Morfologie češtiny znovu a lépe

Jaroslava Hlaváčová a David Kolovratník

Univerzita Karlova v Praze
Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta
Malostranské náměstí 25
11800 Praha
Česká republika

Abstrakt *Nový systém pro automatickou morfologickou analýzu češtiny [1], který zde prezentujeme, vychází ze současného pražského systému. Kromě nové implementace stávajícího slovníku pomocí konečného automatu přidává algoritmy na rozpoznání neznámých tvarů. V současné fázi je to především využití seznamu předpon, které lidé připojují bez větších omezení ke slovům a činí je tak pro automatické metody nerozpoznatelnými. Stačí předponu rozpoznat, odtrhnout a analyzovat zbylý řetězec. Výsledky této analýzy lze pak jednoduše aplikovat na původní tvar s předponou. K rozpoznání neznámých vlastních jmen používáme heuristiku. V příspěvku stručně popíšeme stávající pražský systém a novou implementaci včetně zárodku guessru.*

1 Úvod

Automatická morfologická analýza češtiny již byla řešena, dokonce několikrát. Nejznámější nástroje vznikly v Praze [3] a v Brně [2]. Oba systémy využívají rozsáhlé morfologické slovníky, na jejichž základě jsou schopny rozeznat většinu slovních tvarů vyskytujících se v českých textech. Počet nerozpoznaných slov, tedy slovních tvarů, které nelze odvodit ze slovníků, činí 2 až 3% (viz [4]). Rozšiřování slovníku většinou už nepomůže, naopak zvyšuje homonymii, která potom způsobuje problémy v dalších fázích zpracování textů, především při desambiguaci. Řešením je spíše použití tzv. guessrů, založených na vlastnostech českého tvarosloví, které umožňují s velkou přesností odhadnout charakteristiky (tj. lemma a hodnoty morfologických kategorií) nerozpoznaných slovních tvarů. V článku navrhujeme guesser pro rozpoznání tvarů odvozených pomocí předpony. Využíváme při tom vysoké pravidelnosti českého jazyka při tomto velice produktivním způsobu odvozování.

2 Reprezentace morfologického slovníku

Základní myšlenky implementace vycházejí z reprezentace morfologického slovníku použité J. Hajičem [3]. V následujícím stručně představíme jeho řešení.

Popis morfologie má tři části: (1) slovotvorné vzory, (2) tvaroslovné vzory a (3) vlastní morfologický slovník. Heslo slovníku obsahuje **slovotvorný začátek**, jméno jeho slovotvorného vzoru a **společné lemma**. Pomocí slovotvorného vzoru se generuje seznam (nějak příbuzných) **tvaroslovných začátků** s tvaroslovnými vzory, které v dalším kroku slouží k vytvoření konkrétních slovních tvarů. Každému slovnímu tvaru je na základě tvaroslovného vzoru přiřazena morfologická značka a základní slovní tvar (lemma). Uvedené vztahy přibližuje příklad na obrázku 1.

Kromě výše popsaných hesel slovník obsahuje i konkrétní slovní tvary, které nelze vygenerovat pomocí vzoru (např. substantivum *kůň* s nepravidelným skloňováním), a k nim příslušné morfologické značky.

Odvozování systémem vzorů spolu se slovníkem popisuje všechny možné slovní tvary a k nim příslušné značky a lemmata. Pomocí zmíněných tří částí popisu morfologie lze tedy všechny slovní tvary vygenerovat. To se však nedělá — bylo by to neefektivní. Již původní popis má dobré vlastnosti, které umožňují kompaktní reprezentaci, zachovávající zvlášť začátky a zvlášť vzory, snadno využít k analýze. Tvary se doplňují za pomoci (upravených) vzorů ze slovotvorných začátků až při samotné analýze za běhu.

3 Morfologická analýza

Morfologická analýza je vlastně funkce, která každému slovnímu tvaru přiřadí množinu dvojic [lemma, morfologická značka]. Implementace tedy musí umět rozpoznat slovní tvar a na výstup poslat množinu všech možných příslušných dvojic.

3.1 Implementace analyzátoru

Všechny slovní tvary jsou v analyzátoru uloženy ve formě **trie**[1]. Kdybychom však do trie uložili rovnou celé slovní tvary, bylo by to příliš neefektivní. Z toho

[1] Trie chápeme jako (1,n)-četný zakořeněný strom. Hrany stromu jsou ohodnoceny znaky (případně řetězcem, dojde-li je kompresi). Uzlu stromu lze přiřadit slovo, které vznikne sřetězením znaků na cestě od kořene k danému uzlu. Taková slova označená příznakem tvoří množinu klíčů. Ke klíči je v trie uložena také hodnota.

důvodu se do něj ukládají jen slovotvorné začátky převzaté ze slovníku. Z nich potom vedou dva odkazy. Jeden na složené schema (viz další odstavec), podle kterého se generují slovní tvary, druhý na řetězcovou instrukci, která ze slovotvorného začátku odvozuje **společné lemma** (to může později sloužit jako základ pro lemma analyzovaného tvaru). Řetězcová instrukce zde kompaktně reprezentuje společné lemma uvedené spolu se slovotvorným začátkem ve slovníku. Využívá toho, že si oba řetězce jsou často velmi podobné. Slovotvorný začátek je uložen v trie celý a bude při analýze k dispozici jako výchozí řetězec pro instrukci.

Složené schema v sobě kombinuje oba stupně vzorů. Nahrazuje tak dvoukrokové (viz obrázek 1) odvození jednokrokovým. Každému slovotvornému schematu odpovídá jedno složené. To je organizováno jako hešovací tabulka. Jejími klíči jsou spojení konců pravidel slovotvorného vzoru a příslušných konců vzorů tvaroslovných (například -zen+ej, -zen+ého, -zen+ému, ...).

Hodnotou vloženou ke klíči je opět dvojice odkazů. Jeden identifikuje skupinu značek, které popisují morfologické vlastnosti tvaru, druhý řetězcovou instrukci, která odvozuje lemma ze základu lemmatu. Základem je buď společné lemma nebo slovotvorný začátek. Jeden klíč může být vložen vícekrát s různými dvojicemi odkazů. Všechny výskyty přispívají k celkovému obrazu, doplňují tedy bohatost tvarosloví.

Řetězcová instrukce je trojice [*počet, řetězec, volič základu*]. Slouží ke kompaktnímu uložení řetězce, který sdílí začátek s nějakým známým, zvnějšku získaným, základem. Nový řetězec se získá (odvodí, spočítá) ze základu odmazáním *počtu* znaků od konce a připojením *řetězce*. Jednobitový *volič základu* vybírá jeden ze dvou možných základů. Toho se využívá pro odvození lemmatu — může vycházet ze společného lemmatu nebo slovotvorného základu. Například instrukce [3, „zení“, *společné lemma*] uložená ve složeném schematu ze společného lemmatu (známého základu) „ochladit“ odvodí lemma pro tvary podstatného jména „ochlazení“ odtržením tří znaků a připojením konce „zení“. Společné lemma však bylo dříve spočteno ze slovotvorného základu instrukcí [0, „dit“, *slovotvorný základ*]. Ta je uložena v trie u slovotvorného základu.

3.2 Vyhledání tvaru

Navržené rozdělení datových struktur umožňuje snadné nalezení tvaru. Tvar se postupně znak po znaku vyhledává v trie. V každém uzlu se testuje, zda je v něm nastaven příznak konce slovotvorného začátku a odkaz na složené schema. Pokud ano, hledá se dosud nepřečtený zbytek tvaru v hešovací tabulce prodloužení složeného schematu. Najde-li se, je již přečtená část slovotvorným základem. V hešovací tabulce se potom zjistí množina značek náležících k tvaru. Tímto způsobem se vyhledají všechny možné analýzy každého slovního tvaru.

Lemma je sestaveno pomocí řetězcové instrukce nalezené ve složeném schematu. Ta vychází buď ze slovotvorného začátku (ten je určen rozdělením tvaru na část nalezenou v trie a na zbytek) nebo ze základu lemmatu. Základ lemmatu se odvozuje ze slovotvorného základu řetězcovou instrukcí nalezenou v trie.

Dále tento algoritmus nazýváme **základní analýzou**. Ta je rozšiřována o schopnost rozpoznat předpony.

4 Předpony

4.1 Předpony podchycené ve slovníku

Ve slovníku nejsou předpony obecně nijak podporovány, kromě předpon negace (*ne-*) a superlativu (*nej-*). Negování i stupňování se provádí připojením příslušného morfému před slovo. Proto lze říct, že jistá podpora předpon ve slovníku je. Je však specializovaná.

Zvláštní podpora pro morfémy negace a superlativu ve slovníku má své opodstatnění: (1) negace a stupňování je potřeba i v případě, že není k dispozici jiná podpora předpon (třeba ta níže navrhovaná), (2) slovník přesně říká, se kterými slovy lze morfémy superlativu a negace kombinovat, zatímco dále navrhované řešení předpon omezuje kombinace jen podle slovních druhů a nakonec (3) je to důsledkem toho, že o morfémech superlativu a negace se dozvídáme jinak než o ostatních předponách (slovník versus seznam předpon).

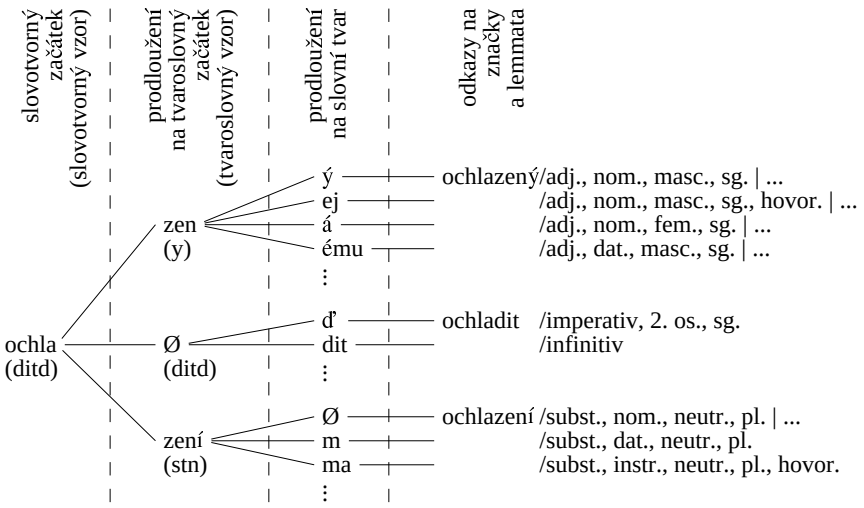
Předpony jsou v češtině běžnou a velmi užívanou součástí jazyka. Analyzátor, který by je zcela ignoroval, by byl stěží použitelný. Jsou tedy podporovány i bez navrhovaného rozšíření. Slova s předponou jsou uvedena ve slovníku jako samostatná hesla, bez uvedení slovotvorných souvislostí. Slova „čistit“ a „vyčistit“ tedy nejsou nijak propojena. Výhodou je opět zejména přesnost (správnost a citlivost) popisu.

4.2 Předpony zavedené seznamem

Námi navrhované rozpoznávání předpon má v mnohém opačné vlastnosti. Předpony jsou vypsány v samostatném seznamu zcela nezávisle na slovníku. Předpona se uvádí s **příznaky**, které říkají, se kterými slovními druhy se může pojit, např.

```
-dobro N                    -dopo NV
-dolno N                    -dovy NV
```

Obrázek 1. Představa odvozování slovních tvarů ze slovotvorného začátku



Příznak N znamená, že uvedená předpona dovoluje spojení se substantivy a adjektivy, příznak V se slovesy.

Mezi předponou a slovníkovým heslem není žádná jiná vazba než příznaky slovního druhu. Možnosti spojování nejsou nijak jinak omezeny. To se může jevit jako chybné řešení, protože s různými slovy lze kombinovat jen některé předpony. Nicméně to není problém, protože morfologická analýza si neklade za cíl vymezovat slovní zásobu jazyka („správná“ ani aktivně užívaná slova). Naopak, je snaha analyzovat co nejvíce slov.

Seznam předpon přichází na řadu až tehdy, když základní analýza pomocí slovníku tvar nerozpozná. Řešení pomocí seznamu umožňují použít dvě důležitá pozorování. Zaprvé je to skutečnost, že tvar se připojením předpony (většinou) nemění uvnitř[2]. Slovotvorný začátek, tvar i lemma se změní pouhým předřazením písmen předpony na začátek. A za druhé se nemění morfologické značky. Nemuselo by tomu tak být. Například předpony mění vid sloves, opět ale můžeme spoléhat na to, že všechny běžné případy jsou zahrnuty ve slovníku. Použití předpon má skutečně jen nabídnout řešení netradičních a nepodchycených slovních tvarů.

Zpočátku jsme chtěli používat analýzu pomocí předpon i u slov, která základní morfologická analýza rozpozná. Může se totiž stát, že některé tvary jsou homonymní s jinými, které slovník neobsahuje. Po prvních experimentech jsme ale od toho upustili, protože vycházelo velké množství falešně rozpoznaných tvarů. Například slovo „ves“ bylo rozděleno na předponu „v“ a substantivum „es“ (plurál od „eso“). Výsledná analýza potom běžnému tvaru „ves“ přiřadila navíc i nic neznamenající lemma „veso“. Podobných „veselých“ analýz se objeví celá řada.

4.3 Implementace předpon

Podmínkou pro rozpoznávání předpon je, že se při kompilaci nachystalo **trie předpon** (je volitelné). Trie se připravuje ze seznamu předpon pro odvozování. Ten se zadává jako zvláštní textový vstup kompilátoru.

Ačkoli trie začátků a trie předpon jsou tentýž typ datové struktury, staví a používají se nezávisle. Spojení by představovalo připojení trie začátků za všechny koncové uzly trie předpon. Protože by napojení bylo všude stejné, nepřineslo by analyzátoru žádnou dodatečnou informaci.

5 Analýza krok za krokem

5.1 Analýza čísel a interpunkce

V prvním kroku se morfologická analýza vypořádává s interpunkcí a posloupnostmi arabských číslic volitelně s desetinnou tečkou. Tyto jednoduché případy se řeší přímo, bez použití slovníku. Ostatní slovní tvary se zpracují následovně.

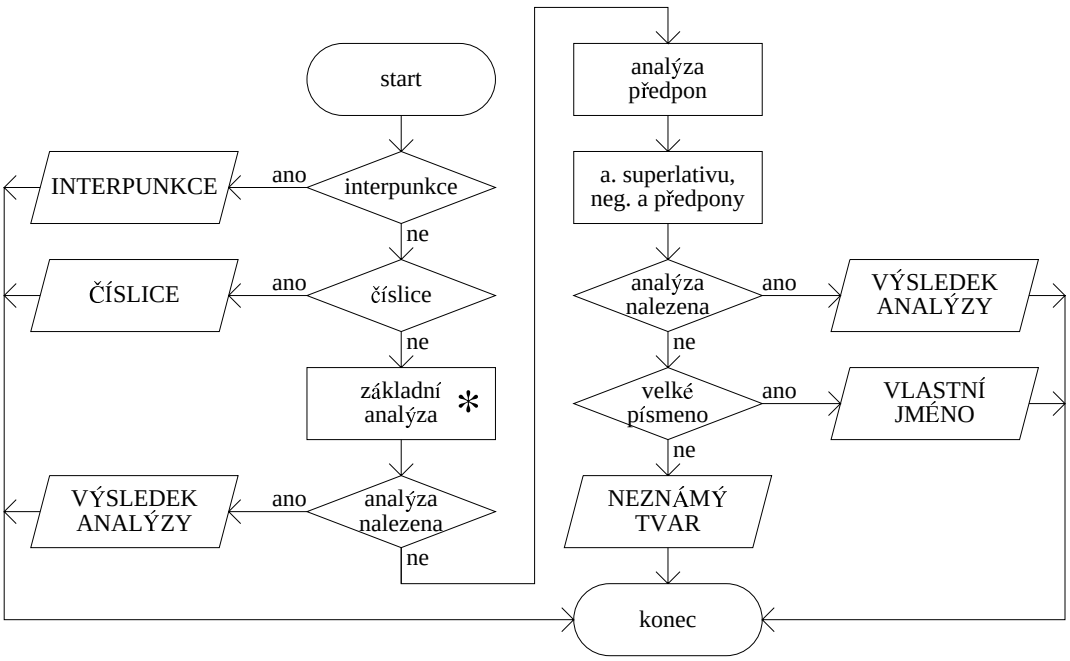
5.2 Základní analýza

Nejprve se řetězec tvaru použije celý tak, jak je, k sestupu do trie, jak bylo popsáno výše. To může vyústit

– v nalezení jednoho či několika lemmat spolu se značkami náležejícími ke tvaru, nebo

[2] Tvary, kde ke změně dochází, jsou již zpravidla ve slovníku zachyceny.

Obrázek 2. Vývojový diagram průběhu analýzy



∗ včetně analýzy případných morfémů negace a superlativu

– v neúspěšnou analýzu (nenalezení žádného slovního tvaru v trie).

Základní analýza je procedura využívaná opakovaně v následujících krocích.

5.3 Negace a stupňování

Pokud tvar začíná morfémem negace a/nebo stupňování, jsou tyto morfémy rozpoznány a na zbylou část tvaru je vždy znovu spuštěna základní analýza.

Uvažují se dva morfémy samostatně a jedna jejich kombinace:

– negativní morfém „ne“, který neguje,
– morfém „nej“, který tvoří z komparativu superlativ, a
– kombinace obou tvořící negované superlativum.

Například slovo „nejnepořádnější“ je rozděleno a analyzováno čtyřmi způsoby:

1. 0 + nejnepořádnější
2. ne + jnepořádnější
3. nej + nepořádnější
4. nejne + pořádnější

První rozdělení odpovídá algoritmu základní analýzy a provádí se vždy (už v předchozím kroku). Další jsou podmíněny přítomností písmen morfémů na začátku slova. V tomto případě uspěje jen poslední analýza, protože jedině komparativ „pořádnější“ je rozpoznán základní analýzou (za předpokladu, že „nepořádnější“ není ve slovníku jako samostatné heslo).

Uspěje-li vyhledávání nějakého rozdělení, může nebo nemusí být uznáno za platnou analýzu. Záleží na slovním druhu, zda jej lze kombinovat s uvažovanými morfémy. Po vyhledání konce tvaru, tedy když je znám slovní druh a značky, se filtrují výsledky.

Spojení morfémů s tvarem může být uznáno ze dvou důvodů. Zaprvé takové spojení mohou naznačovat žolíky ve značkách na pozicích popisujících negaci či stupeň a zadruhé může být obhájeno slovním druhem, dovoluje-li danou kombinaci. Negaci lze použít na afirmativní substantiva, adjektiva a slovesa, superlativa pak lze tvořit z komparativu adjektiv a adverbií.

Jsou-li rozpoznány jak negativní morfém tak morfém pro superlativum, musejí oba nalézt oporu v nějakém pravidle (viz předchozí odstavec). Když nenajdou, analýza se zahodí. Jinak je považována za přijatelnou. Značky se upraví, aby odpovídaly přítomným morfémům.

5.4 Odvozování předponou

Nebyla-li dosud nalezena žádná platná analýza, mohlo by se jednat o slovo odvozené předponou. Dojde-li na

analýzu odvozovaní předponami, hledá se od začátku tvaru předpona v trie předpon. Je-li nějaká nalezena, pokračuje se hledáním zbytku zadaného tvaru v trie začátků a dále běží základní analýza. Najde-li se zbytek jako slovní tvar, prověřuje se, zda jde předpona se slovním druhem dohromady. Pokud ano, je analýza přijata, jinak je odmítnuta. Kombinace zkratek s předponami se nepřipouští.

Analýza pokračuje hledáním delších předpon (a kratších tvarů). Ty by mohly představovat jiné dělení (či odvození) slova.

Analyzátor předpon uvažuje jen jedinou předponu. Vícečetné kombinace se nehledají. Přícházejí-li v úvahu, musí být zadány jako samostatná předpona v seznamu. Tím se navíc podchytí správné uspořádání předpon. Například můžeme „dovysekat trávu před garáží“. Neříká se ale „vydosekat“.

Poté, co se vyčerpají možnosti jedné předpony, zkouší se ještě jedna cesta. Slovo by totiž mohlo být utvořeno předponou v kombinaci s morfémem negace či stupňování.

Morfém superlativu může předcházet negativnímu morfému a jeden nebo oba mohou předcházet odvozovací předponě a ta konečně bude následována tvarem (například „nej—ne—dovy—sekanější“). Slovní druh tvaru musí korespondovat se všemi nalezenými morfémy (mohou být uznány i na základě žolíků ve značce) i předponou. Není tedy dovoleno např. spojení morfému „nej“ se substantivy. Opět se hledá jen jediná předpona.

5.5 Přehled analýzy

Obrázek 2 znázorňuje vývojový diagram průběhu analýzy.

6 Shrnutí a výhledy do budoucna

Představená nová verze morfologické analýzy již byla implementována. V základní formě poskytuje stejné výsledky jako zmiňovaná pražská verze morfologické analýzy. Je třeba ještě provést vyhodnocení výsledků předponového guessru, tj. zda a nakolik se zlepší rozpoznání slovních tvarů, které základní analýza na základě slovníku rozpoznat neumí. Dále plánujeme rozšířit guesser na koncovky a přípony, které jsou v českém jazyce také velmi produktivní.

7 Poděkování

Tato práce byla podpořena granty Informační společnosti č. 1ET101120503 a 1ET101120413 poskytovanými Grantovou agenturou AV ČR a grantem č. 100008/2008 poskytnutým Grantovou agenturou UK.

Reference

1. http://ufal.mff.cuni.cz/morfo/
2. Sedláček R.: Morphological analyser of Czech - ajka. http://nlp.fi.muni.cz/projekty/ajka/
3. Hajič, J.: Disambiguation of Rich Inflection. (Computational Morphology of Czech) Praha, Karolinum 2004.
4. Hlaváčová, J.: Morphological Guesser of Czech Words. Proc. TSD 2001, Springer-Verlag Berlin Heidelberg 2001. str. 70-75.