K počítačové morfologické analýze češtiny

Pavel Šmerk

Centrum zpracování přirozeného jazyka
Fakulta informatiky
Masarykova univerzita

http://nlp.fi.muni.cz/ma, aurora:/nlp/projekty/ajka
tyto slidy: http://www.fi.muni.cz/~smerk/majka

26. 9. 2018

Morfologická analýza

- nejnižší rovina zpracování jazyka v textové podobě
  - (český text lze na slova dělit snadno, až na *gen.*, *byl-li*, *oč/očs* ap.)
- morfologická analýza by měla pro každý slovní tvar vrátit
  - základní slovní tvar (lemma, položka slovníku)
  - možné gramatické významy („značky“) — hodnoty relevantních gramatických kategorií jako např. slovní druh, pád, číslo, osoba atd.
  - např. pro slovní tvar *stroj* lze očekávat
    - *stroj*: podst. jm., mužský neživotný, singulár, nominativ/akusativ
    - *strojit*: sloveso, 2. os. j. č., rozkazovací způsob, nedokonavé
- + syntéza, lemmatizace (vracím jen lemma), . . .
  - naopak nejde o rozklad na morfémy, jak by to někdo mohl chápat
- problém má tři části („osnova“ zbytku slidů)
  - jakou informaci chceme/potřebujeme zachytit, popsat (zde s. 5–6)
  - jak si tuto informaci, tato data budeme organizovat (s. 7–9, 11–22)
  - jak implementovat analýzu či syntézu nad těmito daty (s. 10, 23–27)

Značky

- gramatická informace je reprezentována řetězcem znaků, *tag*em
- poziční systém: značka kóduje jen hodnoty kategorií
  - kategorie je jednoznačně určena *pozicí* ve značce
  - pražský systém — 16 pozic: slovní druh, specifikace, rod, číslo, pád, přivl. rod, přivl. číslo, osoba, čas, stupeň, negace, slovesný rod, volné (13, 14), styl, vid
  - `NNIS4-----A-----`
    - substantivum, obyčejné, muž. neživ., singulár, akuzativ, afirmace
  - http://wiki.korpus.cz/doku.php/seznamy:tagy
- atributový systém: dvojice atribut–hodnota bez ohledu na pořadí
  - brněnský systém — podobné kategorie i hodnoty jako pražský
    - např. atribut c znamená pád a může nabývat hodnot 1 až 7
  - `k1gInSc4` = substantivum, muž. neživ., singulár, akuzativ
    - nezachycena specifikace a afirmace
  - výhody: přehlednější, úspornější, snadno rozšiřitelný, čitelné RE
  - http://nlp.fi.muni.cz/projekty/ajka/tags.pdf

Značky

- „Heterogenní“ systém (Bratislava)
  - vychází z pozičního systému, prázdné pozice jsou vynechávány
  - první znak udává slovní druh, ostatní kódují kategorii i hodnotu
    - tedy odpovídají dvojici znaků atributového systému
  - pořadí je závazné, ale každý znak je užit jen v jednom „významu“
    - pořadí by tedy mohlo být i volné, znaky se ovšem rychle vyčerpají
  - `SSis4`
    - substantivum, subst. deklinace, muž. neživ., singulár, akuzativ
  - výhodou jsou nejkratší značky, na obrazovku se mi vejde víc info
  - nevýhodou je malá rozšiřitelnost a složitější programové zpracování
  - http://korpus.juls.savba.sk/morpho.html
- Jiný typ jazyka, zcela jiné řešení: BNC tagset
  - pevná množina několika desítek „hotových“ značek, např.
    - `AJ0` Adjective (general or positive) (e.g. good, old, beautiful)
    - `AJC` Comparative adjective (e.g. better, older)
    - `AJS` Superlative adjective (e.g. best, oldest)
    - `PNX` Reflexive pronoun (e.g. myself, yourself, itself, ourselves)
  - http://www.natcorp.ox.ac.uk/docs/c5spec.html

Co chceme popisovat

- na první pohled se to zdá být učivo prvního stupně ZŠ
- neshoda může být teoretická (lingvisté) i praktická (aplikace)
  - zejména je reálná: současné dva hlavní popisy *téhož jazyka*, pražský a brněnský nejsou „isomorfní“, vzájemně převoditelné
  - sjednocení se řešilo mnoho let, asi se s tím už všichni smířili
- různé možnosti lemmatizace
  - do jaké míry při určení základního tvaru zohlednit slovotvorbu/flexi
    - otcova ⇒ otcův/otec, učený ⇒ učený/učit, učení ⇒ učení/učit
    - nejstaršího ⇒ starý/nejstarší (vyhledávání: [věk] . . . člověk)
      - (a starší paní může být mladší než stará paní)
    - nebral ⇒ brát/nebrat (úplatky); nemalý ⇒ malý/nemalý
    - bakalářka z VŠMIE: pro online marketing se prý jednotné a množné číslo jmen považují za různá klíčová slova (detaily jsem nezjišťoval)
  - jak naložit s dubletami
    - mysli ⇒ myslet/myslit
    - kapitalismem ⇒ kapitalismus/kapitalizmus
  - o diachronii (všechen/všecken) a varietách (okno/vokno) nemluvě

Co chceme popisovat

- různé možnosti volby gramatických kategorií a jejich hodnot
  - které slovní druhy: zkratky, interpunkce, čísla, speciality (*cos*, *aby*)
  - které gramatické kategorie: druhy zájmen, číslovek, příslovcí, spojek, pád u předložky, životnost *koho*/*čeho*
  - jaké hodnoty kategorií: „duál“ (pes se 4 nohama), druhy zájmen ap.
- vše dosud uvedené je ale ještě to nejmenší
  - větším problémem je, jaká slova budou mít jaké značky
    - ke kterým všem slovním druhům mají patřit *a*, *ani*, *ať*, *až*, ...
  - největším problémem je stanovení pravidel pro určení slovního tvaru v konkrétním větném kontextu
    - může-li mít slovní tvar značky A, B a C, musí být jasné, kterou pro konkrétní výskyt zvolit, mezianotátorská shoda musí být co nejvyšší
    - viz např. konec http://nlp.fi.muni.cz/projekty/desman/, vzít 100 výskytů jednotlivých slov a zkusit, jestli pravidla vždy postačí
    - pokud mluvčí nejsou schopni pravidla spolehlivě aplikovat, je otázka, jestli tato odrážejí nějakou jazykovou realitu

Morfologický analyzátor ajka

- „původní“ řešení (Osolsobě 1996, Sedláček 1999+2005)
- princip organizace dat
  - a priori mám dané, které slovní tvary patří k sobě (viz dříve)
  - slovní tvary lemmatu se rozdělí na společný základ a „koncovky“
  - lemmata mající shodné množiny koncovek patří k témuž vzoru
    - *kluk* je jako *vlk*, ale ne jako *pes* či *slon*

1. p. j. č.	vl-k	p-es	slon-0
2. p. j. č.	vl-ka	p-sa	slon-a
3. p. j. č.	vl-ku	p-su	slon-u
3. p. j. č.	vl-kovi	p-sovi	slon-ovi
	...		
1. p. mn. č.	vl-ci	p-si	slon-i
	...		

  - ve skutečnosti mezi základem a koncovkou ještě intersegment
    - vl-k-0, p-es-0, slon-0-0; ... vl-c-i, p-s-i, slon-0-i; ...
    - ale to už je jen technické řešení, základní princip se nemění

Ukázka slovníku a definice vzorů

- slovník
  - formát lemma:vzor, ! lze negovat, % reflexiva tantum + poznámky

```
hanbit:barvit!%|793.1,167.1
zelený:nový!|148.1
osel:orel|180.1
...
```

- příklad definice vzoru
  - lemma vzoru + <intersegmenty> + seznam koncovkových množin

```
+barvit
        <i> NEWES717, NEWES744, konc44
        <en> NEWES710
        <il> NEWES705, NEWES778
        <ě> NEWES757
        <íc> NEWES759
        ...
```


Ukázka slovníku a definice vzorů

- příklad koncovkových množin
  - jména jsou arbitrární, generovaná nějakým programem
  - množina dvojic koncovka + jí odpovídající značka

```
=NEWES717
        {t, k5aImF}
=NEWES705
        {y, k5aImAgFnP}
        {i, k5aImAgMnP}
        {a, k5aImAgFnS}
        ...
```

- interpretace
  - z lemmatu odtrhnu první intersegment a koncovku vzoru, čímž dostanu slovní základ, k němu připojuju intersegmenty a koncovky
  - hanbit ⇒ hanb + -i-t
  - ⇒ hanb-i-t k5aImF, ..., hanb-il-i k5aImAgMnP, ...

Obecné statistiky

koncovky	83
intersegmenty	3.265
kmeny	389.793
značky	1.201
konc. množiny	1.340
vzory	1.838
generované tvary	6.294.591
včetně hovorových	11.693.520

Vesměs automaticky generovaná část slovníku

- Substantiva – deverbativa (32%)
- Adjektiva – posesiva mužská (12%), ženská (5%), deverbativa (64%)
- Slovesa – prefigovaná (78%)
- Adverbia – odvozená z adjektiv (96%)

Počet kmenů, vzorů a tvarů jednotlivých SD

Slovní druh	Kmenů	Vzorů	Tvarů	Včetně hovor.
Podstatná jména	131.776	778	967.231	1.217.442
Přídavná jména	170.771	69	3.831.134	8.167.371
Zájmena	199	104	2.150	3.035
Číslovky	217	44	1.699	1.699
Slovesa	42.720	758	2.014.122	2.155.125
Příslovce	41.587	71	146.244	146.247
Předložky	333	6	350	350
Spojky	195	2	213	213
Částice	251	1	264	264
Citoslovce	1.039	1	1.085	1.085
Zkratky	689	2	689	689

Počty vzorů podle počtu kmenů

Počet kmenů	Počet vzorů	Příklad vzorů
1	580	den, hůl, křest, vrzat
2	208	křemen (skřemen), líh (klíh)
3	120	okres (ples, expres)
4–10	345	
...	...	...
14071	1	nově
14199	1	nový
18634	1	otcův
33335	1	nesen
37689	1	stavení

Vztah ke klasickým vzorům – příklad

Mužský životný	Počet vzorů	Mužský neživotný	Počet vzorů
pán	44+28	hrad	49+14
muž	22+2	les	15+1
předseda	15+3	stroj	17
soudce	2	hrad/les	14+1
pán/muž	4	les/stroj	2
		stroj/hrad	6
výjimky	6		3
ind./adj./pl.t.	2+8+5		2+1+28
celkem	141		153

Systém vzorů – příklad

Klasický vzor *pán*:

- kmen se nemění – nom. pl. *-i, -ové* (slon), *-é* (občan), *-i* (docent), *-i, -é* (akrobat), *-ové* (filosof)
- samohl. alternace kmene – nom. sg./zbytek (pes), sg./pl. (přítel)
- souhl. alternace finály – *k-c* (vlk), *h-z* (vrah), *ch-š* (hroch), *r-ř* (doktor),
  r-ř (mistr), *g-z* (archeolog), *k-č* (člověk), *h-z-ž* (bůh)
- alternace finální skupiny – medvídek, daněk, Achilles, brontosaurus, génius
- cizí koncovka nom. sg. – Fero, Antonio

Princip analýzy nad uvedenými daty

- analyzované slovo $w_1 w_2 \ldots w_i = Z + I + K$
- základ Z, intersegment I i koncovka K mohou být nulové
  - např. `slon-0-0`, naopak `0-člověk-0, 0-lid-é`
- základem tedy může být ϵ, w_1, …, $w_1 \ldots w_i$
- pro každý základ $Z = w_1 \ldots w_n$ nalezený v seznamu základů se v jeho vzoru zkusí dohledat kandidáti na $w_{n+1} \ldots w_i = I + K$
- značky příslušné k nalezeným trojicím $Z + I + K$ jsou výstupem
- ve skutečnosti se ještě počítá s možnými prefixy `nej` a `ne` a postfixy, např. `s` v `Byls tam`?

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
  - (je to podobné jako mít konstanty přímo v programu)
  - příklad (vše m. živ.): doplnění hovorového Gsg *-a*: *muža*
    - 217 vzorů, tedy nutno automaticky, Gsg *-e* → *-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
  - (podobné problémy má i IJP či SSJČ, obecně cokoli tvořené ručně)
- 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, později jen technické řešení

Nový formát dat

- zůstává slovník a soubor vzorů
  - snaha oddělit pravidelné (vzory, program) a nepravidelné (slovník)
    - slovník: specifické pro jednotlivá lemmata, co si musím pamatovat
    - vzory: vlastnosti koncovek, program: pravidla pro skládání
  - snaha o „interpretovatelnost“
    - různé cesty k témuž výsledku mohou mít odlišnou interpretaci
    - ovšem za předpokladu, že to vůbec chci nějak interpretovat
- základy (`slon:pán`) ve slovníku, koncovky uspořádané do vzorů
  `pán k1gM`
  ```
  nSc1    0
  nSc2    a
  nSc3    u, ovi
  ...
  ```
- základy se spojují s koncovkami: `slon-0`, `slon-a`, ...
- odpovídající značky dostanu spojením části společné pro celý vzor a části specifické pro použitou koncovku: `k1gMnSc1`, ...

Nový formát dat

- po spojení základu s koncovkou (`slon-0`) se slovní tvar získá aplikací předdefinovaných pravidel
  - triviálně je potřeba odstranit `-` a `0`
  - `ňe` → `ně`: `tuleň-e` → `tuleňe` (nebo `tulen-ě`) → `tuleně`
    - na pořadí pravidel někdy nezáleží z hlediska výsledky, ale může záležet z hlediska mezivýsledku, zde např. „zvuková“ podoba
  - *Ábel* × *ďábel* ⇒ `Ábel` × `ďáb.el`: `.eC-0` → `eC-0`, `.eC-V` → `C-V`
    - (u `Luď.ek` lze tvrdit, že jde o kontext, u `ďáb.el` zjevně ne)
  - `vlk-i` → `vlc-i` (ale také `pán-i` → `páň-i` → `páňi` → `páni`)
- 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`

Nový formát dat

- dále
  - tvorba vzorů děděním:
    `soudce:muž`
    ```
    nSc1    e
    nSc5    e
    ```
    - možné koncovky se při tvorbě vzoru defaultně přepisují
    - pokud bych před část značky uvedl `+`, přidají se
  - omezené vzory: `despota:pán_nP` + singulárové koncovky
  - pomocné vzory pro koncovky:
    `-ové k1gM`
    ```
    nPc1    ové
    ```
  - 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
    - výhledově taky slovník, není to mnoho slov, ale jedna přípona
    - „výjimkou“ je totiž spíše *=tel*, než *datel*
    - *datel* se skloňuje stejně jako ostatní `k1gM` *-l*
  - 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 detailně zpracována jen životná maskulina
- nejčastější popisy slov ve slovníku z celkem 19975 lemmat k1gM
  - (komentář k tabulce je na další stránce)

# 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`

- popis „vzorů“ je 13x menší než odpovídající část původních dat
  - pokud se nepočítají části společné s jinými rody, tak dokonce 24x

Data v novém formátu v číslech

- i z těchto částečných dat (>100 lemmat) je vidět, že pro >90 % životných maskulin stačí část značky, nebo i jen vyznačení přípony
- to asi odpovídá realitě lépe než předchozí model dat
  - lidé si ke slovu nepamatují vzor, natož jeden z cca 2000
  - dítě umí skloňovat i s výjimkami ještě než jde do školy
  - stačí mi vědět, že gaučo je mužský životný a umím jej vyskloňovat
    - k1gM možná odvozuju z nějaké sémantiky, ale to už je celkem jedno
- ani u kok.eš-e si nepamatuju vzor, jen drobné upřesnění defaultu
  - navíc, *v principu* skutečně jde o slovníkovou informaci
    - (tedy nikoli vzor rozexpandovaný do slovníku)
    - já totiž nevím, proč to tak je, prostě to tak je v nějakém Zdroji, musím se to naučit a pamatovat si to: kokšové, nikoli kokši — a toto si musím pro každé takové slovo pamatovat zvlášť
    - a nebo to důvod má, ta slova mají něco společného (a nemám je naučená zvlášť), no a pak je potřeba to adekvátně popsat, tím lépe

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

Nový morfologický analyzátor majka

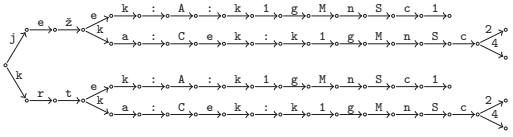
- `ajka` byla už příliš složitá, a proto v podstatě nerozšiřitelná [allt]
- využití přístupu popsaného v disertační práci Jana Daciuka
  - analýza je realizována pouhým vyhledáním tvaru v seznamu WLT
  - data jsou vlastně seznam dotaz:odpověď ve formátu

```
ježek:A:k1gMnSc1      ← ježek:ježek:k1gMnSc1
ježka:Cek:k1gMnSc2    ← ježka:ježek:k1gMnSc2
ježka:Cek:k1gMnSc4    ← ježka:ježek:k1gMnSc4
krtek:A:k1gMnSc1      ← krtek:krtek:k1gMnSc1
krtka:Cek:k1gMnSc2    ← krtka:krtek:k1gMnSc2
krtka:Cek:k1gMnSc4    ← krtka:krtek:k1gMnSc4
```

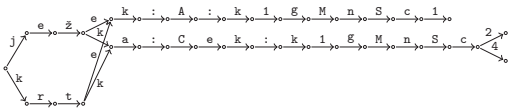
  - seznam lze chápat jako konečný jazyk ⇒ existuje pro něj DAFSA
    - musí být minimalizovaný, jinak by byl nepoužitelně velký (viz i dále)
    - lemma je potřeba kódovat, jinak by minimalizace nepomohla
    - Daciuk nabízí inkrementální tvorbu zachovávající minimalizovanost
  - (toto je nezávislé na předchozí části: z původních dat `ajky` lze generovat WLT, naopak z nových dat lze generovat data pro `ajku`)

Nový morfologický analyzátor majka

- deterministický automat neminimalizovaný



- deterministický automat minimalizovaný



- „analýza“ je jen rychlé a jednoduché procházení tohoto FSA
  - deterministický průchod dle „dotazu“ + dohledání všech „odpovědí“

Nový morfologický analyzátor majka

- obdobně data pro lemmatizaci, generování, segmentaci atp.
- lemmatizace: `krtek:A`, `krtka:Cek`
- generování: `krtek:A:k1gMnSc1`, `krtek:Cka:k1gMnSc2`
- nebo generování z lemmatu a značky: `krtek:k1gMnSc2:Cka`
- převod na původní strukturu: `krtek:C.ek-0`, `mužova:D=%ov-a`
  - až po aplikaci některých pravidel: `krtek:Cek-0`, `krtka:Ck-a`
- prefixy: `nemalý:CA:k2*, malý:Ane:A:k2*/malý:ACneA:k2*`
- pro čísla a složeniny (trojciferný, českopolský) gramatika
- FSA využitelné i obecně (frekvence slov v aplikaci Deriv) [judy?]
- obava z velkého seznamu (Gelbukh '03) není odůvodněná

Charakteristiky a výsledky analyzátoru majka

- statistické informace o (některých) slovnících

slovník	řádků	zdroj MB	slovník MB	bytů/řádek
w	13,609,590	186	3.3	0.240
w → l	14,101,767	240	4.0	0.287
w → l+t	80,303,929	2,478	4.4	0.054
w → w	957,464,060	19,993	6.1	0.006

- porovnání s morfologickým analyzátorem `ajka`

	velikost dat		čas v sekundách		
	`ajka`	`majka`	`ajka`	`majka`	poměr
analýza	3.1	4.4	**18.22**	**2.88**	**6.3x**
lemmatizace		4.0	16.76	1.57	10.7x
tvary		6.1	55.33	8.42	6.6x
diakritika		3.3	8698.80	1.61	5403x

- analýza 4.6x rychlejší proti pražskému analyzátoru Morfo (11 MB)
- `majka` je používána např. v Seznam.cz a projektech IS MU

Výhody a přínosy nového řešení

- naprosto zásadní výhodou je jednoduchost: průchod automatem je nezávislý na konkrétních datech, funkcionalitu rozšiřuji, případně měním datovými soubory, nikoli změnami kódu analyzátoru
  - výjimkou mohou být složeniny
  - obslužné kódy pro jednotlivé datové soubory jsou nezávislé
  - to vše je obrovský rozdíl například proti analyzátoru `ajka`
- příjemnou výhodou je samozřejmě výrazné zrychlení
  - přičemž se nejedná o okrajový problém, který by dosud jen nebyl dostatečně řešen
- teoretický přínos
  - naprosté oddělení popisu dat a analyzátoru
  - prokazuji, že pro realizaci počítačové morfologické analýzy jazyků, jako je čeština (s morfologií na konci slova), nejsou potřeba žádné speciální datové struktury či algoritmy

Související a navazující aplikace

- umožňuje perfektní hašování $X \to 1..|X|$
  - $\Rightarrow$ využito pro indexaci korpusů (obecně jazykových dat)
  - komprese dat srovnatelná se zip*, viz
- guesser
- derivační vztahy: derivancze
- desambiguace: test, pro uplatnění např. viz CzTenTen