, který je neutrální, bezpříznakový, tedy např. zástupcem vícenásobného lemmatu {*okno, vokno*} by bylo *okno*. Často je však toto kritérium subjektivní, někdy jich je bezpříznakových víc — to se týká vlastně všech ostatních dosud uvedených příkladů.

Objektivnějším kritériem je frekvence. Zástupce vícenásobného lemmatu by byl ten prvek, který je nejfrekventovanější. Zde se však okamžitě nabízí otázka: kde? Nejlepší odpovědí by bylo: v jazyce, ale takovou odpověď nikdo nezná. K tomu by bylo třeba zvolit nějaký referenční korpus, jehož frekvence by byly určující. I toto řešení má svá úskalí. Volba takového korpusu opět není stoprocentně objektivní. Vždy jde o nějaký výběr, jakkoli se tvůrci korpusů snaží o různé druhy vyváženosti. Navíc žádný korpus není dostatečně velký na to, aby obsahoval všechny možné mutace. A konečně, při malých frekvencích málo běžných slov přestává být frekvence v korpusu objektivní (frekvence 1 a 2 v mnohamilionovém korpusu neznamená, že jedno slovo je skutečně dvakrát četnější než druhé).

Identifikátorem vícenásobného lemmatu by mělo být jedno z variantních lemmat a mělo by být zvoleno podle jednoznačného objektivního kritéria, které nevyužívá žádných informací, které by závisely na nějakém dalším ne zcela objektivním zdroji. Jako nejpřirozenější se nám nakonec jeví zvolit za identifikátor ten prvek, který je při lexikografickém uspořádání první. Identifikátorem vícenásobného lemmatu {*okno, vokno*} je tedy *okno*, podobně {*Afghánistán, Afghánistan, Afgánistán, Afghanistán, Afganistán, Afganistan, Afgánistan, Afghanistan*} → *Afganistan*, {*citron, citrón*} → *citron*. Takto zvolený reprezentant vícenásobného lemmatu je vždy jednoznačný, nezávisle na jakýchkoli vnějších kritériích.[3]

5 Praktické využití vícenásobného lemmatu

Napřed popíšeme, jak by měla vypadat implementace vícenásobného lemmatu v korpusových manažerech. Potom naznačíme použití při generování, které je potřebné např. při automatickém překladu.

5.1 Zobrazování

Zobrazení vícenásobného lemmatu by mělo mít dva módy. Prvním je zobrazení identifikátoru, druhým zobrazení celé množiny všech variantních lemmat.

Při přípravě korpusu pro vstup do korpusového manažeru by u většiny jazyků zřejmě stačilo, aby v datech bylo každé vícenásobné lemma zastoupeno

[3] Poněkud zvláštně však vypadá např. identifikátor *tejden* pro vícenásobné lemma {*týden, tejden*}. Běžný uživatel korpusu však identifikátor vlastně nemusí znát, viz dále sekci 5.2.

pouze svým identifikátorem. K tomu je třeba zobrazení L1:

L1: slovní tvar → identifikátor

V případě jednoduchého lemmatu je identifikátorem samozřejmě vlastní lemma (jediný prvek jednoprvkové množiny). Ty identifikátory, které zastupují vícenásobné lemma, by byly obsaženy v přídavné tabulce a indexovány, aby se v nich dalo vyhledávat. Zavedeme tedy zobrazení L2, které odkazuje na příslušnou množinu variantních lemmat.

L2: identifikátor → vícenásobné lemma

K vyhledání celého vícenásobného lemmatu, tedy množiny variantních lemmat, by docházelo jen na přání uživatele, který má možnost si zvolit, v jakém tvaru chce který atribut zobrazit.

Upřesněním hodnoty globální mutace je samozřejmě možné zobrazit pouze některé prvky vícenásobného lemmatu.

5.2 Vyhledávání

Zatímco při zobrazování lemmatu stačilo mít ke každému identifikátoru přiřazenu množinu variantních lemmat, při vyhledávání je třeba, aby existovalo zobrazení inverzní, které každému variantnímu lemmatu přiřadí příslušný identifikátor. Máme tedy zobrazení

id: variantní lemma → identifikátor

Jestliže chce uživatel vyhledávat podle lemmatu, nemůžeme předpokládat, že bude znát jednoznačný identifikátor, neboť, jak bylo uvedeno výše, nemusí znát celou množinu všech variantních lemmat. Ať zadá jakékoli variantní lemma, je třeba k němu nalézt podle zobrazení id příslušný identifikátor. Ten je už obsažen a indexován v datech jako lemma k příslušným slovním tvarům, korpusový vyhledávač už tedy může podle něj vyhledávat.

Pokud si uživatel přeje vyhledat jen tvary konkrétní globální mutace, použije k vyhledání hodnotu globální mutace. Např. pro vyhledání všech tvarů lemmatu *citron*, ale ne *citrón* může dotaz vypadat schematicky takto:

lemma = *citron* & glob. mutace = *k*

5.3 Generování

V případě generování textů je problém složitější. Které lemma má generátor z množiny lemmat vybrat? Identifikátor nemusí být vždy správným řešením. Kdybychom např. chtěli generovat nespisovnou řeč, měli bychom z vícenásobného lemmatu {*okno, vokno*} vybrat *vokno*, přestože identifikátor je *okno*. V tomto případě by se hodilo mít každé variantní lemma nějakým způsobem popsané, např. spisovnost, archaičnost, slang, dále příslušnost k určitému oboru. Výše jsme však uvedli, že takové popisy do morfologického slovníku nepatří. Nic však nebrání vytvoření dalšího zdroje informací, který by sloužil k určitému účelu, v tomto případě ke generování textu. Takové přídavné informace je ovšem třeba zachytit odděleně od morfologického slovníku, a to především z důvodu nemožnosti objektivního rozhodnutí o zařazení lemmatu. Každá aplikace, nebo i každý uživatel, může mít svá přídavná data.[4]

6 Závěr

Chápeme-li pojem slovo jako řetězec znaků s určitým významem, měly by být různé zápisy téhož slova reprezentovány stejně. Vzhledem k obtížnému stanovení objektivního kritéria pro výběr takového reprezentanta jsme zavedli **vícenásobné lemma** jako množinu všech možných variant zápisu daného slova. Variantám říkáme **mutace**, neboť pojem varianta se v lingvistice používá i v jiných, navíc často různých kontextech. Naznačili jsme, jakým způsobem lze nově zavedené koncepty použít v lingvistické praxi.

Nejbližším plánem je jejich implementace a ověření jejich užitečnosti v praxi.

7 Poděkování

Příspěvek vznikl na základě grantu P406/2010/0875 Grantové agentury ČR a výzkumného záměru MŠMT ČR číslo MSM 0021620838.

Reference

1. Mluvnice češtiny 2, Academia, Praha 1986.
2. Český národní korpus - SYN. Ústav Českého národního korpusu FF UK, Praha 2010. Dostupný z WWW: http://www.korpus.cz.
3. Jaroslava Hlaváčová: Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Disertační práce, MFF UK v Praze, 2009.
4. Jaroslava Hlaváčová: Slovní varianty a morfologická anotace korpusů in Grammar & Corpora / Gramatika a korpus, pp. 161–168, Academia, Praha 2008
5. Karel Kučera: Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora, in Computer Treatment of Slavic and East European Languages, pp 121–125, Slovak Academy of Sciences 2007
6. Jana Marie Tušková: Variantní a dubletní tvary v současné deklinaci apelativních feminin. Spisy Pedagogické fakulty MU, sv. č. 98, Masarykova univerzita, Brno, 2006

[4] Tyto nemorfologické informace je samozřejmě možné přiřadit i jiným než vícenásobným lemmatům.