Osolsobě, Klára

Automatické rozpoznávání a generování určitých číslovek a od nich odvozených číselných pojmenování na počítači

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 1995, vol. 44, iss. A43, pp. [31]-48

ISBN 80-210-1192-0
ISSN 0231-7567

SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 43, 1995

KLÁRA OSOLSOBĚ

AUTOMATICKÉ ROZPOZNÁVÁNÍ A GENEROVÁNÍ ČESKÝCH URČITÝCH ČÍSLOVEK A OD NICH ODVOZENÝCH ČÍSELNÝCH POJMENOVÁNÍ NA POČÍTAČI

V tomto článku se budeme zabývat automatickou morfologickou analýzou a syntézou českých číslovek a číselných pojmenování pomocí počítače. Výsledky naší práce jsou součástí algoritmického zpracování české formální morfologie pro účely automatické morfologické analýzy a zároveň prvním pokusem o přesah do oblasti slovotvorby a syntaxe.

1. Úvod

Číslovky jakožto slovní druh představují zajímavý lingvistický fenomén. Mluvnice je definují jako slova, která mají číselný význam a označují počet, pořadí atd. Na rozdíl od ostatních slovních druhů jsou číslovky přesně vymezitelné sémanticky a zároveň vykazují značnou pestrost právě v oblasti formální morfologie. Podle sémantických kriterií se dále dělí na jednotlivé druhy číslovek, které se buď řadí k adjektivním, popřípadě substantivním skloňovacím typům, a nebo mají vlastní skloňování. Některé druhy číslovek jsou nesklonné, mají charakter adverbií.

Sémantická jednoznačnost a omezený repertoár číslovkových kmenů na straně jedné a velká formální pestrost na straně druhé nás vedly při algoritmickém zpracování české formální morfologie číslovek ke kombinaci popisu formální morfologie a slovotvorby, který jsme u ostatních slovních druhů zavrhli pro nesnadnou formalizaci analogického postupu.

Algoritmický popis morfologie českých číslovek je součástí algoritmického popisu formální morfologie češtiny (Osolsobě, 1994 – aktuální verze), který spolu se strojovým slovníkem kmenů českých slov obsahujícím cca. 160 000 položek tvoří lingvistickou bázi programů pro práci s přirozeným jazykem (český automatický korektor – Osolsobě, Ševeček 1991, automatický lemmatizátor češtiny – Osolsobě, Ševeček, 1993, je rovněž využíván počítačovým thesau-

rem Pala, Ševeček 1993). Popis sám je základem programu umožňujícího transkripci číslic do slovní podoby a naopak (Osolsobě, Ševeček 1994).

2. Segmentace slova pro potřeby algoritmického popisu formální morfologie a slovotvorby

Základem segmentace slova pro potřeby strojového popisu morfologické analýzy bylo dvoufázové ternární členění slovního tvaru od jeho konce. V první fázi se slovní tvar rozdělí na *kmen* (KM) a *koncovku* (T).

Pod pojem *kmen* (KM) zahrnujeme *slovotvorné základy*, které mohou být reprezentovány kupř. kořennými morfémy *(pět–i)*, nebo u odvozených a složených slov různě rozsáhlým komplexem morfémů a konektémů *(((deset–)in)–a)*, *(((desat–er–o)–(násoh)–n)–ý)*.

Budeme-li nadále v naší práci hovořit o *koncovce* (T), máme na mysli koncovku, jak ji definuje MČ (2), tedy tvarotvornou příponu stojící v absolutním konci slova, která je nositelkou gramatických významů příslušných gramatických kategorií pro příslušný slovní druh. Za koncovku považujeme i *nulovou koncovku*. *Nulová koncovka* (*–0*) v našem pojetí zahrnuje jak případy, kdy kmen (tvarotvorný základ) zůstává beze změny ve srovnání s tvary s *nenulovou koncovkou*, tak i případy, kdy v souvislosti s výskytem *nulové koncovky* dochází k alternaci uvnitř kmene, nebo na jeho konci.

Ve druhé fázi se *kmen* (KM) dále segmentuje na dvě části. Pracovně jsme je nazvali *kmenový základ* (KMZ) a *intersegment* (IS). Jako IS označujeme *finální skupinu kmene* (FSK), která se během flexe mění.

Cílem algoritmického popisu české formální morfologie bylo vytvořit dynamický popis skloňování a časování. V průběhu práce na přípravě strojového popisu jsme však narazili na některé případy stojící na pomezí mezi formální morfologií a tvořením slov. Zvláštní formu kombinovaného popisu jsme proto zvolili právě u číslovek. Vzhledem k tomu, že se jedná o úzký sémanticky vymezený okruh lexémů, zkombinovali jsme popis formální morfologie jednotlivých druhů číslovek a vytvořili komplexní popis flexe a slovotvorby číslovek.

Kmen (KM) rozložený pro účely strojového popisu na KMZ a FSK představuje u číslovek jednak tvarotvorný, jednak slovotvorný základ. Z tohoto přístupu vyplývá rozlišování IS prvního a druhého řádu (IS1, IS2).

IS1 – intersegmenty prvního řádu jsou části KM izolované od konce, které se během skloňování základních číslovek mění. Pokud zůstává KM během flexe neměnný, uvažujeme KM=KMZ+0 a počítáme tedy s nulovým IS1.

IS2 – intersegmenty druhého řádu jsou příslušné slovotvorné sufixy stojící mezi KMZ+IS1 a koncovkou derivovaného tvaru.

IS (IS1+IS2) je tedy souborný název pro všechny segmenty, které stojí mezi koncovkami číslovek základních, řadových, druhových, souborných, úhrnných,

násobných, názvů zlomků, názvů číslic a n–tic odvozených od číslovek základních a KMZ.

Intersegmenty prvního řádu (IS1)

K IS prvního řádu řadíme u číslovek *finální skupinu kmene* (FSK). U číslovek jsme pod pojem FSK zařadili všechny případy alternací posledních maximálně tří písmen kmene (des–**et**–0, des–**ít**–i, des–**at**–er–ý/ čt–**yř**–i, čt–**vrt**–ý, čt–**v**–er–ý).

Stejné postavení v našem systému zaujímají *kmenová finála* (KF), tedy poslední hláska (písmeno) kmene, u níž dochází k alternaci (t–r–oj–**k**–a, t–r–oj–**c**–e), a *kořenová finála* (KOF), tedy poslední písmeno kořenového morfu (t–**ř**–i, t–**r**–oj–í).

Do této skupiny (IS1) dále řadíme finální dvojici samohlásky (písmene) *–e–* a libovolného konsonantu *–C–* (*e*+*C*), která se vyskytuje u substantiv, je někdy součástí kmene neodvozeného slova, jinde se jedná o slovotvornou příponu. Společným rysem těchto „skupin" (EC), ať už je jejich genetický původ jakýkoli, je to, že u nich dochází k alternaci $e+C>0+C$. Tato skupina se vyskytuje v derivačním sufixu názvů číslic (šest–**k**–a, šest–**ek**–0). U číslovek tedy funguje v roli IS2.

Intersegmenty druhého řádu (IS2)

Zavedení pojmu intersegmentů druhého řádu do našeho popisu je branou pro plánované propojení popisu formální morfologie a slovotvorby. Do popisu české formální morfologie jsme integrovali popis derivace číslovek a číselných pojmenování. Kritériem byla snadná formalizovatelnost těchto jevů.

V rámci popisu formální morfologie číslovek jsme se pokusili vytvořit jednotný popis flexe číslovek základních *(šest–0)*, od nich odvozených číslovek řadových *(šest–ý)*, druhových *(šest–er–ý)*, souborných *(šest–er–y)*, úhrnných *(šest–er–o)*, násobných *(šest–krát, šest–i–násobn–ý, dv–oj–ak–ý, dv–oj–it–ý, dv–oj–m–o, dv–oj–násob)*, názvů zlomků *(šest–in–a)*, pojmenování číslic *(šest–k–a)* a n–tic *(šest–ic–e)*. Poslední tři jsou fakticky substantiva s číselným významem. Jednotlivé segmenty slov stojící mezi číslovkovou, adjektivní nebo substantivní koncovkou a kmenovým základem sloužící k derivaci příslušných číselných pojmenování budeme nazývat: *derivační sufix druhových číslovek* (DSDN), *derivační sufix číslovek souborných* (DSSN), *derivační sufix číslovek úhrnných* (DSUN), *derivační sufix číslovek násobných* (DSNN), *derivační sufix názvů zlomků* (DSZN), *derivační sufix pojmenování číslic* (DSCN) a *derivační sufix názvů n–tic* (DSTN). Tabulka zachycuje segmentaci číslovek a číselných výrazů.

kmen					koncovka
KMZ	FSK	DSN			
			EC	(K)F	
dev–	–ět–				–0
dev–	–ít–				–i
dev–	–át–				–ý
dev–	–at–	–e–		–r–	–ý
dev–	–at–	–e–		–ř–	–í
dev–	–at–	–e–		–r–	–y
dev–	–at–	–e–		–r–	–o
dev–	–ět–				–krát
dv–		–oj–it–			–ý
dv–		–oj–at–			–ý
dv–		–oj–a–		–k–	–ý
dv–		–oj–a–		–c–	–í
dv–		–oj–in–			–o
dv–		–oj–			–násob
dev–	–at–	–er–o–násob–n–			–ý
dev–	–ít–	–in–			–a
dev–	–ít–	–ic–			–e
dev–	–ít–		–0–	–k–	–a
dev–	–ít–		–e–	–k–	–0
dev–	–ít–		–0–	–c–	–e

Strukturu popisu číslovek a číselných pojmenování lze zapsat pomocí následujících pravidel:

NUM ⟶ *KMN* + *TN*
NUMK ⟶ *KMZNK* + *IS1* + *TNK*
NUMO ⟶ *KMZNO* + *IS1* + *TA*
NUMD ⟶ *KMZND* + *IS1* + *IS2* + *TA*
NUMS ⟶ *KMZNS* + *IS1* + *IS2* + *TA*
NUMU ⟶ *KMZNU* + *IS1* + *IS2* + *TS*
NUMN ⟶ *KMZNN* + *IS1* + *IS2* + *TA*
NUMN ⟶ *KMZNN* + *IS1* + *IS2* + *TN*
SNUMZ ⟶ *SKMZNZ* + *IS1* + *IS2* + *TS*
SNUMT ⟶ *SKMZNT* + *IS1* + *IS2* + *TS*
SNUMC ⟶ *SKMZNC* + *IS1* + *IS2* + *TS*

3. Systematický popis koncovek českých číslovek a číselných pojmenování

Definujme nejdříve pojmy, které se vztahují ke zkoumané problematice. Hovoříme-li o *flektivních koncovkách*, máme na mysli množinu všech českých flektivních koncovek, tedy z hlediska formálního všech řetězců nula–, jedno–, dvou–, až třípísmenových, jež lze odtrhnout od konce řetězce písmen, které se zbytkem řetězce (kmenem) tvoří správné tvary českých ohebných slov. Tuto množinu lze dále podle různých kritérií členit na jednotlivé podmnožiny (např. na jedno–, dvou–, třípísmenové koncovky). Základním kritériem našeho popisu bylo členění na množiny koncovek, které tvoří tvaroslovnou charakteristiku (vzor).

Obecný princip, který jsme uplatnili pro všechny koncovkové systémy i podsystémy, byl systém rozdělení koncovkových množin definovaných jakožto tvaroslovné charakteristiky (vzory) do strukturovaného systému podmnožin.

Koncovky číslovek a číselných výrazů

Popis systému koncovek českých číslovek je složen z popisu koncovek základních číslovek, dále je použit popis adjektivních a substantivních koncovek, a to v případech, že se číslovky skloňují jako adjektivum (číslovky řadové, druhové) nebo jako substantivum (názvy zlomků,číslic a n–tic).

Konkrétní rozdělení množiny koncovek českých číslovek

Množinu koncovek českých číslovek jsme rozdělili na podmnožiny koncovek číslovek základních, řadových, druhových, souborných, úhrnných a násobných. V rámci našeho popisu se dále zabýváme koncovkami názvů zlomků, číslic a n–tic.

Koncovky číslovek základních

Koncovky základních číslovek 5–99 zahrnují jeden inventář koncovek, který v některých případech nezpůsobuje alternace kmene, v některých kmen alternuje, a množina koncovek je tudíž rozdělena na dvě podmnožiny, zvláštní systém podmnožin mají číslovky 1–4 a číslovky sto, tisíc, milion a miliarda.

Popis jednotlivých množin

* Koncovky základních číslovek (5–99)
  * Kompletní množina plurálových koncovek – množina 4A
  * Koncovky nom., ak., vok. pl – množina 4B
  * Koncovky gen., dat., lok., instr. pl – množina 4C
* Koncovky základní číslovky jeden, jedna, jedno
  * Koncovky zájmen tradičně skloňovaných podle vzoru *ten*, nom. sg. mask. živ. a nom. a ak. sg. mask. neživ. množina PD1
  * Koncovky zájmen tradičně skloňovaných podle vzoru *ten*, gen., dat., lok., instr. sg. mask. živ., gen., dat., ak., vok., lok., instr. sg. mask. neživ., všech-

ny koncovky sg. a pl. fem. a neu., všechny koncovky pl. mask. živ. i neživ. – množina PD2

* Koncovky základní číslovky dvě, obě
  * Koncovky nom., ak., vok. pl. životných maskulin – množina 4D1
  * Koncovky nom., ak., vok. pl. neživotných maskulin – množina 4D2
  * Koncovky nom., ak., vok. pl. neuter – množina 4D3
  * Koncovky nom., ak., vok. pl. feminin – množina 4D4
  * Koncovky gen., dat., lok., instr. pl. – množina 4E
* Koncovky základních číslovek tři, čtyři
  * Koncovky nom., dat., ak., vok., lok. pl. – množina 4F
  * Koncovky gen., instr. pl. – množiny 4G1, 4G2
* Koncovky základní číslovky sto
  * Koncovky tvrdých neuter tradičně skloňovaných podle vzoru *město*
  * Koncovky nom., gen., dat., ak., instr. sg. a nom., dat., ak., instr. pl. tvrdých neuter – množina V11
  * Koncovka lok.pl. tvrdých neuter – množina VZ
  * Koncovka lok. sg. tvrdých neuter – množina V211E
  * Koncovka gen. pl. tvrdých neuter – množina V24781112X
  * Koncovky nom., ak. pl. – množina 4JP1
  * Koncovky nom., ak. pl. – množina 4JP2
  * Koncovky gen. pl. – množina 4JP3
  * Koncovky lok. pl. – množina 4JP4
  * Koncovky dat. pl. – množina 4JP5
  * Koncovky instr. pl. – množina 4JP6
  * Koncovky nom., ak., vok. pl. – množina 4JP7
* Koncovky základní číslovky tisíc
  * Koncovky neživotných měkkých maskulin tradičně skloňovaných podle vzoru *stroj*: dat., vok., lok. sg. a lok., instr. pl. – množina V4
  * Koncovky nom., ak. a vok. sg. neživotných měkkých maskulin – množina V24910X
  * Koncovky gen., dat. pl. neživotných měkkých maskulin – množina V4U
  * Koncovky gen. a instr. sg. a nom., ak., vok. pl. neživotných měkkých maskulin – množina V4A
* Koncovky základní číslovky milion
  * Koncovky neživotných tvrdých maskulin tradičně skloňovaných podle vzoru *hrad*: dat., lok., instr. sg. a nom., gen., dat., ak., instr. pl. – množina V2
  * Koncovky nom., ak. a vok. sg. neživotných tvrdých maskulin – množina V24910X
  * Koncovky lok. pl. tvrdých maskulin a neuter – množiny VZ
  * Koncovky gen. sg. tvrdých neživotných maskulin – množiny V2C
  * Koncovky lok. sg. tvrdých neživotných maskulin a neuter – množiny V211E
* Koncovky základní číslovky miliarda

* Koncovky tvrdých feminin tradičně skloňovaných podle vzoru *žena*: nom., gen., ak., vok., instr. sg. a nom., dat., ak., vok., lok., instr. pl. – množina V7
* Koncovky dat. a lok. sg. tvrdých feminin – množina V7B
* Koncovky gen. pl. tvrdých feminin a neuter – množina V24781112X

Koncovky číslovek řadových

Koncovky řadových číslovek jsou koncovkami tvrdých, případně měkkých adjektiv.

Popis jednotlivých množin

* Koncovky řadových číslovek se pojí ke kmeni číslovek základních. Číslovky řadové *prvn–í, druh–ý* mají alternativní kmeny
  * Koncovky I.stupně tvrdých adjektiv – nom. pl. životných maskulin – množina PRT1, koncovky ostatních pádů – množina PRT2
  * Koncovky I.stupně měkkých adjektiv – množina PRM

Koncovky číslovek druhových

Koncovky druhových číslovek jsou rovněž koncovkami adjektivními a platí o nich totéž, co bylo řečeno o koncovkách číslovek řadových.

Popis jednotlivých množin

* Koncovky druhových číslovek pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –er– nebo –oj–
  * Koncovky I.stupně tvrdých adjektiv – nom. pl. životných maskulin – množina PRT1, koncovky ostatních pádů – množina PRT2
  * Koncovky I.stupně měkkých adjektiv – množina PRM

Koncovky číslovek souborných

* Koncovky druhových číslovek pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –er–
  * Koncovková množina 4S1
* Koncovky druhových číslovek pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –oj–
  * Koncovková množina 4S2

Koncovky číslovek úhrnných

* Koncovky druhových číslovek pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –er–
  * Koncovková množina 4U1
* Koncovky druhových číslovek pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –oj–
  * Koncovková množina 4U2

Koncovky číslovek násobných

Jedná se spíše o slovotvorný sufix pro derivaci číslovek majících charakter adverbií *–krát, –mo, –násob*. Ostatní deriváty mají charakter adjektiv a stojí na hranici mezi kompozicí a derivací.

Popis jednotlivých množin

* Koncovky číslovek násobných mající adverbiální charakter

* Koncovka, nebo spíše slovotvorný formant – množina 6I
* Koncovková množina pro tvoření I. stupně adverbií – 6C
* Koncovková množina pro tvoření I. stupně adverbií – 6M
* Koncovková množina pro tvoření I. stupně adverbií – 6K

* Koncovky číslovek násobných pojící se ke kmeni rozšířenému o derivační sufix –at–, –ak–, –násob–n–.
  * Koncovky I.stupně tvrdých adjektiv – nom. pl. životných maskulin – množina PRT1, koncovky ostatních pádů – množina PRT2

Koncovky názvů zlomků

Koncovky názvů zlomků jsou po formální stránce koncovkami tvrdých feminin tradičně skloňovaných podle vzoru *žena.*

Popis jednotlivých množin

* Koncovky zlomků pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –in–
  * Koncovky tvrdých feminin tradičně skloňovaných podle vzoru *žena*: nom., gen., ak., vok., instr. sg. a nom., dat., ak., vok., lok., instr. pl. – množina V7
  * Koncovky dat. a lok. sg. tvrdých feminin – množina V7B
  * Koncovky gen. pl. tvrdých feminin a neuter – množina V24781112X

Koncovky názvů číslic

Rovněž koncovky názvů číslic jsou po formální stránce koncovkami tvrdých feminin tradičně skloňovaných podle vzoru *žena.*

Popis jednotlivých množin

* Koncovky názvů číslic pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –k–(–c–,–ek–) nebo –ičk–(–ičc–, –iček–)
  * Koncovky tvrdých feminin tradičně skloňovaných podle vzoru *žena*: nom., gen., ak., vok., instr. sg. a nom., dat., ak., vok., lok., instr. pl. – množina V7
  * Koncovky dat. a lok. sg. tvrdých feminin – množina V7A
  * Koncovky gen. pl. tvrdých feminin a neuter – množina V24781112X

Koncovky názvů n–tic

Koncovky názvů n–tic jsou po formální stránce koncovkami měkkých feminin tradičně skloňovaných podle vzoru *růže.*

Popis jednotlivých množin

* Koncovky názvů n–tic pojících se ke kmeni základních číslovek rozšířenému o tvarotvorný sufix –ic–
  * Koncovky nom., gen., sg. a nom., ak., vok., instr. pl. měkkých feminin – množiny V8A
  * Koncovky měkkých feminin tradičně skloňovaných podle vzoru *růže:* dat., ak., lok., instr. sg. a dat., lok. pl. – množina V8
  * Koncovky gen. pl. měkkých feminin a neuter – množina V24781112X

4. Intersegmenty

Intersegment (IS) je společné označení pro řetězec 0–4 znaků (písmen), předcházejících bezprostředně před koncovkou (T), která může mít 0–3 znaky.

IS je součástí tvarotvorného základu (kmene), která se často mění v souvislosti s ohýbáním slova. Oddělením intersegmentu IS od kmene KM vzniká kmenový základ KMZ, tj. část tvarotvorného základu slova uložená ve slovníku kmenů, sloužící k identifikaci analyzovaného slovního tvaru.

Jak již bylo řečeno výše, rozlišujeme v našem popisu IS prvního řádu a IS druhého řádu. IS prvního řádu dále dělíme na alternační intersegmenty (AIS), k nimž patří u číslovek FSK, KOF, KF a EC, a gramatické intersegmenty (GIS), s nimiž se u číslovek nepracuje. K IS druhého řádu patří u numerálií DSDN, DSSN, DSUN, DSNN, DSZN, DSCN, DSTN.

IS prvního řádu mohou být –0–. IS druhého řádu jsou naopak vždy nenulové. Praktická analýza, na jejíž bázi pracují aktuální programové produkty, zkoumá IS prvního a druhého řádu spojitě jako jeden jediný segment. „Čisté" IS druhého řádu se tudíž při analýze vyskytují pouze v případech nulových IS prvního řádu.

Teoretický přínos ternárního členění ohebných slov na KMZ+IS+T spočívá ve vytvoření dynamického popisu flektivní morfologie zahrnujícího nejen pravidla změn koncovek (ohýbání slov), ale i pravidla změn tvarotvorných základů v závislosti na flexi (morfologických alternací). Zavedení pojmu IS2 a jejich zapracování do navrženého modelu jsou ukázkou propojení flektivní a derivační morfologie.

Praktickým důsledkem je podstatné zmenšení rozsahu slovníku kmenů. To je dáno v případě IS1 zmenšením počtu variant kmene, od nějž se tvoří paradigma. Zavedením IS2 se snižuje počet hesel ve slovníku kmenů a současně roste počet automaticky generovatelných slov a jejich slovních tvarů.

Intersegmenty prvního řádu

FSK

KOF, KF

U FSK, KF a KOF se setkáváme se změnami, které jsou pro strojový popis důležité. Jedná se o tři typy alternací:

* alternace, které mají zvukovou a grafickou realizaci
* alternace, které mají zvukovou, avšak nikoli grafickou realizaci
* alternace, které nemají zvukovou, ale mají grafickou realizaci

Zatímco klasické mluvnice uvažují o alternacích pouze v prvním a druhém případě, třetí případ se v nich vlastně vůbec neuvádí, bylo pro potřeby strojového popisu zaměřeného k analýze a generování psané podoby přirozeného jazyka nutné vzít v úvahu třetí případ, zatímco druhý případ zůstává pro model zaměřený na psanou podobu jazyka prozatím stranou.

Přehled alternací FSK:

* alternace, které mají zvukovou a grafickou realizaci
* alternace FSK

    * et>ít>át> (des–**et**–0, des–**ít**–i, des–**át**–ý)
    * ět>ít>át (dev–**ět**–0, dev–**ít**–i, dev–**át**–ý)
  * alternace kmene a rozšiřování kmene při odvozování
    * yř>vrt>v (čt–**yř**–i, čt–**vr**–t–ý, čt–**v**–er–ý)
    * ř>řet>r (t–**ř**–i, t–**řet**–í, t–**r**–oj–í)
* alternace, které mají zvukovou a nemají grafickou realizaci
  * číslovky a číslovková pojmenování se substantivní a adjektivní flexí
    * alternace měkčení t>ť (čt–vrt–ý, čt–vrt–í)
* alternace, která nemají zvukovou, ale mají grafickou realizaci, se u českých číslovek nevyskytují

Intersegmenty druhého řádu

Derivační sufixy pro tvoření číslovek a číselných pojmenování (DSDN, DSSN, DSUN, DSNN, DSZN, DSCN, DSTN)

Na základě spojení flexe a odvozování jednotlivých druhů číslovek od kmenů základních číslovek jsme vytvořili v rámci algoritmického popisu podmodel, který ukazuje, kterým směrem by se měly rozvíjet lingvistické úvahy, na jejichž základě je algoritmický popis formální morfologie vybudován. Popis flexe a derivace číslovek bylo možno propojit, protože se jedná o velmi omezenou množinu kmenů. Derivace příslušných slov a jejich tvarů je naprosto pravidelná pro většinu prvků dané množiny a lze ji tudíž poměrně snadno formalizovat.

U jednotlivých typů DSN se setkáváme s některými typy alternací popsanými výše, jako jsou alternace KF, jí odpovídá alternace finálního písmene DSN a alternace typu EC, která se vyskytuje v případě DSCN, kde derivační sufix má vlastnosti IS typu EC.

Přehled všech typů DSN

* DSDN–
  * –er–/–eř–
* DSSN
  * –er–
  * –oj–
* DSUN
  * –er–
  * –oj–
* DSNN
  * –0–
  * –oj–m–
  * –násob–n–
  * –o–násob–n–
  * –i–násob–n–
  * –oj–násob–n–
  * –násob
  * –i–násob

  * –oj–násob
  * –oj–it–
  * –oj–at–
  * –oj–ak–/–ac–
* DSZN
  * –in–
* DSCN
  * –k–>–ek>–c–
  * –ič–k–>–ič–ek>–ič–c–
* DSTN
  * –ic–

5. Vzory

Při popisu tvaroslovného systému je základním termínem *paradigma. Morfologické (tvaroslovné) paradigma* je soubor tvarů ohebného slova vyjadřující systém jeho mluvnických kategorií.

MČ (2) definuje *vzor* jako reprezentaci tvaroslovného paradigmatu paradigmatem určitého konkrétního slova. Pod pojmem *vzor* rozumíme v naší práci definici zahrnující popis tvoření všech paradigmatických, popřípadě odvozených forem pro konkrétní vzorové slovo. Vzorové slovo reprezentuje množinu všech slov, která tvoří ohebné i derivované tvary pomocí identického inventáře koncovek. V souvislosti s ohýbáním a derivací u nich dochází ke stejným změnám FSK. Takto definované vzorové slovo je ve slovníku kmenů přiřazeno všem kmenům, které reprezentuje.

Definice jednotlivých vzorů je hlavní součástí algoritmického popisu české formální morfologie a odvozování některých slovních tvarů. Algoritmický popis zahrnuje na prvním místě definice koncovkových množin, jak o nich byla řeč výše. Vzory jsou pak definovány prostřednictvím vzorových slov, která se rozpadají na neměnnou část vzorového slova (KMZ), IS – proměnlivou část vzorového slova a koncovkové množiny MK zahrnující všechny koncovky, s nimiž se KM vzorového slova může spojit, aby vzniklo správné české slovo.

Při popisu vzorů jsme vyšli z klasických popisů, jak jsou uváděny v mluvnicích, např. MČ 2 atd., tedy v podstatě ze vzorů učebnicových. Pro potřeby algoritmizace se tento popis již na začátku ukázal jako naprosto nedostatečný, a to ze dvou důvodů. Prvním z nich je statičnost klasických popisů. Vzor je definován jako „tabulkové" paradigma vzorového slova, u některých vzorů jsou navíc uvedena nejvýraznější kolísání v koncovkách jistých pádů a další „vedlejší" vzorové slovo. Všechny ostatní gramatické jevy (alternace kmene, alternativní koncovky, dublety, výjimky) jsou uvedeny v jednotlivých odstavcích, prostě se konstatují, jsou uváděny (většinou neúplné) seznamy příkladů ilustrujících zmíněný jev atd. Druhým důvodem je potřeba algoritmického popisu, který je zamě-

řen na morfologickou analýzu grafické podoby jazyka, jíž se v běžných popisech z pochopitelných důvodů nevěnuje dostatečná pozornost.

Zajímavé je srovnání přístupu ke vzorům v klasických mluvnicích se systémem *vzorů*, který je základem počítačového zpracování přirozeného jazyka.

Flexe číslovek určitých a jejich vzory

Číslovky jsou jako slovní druh definovány spíše sémanticky než morfologicky. Speciální číslovkovou flexi mají geneticky pouze základní číslovky dvě, tři a čtyři, číslovka jedna má skloňování zájmenné, základní číslovky od pěti nahorů jsou původně substantiva, u číslovek sto, tisíc, milion a miliarda je jejich substantivní původ dosud transparentní.

Číslovky řadové jsou z formálního hlediska tvrdá nebo měkká adjektiva, číslovky druhové jsou formálně tvrdá adjektiva. Jmenné adjektivní tvary představují číslovky souborné a úhrnné. Číslovky násobné jsou buď ustrnulé neohebné tvary mající příslovečný charakter, a nebo deriváty, popřípadě kompozita s adjektivní flexí.

Přehledný popis číslovkových vzorů

Mluvnice běžně uvádějí vzory pro skloňování číslovek základních určitých a číslovek neurčitých, protože ostatní druhy číslovek se vesměs skloňují jako adjektiva nebo substantiva, nebo zůstavají nesklonné. Rovněž náš popis vychází z kmene číslovky základní, od něhož se pak dále prostřednictvím derivačních sufixů, které jsou v našem popisu reprezentovány IS2, tvoří ostatní druhy číslovek a jejich paradigmatické tvary.

Přehledný popis vztahu nových číslovkových vzorů ke klasickým učebnicovým vzorům

Klasický vzor „ten“

Podle tohoto vzoru se skloňuje číslovka *jeden, jedna, jedno*. Vzhledem ke specifikům číslovkových vzorů, o nichž jsme se zmiňovali výše, jsme do našeho popisu zařadili samostatný číslovkový vzor, který je ovšem definován koncovkovými množinami společnými pro definice zájmenných vzorů. Navíc jsou součástí definice koncovkové množiny pro derivování všech příslušných odvozených číslovek a číselných pojmenování, tedy i supletivní kmen pro tvoření číslovek řadových.

* vzor [jed–en–]
* vzor [prvn–0–]

Klasický vzor „dva, oba“

Náš popis obsahuje vzor pro tvoření tvarů číslovky *dvě, obě*.

* vzor [stod--v–]
* vzor [ob–0–]

Klasický vzor „tři“

V rámci algoritmického popisu mu odpovídá vzor pro tvoření tvarů číslovky tři a příslušných odvozených numerálií.

* vzor [stot]

Klasický vzor „čtyři“

Algoritmický popis obsahuje stejně jako klasické popisy samostatný vzor pro generování tvarů číslovky čtyři a od ní odvozených číslovek a číselných pojmenování.

* vzor [stočt]

Klasický vzor „pět“

Podle tohoto vzoru se skloňují číslovky od pěti výše až do devadesáti devíti a všechny komponenty vyšší číslovky složené z těchto číslovek. V rámci našeho popisu se tento vzor rozpadá do podvzorů: *pět, šest, sedm, devět, deset, padesát.*

* Číslovkový kmen nealternuje ani při tvoření tvarů paradigmatu číslovek základních, ani při derivaci ostatních druhů číslovek a číselných pojmenování. Podle tohoto vzoru se ohýbají a odvozují tvary číslovek šest a číslovek jedenáct až devatenáct.
  * vzor [šest]
* Číslovkový kmen nealternuje při tvoření tvarů paradigmatu číslovek základních, ale alternuje při derivaci ostatních druhů číslovek a číselných pojmenování.
  * vzor [stop–ět / stop–át–ý / stop–at–er–ý]

  Podle tohoto vzoru se tvoří tvary číslovky pět.
  * vzor [pades–át/ pades–at–er–ý]

  Podle tohoto vzoru se tvoří tvary číslovek padesát, šedesát, sedmdesát, osmdesát a devadesát.
* Číslovkový kmen alternuje při tvoření tvarů paradigmatu číslovek základních i při derivaci ostatních druhů číslovek a číselných pojmenování.
  * vzor [dev–ět / dev–ít–i / dev–át–ý / dev–at–er–ý]

  Podle tohoto vzoru se tvoří tvary číslovky devět.
  * vzor [des–et/ des–ít–i / des–át–ý / des–at–er–ý]

  Podle tohoto vzoru se tvoří tvary číslovek deset, dvacet, třicet a čtyřicet.
* Číslovkový kmen nealternuje při tvoření tvarů paradigmatu číslovek základních, ani při derivaci ostatních druhů číslovek a číselných pojmenování. Má však odlišný derivační sufix pro tvoření pojmenování číslic.
  * vzor [sedm / sedm–ič–k–]

  Podle tohoto vzoru se tvoří tvary číslovek sedm a osm.

Další číslovkové vzory pro skloňování a derivaci číslovek běžně řazených k substantivním vzorům

* vzor [s–t–o]
* vzor [tisíc]
* vzor [milion]
* vzor [miliard]

6. Příklady algoritmického popisu českých číslovkových koncovek a vzorů

Definice koncovkových množin

Definice koncovkové množiny má pevnou formu. Začíná znakem „=", po něm následuje bez mezery jméno koncovkové množiny (MK) složené z číslic a velkých písmen latinské abecedy. Ve třetím sloupci druhého řádku definice je v páru hranatých závorek [] uzavřena informace o slovním druhu, čísle a rodě, gramatických významů společných pro více koncovek. Další řádek začíná čtvrtým sloupcem. V kulatých závorkách () je uvedena příslušná koncovka, za ní následuje čárka a po ní číslice označující gramatický význam pádu vyjadřovaný příslušnou koncovkou v rámci definované množiny koncovek.
Obecně můžeme definici koncovky znázornit následujícím schématem:

```
+JT
  [SD,NU,GEN]
    (T,CAS)
```

Konkrétní příklad popisu definice koncovky:

```
;..... NUMERÁLIA .....
#type=4
=4A
  [Tpx]
    (_,1)
    (i,2)
    (i,3)
    (_,4)
    (_,5)
    (i,6)
    (i,7)
```

Definice vzorů

Definice vzorů mají pevnou stavbu. První řádek definice začíná symbolem „+", za ním bez mezery následuje KMZ vzorového slova. (KMZ jsou případně upraveny tak aby, byly tvořeny pokud možno více než jedním písmenem. Při počtu nad 500 vzorů v kompletním popisu české flexe by v případě jednopísmenových KMZ vzorových slov docházelo k opakování a tedy nejednoznačnostem v definování vzorů.) Každý následující řádek definice začíná třetím sloupcem a skládá se z páru <> závorek, v nichž jsou uzavřeny jednotlivé IS1 a IS2. Po <IS> následuje mezera a jména jednotlivých koncovkových množin, které se pojí s příslušným KMZ+IS. Mezi jednotlivými jmény MK jsou čárky, za čárkou

bez mezery následuje další jméno MK. Za posledním jménem koncovkové množiny není čárka ani jiný oddělovač.

Obecně můžeme definici vzoru znázornit následujícím schématem:

```
+KMZ
  <IS> MKa,.....,MKn


  <ISn> MKb,.....,MKz
```

Konkrétní příklad popisu definice vzoru:

```
;------ vzory:
+stop
   <ět> 4A,6J
   <át> PRT1,PRT2
   <ětin> V7,V7B,V24781112X
   <ater> PRT1,4S1,4U1
   <atero> 6K
   <ateř> PRT2
   <ětic> V8,V8A,V24781112X
   <ětk> V7
   <ětek> V24781112X
   <ětc> V7A
```

7. Strojový slovník kmenů číslovek určitých

Struktura strojového slovníku má svá přesná pravidla. První řádek hesla obsahuje KMZ–IS1+IS2–T(, KMZn–IS1n+IS2n–Tn) slova (slov), následuje mezera, symbol „>„ a vzorové slovo (slova) podle definice vzorů v popisu vzorů. Druhý řádek hesla začíná třetím sloupcem symbolem „#“, po němž následuje mezera a seznam prefixů, popřípadě prvních částí kompozit derivovaných od KMZ.

Konkrétní příklad strojového slovníku číslovkových kmenů:

```
jed–en, prvn–í                >jed,prvn
d–v–a, ob–a                   >stod,ob
t–ř–i                         >stot
čt–yř–i                       >stočt
p–ět                          >stop
šest                          >šest
sedm                          >sedm
osm                           >sedm
dev–ět                        >dev
des–et                        >des
jedenáct                      >šest
```

dvanáct >šest
třináct >šest
čtrnáct >šest
patnáct >šest
šestnáct >šest
sedmnáct >šest
osmnáct >šest
devatenáct >šest
dvac–et >des
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
tři c–et >des
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
čty řic–et >des
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
pades–át >pades
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
šededes–át >pades
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
sedmdes–át >pades
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
osmdes–át >pades
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
devades–át >pades
_, jedna, dvaa, tři a, čtyři a, pěta, šesta, sedma, osma, devěta,
s–t–o, >cs
tisíc >tisíc
milion >milion
milión >milion
miliard–a >miliard
bilion >milion
bilión >milion
trilion >milion
trilión >milion
kvadrilion >milion
kvadrilión >milion
kvintilion >milion
kvintilión >milion
sextilion >milion
sextilión >milion
septilion >milion
septilión >milion

Na základě ani ne padesátiřádkového slovníku kmenů určitých číslovek je možné pomocí počítače vygenerovat cca 2000 slovních tvarů a v podstatě nekonečnou řadu jedno– a víceslovných pojmenování, či odpovídajících cifer.

8. Závěr

Propojení formální morfologie a slovotvorby aplikované na algoritmický popis směřující k automatické analýze a syntéze českých číslovkových tvarů a tvarů číselných pojmenování pomocí počítače je dalším stupněm strojového zpracování přirozeného jazyka. V současné době je výše uvedený algoritmický popis formální morfologie a slovotvorby číslovek a číselných pojmenování součástí algoritmického popisu české formální morfologie jako celku, který je v součinnosti se strojovým slovníkem kmenů lingvistickou bazí konkrétních softwarových produktů (český spelling–checker Osolsobě, Ševeček 1991, automatický lemmatizátor Osolsobě, Ševeček 1993). Programové zpracování navíc umožňuje transkripci jednotlivých číslovek a číselných pojmenování formou číslic (Osolsobě, Ševeček 1994). Vzhledem k tomu, že se formálně jedná o zpracování víceslovných přesně definovatelných pojmenování, je takové zpracování inspirativní pro řešení problému automatického rozpoznávání a generování syntaktických celků s pevnou slovoslednou stavbou, a tudíž jakousi branou k automatické strojové syntaktické analýze a syntéze.

DAS AUTOMATISCHE ERKENNEN UND DIE GENERIERUNG DER TSCHECHISCHEN ZAHLWÖRTER MIT PC

Die Integration der Form– und Wortbildung im Rahmen der algoritmischen Beschreibung der tschechischen Numeralien und der Ausdrücke mit der Zahlbedeutung stellt die erste Stufe auf dem Weg zur automatischen Analyse und Synthese nicht nur der Flexionsmorphologie, sondern auch der Wortbildung dar.

Die oben genannte algoritmische Beschreibung ist ein Bestandteil der algoritmischen Beschreibung der tschechischen Morphologie als Ganzen, die mit dem Computer-Stammwörterbuch die Basis der konkreten Softwareprodukte (der automatische Spelling-checker – Osolsobě, Ševeček 1991, der automatische Lemmatizator Osolsobě, Ševeček 1993) bildet.

Die Programmbearbeitung der Beschreibung der Form– und Wortbildung der tschechischen Numeralien ermöglicht die automatische Transkription der Zahlwörter in den Ziffern und umgekehrt (Osolsobě, Ševeček, 1994). Dieses Fakt hat eine Bedeutung für das automatische Erkennen und die Generierung der syntaktischen Konstruktionen mit der festen Wortfolge.

LITERATURA

ČERMÁK, F., Syntagmatika a paradigmatika českého slova I., II., UK, Praha, 1990.

HAVRÁNEK, B., JEDLIČKA, A., Česká mluvnice, SPN, Praha, 1981.

KOMÁREK, M., Ke dvěma koncepcím stavby jednoduchých slovesných tvarů v češtině, Acta Universitatis Olomucensis, Studia Bohemica IV, SPN, Praha 1987.

LAMPRECHT, A., Praslovanština, Univerzita J. E. Purkyně, Brno, 1987.

OSOLSOBĚ, K., PALA, K., FRANC, S.: Česká morfologie a syntax v PROLOGU, sb.semináře SOFSEM 1987, VUSEIAR, Bratislava 1987.

HALASOVÁ–OSOLSOBĚ, K.: Algoritmický popis české formální morfologie substantiv a adjektiv, SPFFBU, A 37–38, 1989–90, s.83–97.

OSOLSOBĚ, K. , PALA, K.: Czech Stem Dictionary for IBM PC XT/AT, Conference on Computer Lexicography, Balatonfüred, September 1990.

OSOLSOBĚ, K.: Popis systému českých substantivních a slovesných vzorů, rukopis, Brno, 1991.

OSOLSOBĚ, K. , PALA, K.: Czech Stem Dictionary, SPFFBU, A 41, 1993, s. 70–83 .

PETR, J., kol., Mluvnice češtiny I.,II., Academia, Praha, 1986.

ROMPORTL, S., Návrh principu automatického šifrování a dešifrace gramatických příznaků českého slovesa při překládání z češtiny a do češtiny, In: SbVUT, Brno 1961.

ROMPORTL, S., Struktura gramatické složky slovesných tvarů určitých v češtině, Academia, Praha, 1970.

SGALL, P., Generativní popis jazyka a česká deklinace, ČSAV, Praha, 1967.