K počítačové morfologické analýze češtiny

Pavel Šmerk

Centrum zpracování přirozeného jazyka
Fakulta informatiky
Masarykova univerzita

`http://nlp.fi.muni.cz/ma`

3. 12. 2010

Nevýhody stávajícího formátu dat morf. analyzátoru

- současný stav: „pražský“ a „brněnský“ analyzátor
- i přes dílčí odlišnosti je organizace dat v principu shodná
  - slovník základů + soubor vzorů, množin koncovek se značkami
  - pro každý základ jsou specifikovány vzory, připojením jejich koncovek se získají tvary se značkami
  - základy i koncovky jsou řetězce, které se jen skládají k sobě
- z posledního plyne zásadní nevýhoda: redundance popisu
  - *Luděk/Luďka*, *Staněk/Staňka*, *vrah/vraha*, *medvídek/medvídka* atp. se skloňují stejně či podobně, ale kvůli drobným odlišnostem vyžadují vlastní řešení (v Brně extra vzor, v Praze vzor či výjimky)
- redundance vede k nekonzistenci při doplňování či opravách
  - příklad (vše m. živ.): doplnění hovorového Gsg *-a*: *muža*
    - 217 vzorů, tedy nutno automaticky, Gsg *-e* $\rightarrow$ *-a*
    - ovšem u cca 10 vzorů je *-ě* místo *-e*; u *strašpytel* a *neumětel* *-a* už je
    - kontrola obtížná, ne-li nemožná

Nevýhody stávajícího formátu dat morf. analyzátoru

- takových nekonzistencí nejrůznějších druhů je celá řada
  - (v Praze předpokládám podobný stav)
- na druhou stranu, jde vesměs o okrajové věci
  - nikdo to „nereklamuje“, vyvstalo až při přeuspořádání
- takže jakékoli řešení (ať už prevence, nebo lék) je *příliš* drahé, protože náklady budou velké, ale reálný přínos bude malý

- méně závažnou nevýhodou je formální, strukturní nekonzistence
  - tedy možnost popsat tutéž věc různými způsoby
  - důsledek skutečnosti, že struktura dat nemá interpretaci
    - původně byla daná hranice mezi intersegmentem a koncovkou a koncovkové množiny byly tvořeny podle pevných pravidel, teď částečně technické řešení

Nový formát dat

- zůstává slovník a soubor vzorů
- snaha oddělit pravidelné (vzory, program) a nepravidelné (slovník)
- snaha o „interpretovatelnost“
  - různé cesty k témuž výsledku mohou mít odlišnou interpretaci
  - ovšem pouze za předpokladu, že to vůbec chci nějak interpretovat
- základy (`slon:pán`) ve slovníku, koncovky uspořádané do vzorů

```
pán k1gM
        nSc1    0
        nSc2    a
```

- ale po jejich spojení (`slon-0`) se aplikují předdefinovaná pravidla
  - triviálně je potřeba odstranit `-` a `0`
  - `ňe` → `ně`: `tuleň-e` → `tuleňe` (nebo `tulen-ě`) → `tuleně`
  - *Ábel* × *ďábel* ⇒ `Ábel` × `ďáb.el`: `.eC-0` → `eC-0`, `.eC-V` → `C-V`
  - `vlk-i` → `vlc-i` (ale také `pán-i` → `páň-i` → `páňi` → `páni`)

Nový formát dat

- použitelnost koncovek lze omezit podmínkou na konec základu
  - např. `nPc6 ech, ích/[ghk]|ch` (ve vzoru)
- už jen toto málo stačí pro popis mnoha dosud oddělených vzorů
  - `Luď.ek-0` → `Luďek-0` → `Luďek` → `Luděk`
  - `pejs.ek-ích` → `pejsk-ích` → `pejsc-ích` → `pejscích`

- dále
  - tvorba vzorů děděním: `soudce:muž` + `e` pro `nSc1` a `nSc5`
  - omezené vzory: `despota:pán_nP` + singulárové koncovky
  - odvození z více vzorů: `filozof:pán,-ové`, `dřevokaz:pán,+muž`
- příklad rozdílné interpretace téhož výsledku *g* ⇒ Npl jen *g-ové*
  - `nPc1 i/[^g], ové/` — tvary typu **mázi* systémově nemožné
  - `mág:filozof` — shodou okolností takové slovo aktuálně neexistuje

Nový formát dat

- dále
  - hovorové tvary: Npl (a Vpl) *?učitelové*, ale **pokrytcé*
    - obecně: 1) ne/lze *-é*; 2) které z koncovek *-i* a *-ové* jsou spisovné
    - `filozof:pán,<-ové`; `občan:pán,<-é`; `akrobat:pán,<-i,+-é`
    - (bez < bych musel substandardní koncovky definovat ve vzorech `-é`)
  - více slovních základů, nepravidelné tvary (tedy slovník)

    ```
    přítel:muž,<-é
            <přátel:muž_nP,<-é
            <přátel-0 nPc2
    ```

    - `wH` tvary dokládá Google, jen spisovné tvary by byly bez <
    - pořadí ovlivňuje výsledek (dosud data neuspořádaná)
    - vyjadřuje, co je základní a co specifické (dosud tvary rovnocenné)
    - (Google: *přítelů* < *přátelů* < *přátel*, podobně i pro nepřítele)
    - `pejs.ek` je ve „struktuře“ vždy stejný, ale lze i

      ```
      pejsk:pán
              pejsek-0 / pejsek / pejsek:pán nSc1
      ```

    - ovšem zde nelze <, nemluvě o tom, že by to komplikovalo data

Nový formát dat

- dále
  - zachycení rozdílů mezi zápisem a výslovností
    `Smith[t:pán,-ové`
    `+Smith[s:muž,-ové`
- dosavadní umožňuje popis pomocí tradičních mluvnických vzorů, případně s upřesněními, bez nichž se ale neobejdou ani mluvnice
- ztotožňování shodných koncovek
  - falešný vzor `$shoda`

`c1`	`c5`
`k1gMnS\Kc3`	`c6`

  - `Marcel:pán,<-ové,muž_nSc5` ⇒ *Marceli* i *Marcelu*
  - `despot:žena_nS,-ovi,pán_nP gM`
  - `gigol:město_nS,+-ovi,pán_nP gM` (`ě/!gM`)
- (skládání značky, implicitní značka, implicitní vzor, ...)

Od slovníku vzorů ke slovníku rysů

- lze si ale myslet, že lidé si nepamatují vzory, ale ohýbají slova podle jiných vlastností: sémantických, strukturních či hláskových
  - u vlastních jmen je preferována `-ové` před `-i`
  - slova odvozená příponou `tel` jsou `muž,<-é`
  - životná maskulina zakončená v Nsg na *d* se skloňují tvrdě
- skloňování určované slovotvornými příponami
  - `=tel:muž,<-é` do souboru vzorů
  - ve slovníku pak postačí `uči=tel` nebo např.
    `pří=tel`

    ```
    <přá=tel        nP
    <přá=tel-0      nPc2
    ```
  - `=í:adj` ⇒ `krejč=í`
  - pokud sufixy připustím i v seznamu vzorů, mám derivaci
  - např. `k1gM:=%ov`, kde `k1gM` bude „předek“ mužských vzorů

Od slovníku vzorů ke slovníku rysů

- implicitní pravidla: typické, pravidelné chování podle zakončení základu nebo jeho rysů vyjádřených značkou ve slovníku

```
$k1gM
    \Ko      město_nS,+-ovi,pán_nP,muž_nP/$M|i,-ové
    s/qJO    muž,<pán_nPc[67],+pán_nPc4
```

- `$M` a pod. jsou zkratky za regulární výrazy (měkké souhlásky)
  - také definované v datech pomocí falešného vzoru
- pak ve slovníku

```
gigolo   k1gM
Klaus    k1gMqJOP
```

Data v novém formátu v číslech

- zatím zpracována jen životná maskulina
- nejčastější popisy slov ve slovníku (z 19975 lemmat)

# lemmat	% z celku	příklad
13871	69.17	`gaučo k1gM`
2207	11.01	`Ionesc[ko k1gMqJOP`
1654	8.25	`Severo+evrop=an`
683	3.41	`Mario k1gMqJO`
440	2.19	`kok.eš:-ové k1gM`
321	1.60	`sob.ěk:-i k1gM`
146	0.73	`uniat:-é k1gM`

- i z těchto částečných dat (>100 lemmat) je vidět, že pro >90 % životných maskulin stačí část značky, nebo i jen vyznačení přípony
- popis „vzorů“ je 13x menší než odpovídající část původních dat
  - pokud se nepočítají části společné s jinými rody, tak dokonce 24x

Vlastnosti a přínos nového formátu

- významná redukce dosavadní redundance
- výrazně vyšší „lingvistická přijatelnost“
  - slova lze řadit k tradičním vzorům
  - hranice mezi kmenem a koncovkou může odpovídat mluvnicím
  - lze zachytit pravidelné hláskové změny (alternace)
  - formát umožňuje slovotvorné vztahy a morfematickou analýzu
  - umožňuje rozlišit pravidelné, typické jevy od okrajových, u kterých navíc stačí popsat jen odchylku od většinového chování
- různé zápisy téhož lze zpravidla i různě interpretovat
- jednotlivé možnosti jsou vzájemně nezávislé, lze tedy některé nepoužívat
- celkově prokazuji, že pro popis dat nejsou potřeba žádná „technická“ řešení, že není nutný zásadní rozdíl mezi lingvistickým popisem a popisem vhodným pro počítač