Trojrekční předložka „za“ a automatické rozpoznávání její rekce v českých textech

Vladimír PETKEVIČ

Univerzita Karlova, Filozofická fakulta, Praha

Résumé : Je voudrais faire plaisir à cher et éstimé professeur Patrice Pognan avec une petite étude de cas qui devrait montrer comment il est possible d'utiliser les outils informatiques automatiques de la linguistique de corpus pour déterminer les valeurs de certaines catégories morphologiques dans une langue à flexion. Je vais me concentrer sur une langue qui est extrêmement proche de l'esprit, du cœur et de l'intérêt de Patrice, une langue qu'il connaît parfaitement : le tchèque.

Mots-clés : TAL, prépositions tchèques, linguistique tchèque

1 Úvod

Chtěl bych váženého a milého profesora Patrice Pognana potěšit menší případovou studií, která by měla ukázat, jak je možné pomocí automatických počítačových nástrojů korpusové lingvistiky určit hodnoty některých morfologických kategorií ve flektivním jazyce. Zaměřím se na jazyk Patricovu duchu, srdci a zájmu nesmírně blízký, na jazyk, který přitom dokonale ovládá: na češtinu.

Již dlouhá léta se spolu s kolegy v Ústavu teoretické a komputační lingvistiky FF UK zabývám úkolem nadmíru obtížným: morfologickou anotací korpusů současné češtiny – korpusů řady SYN, např. korpusu SYN2015 (Křen *et al.*, 2015) či různých verzí korpusu SYN (poslední verzí této řady je SYN v8 z konce roku 2019, Křen *et al.*, 2019). Spolu s kolegy vytváříme pravidla tzv. automatické morfologické disambiguace českých textů, tj. určujeme podle kontextu (i) hodnoty morfologických kategorií v češtině (slovní druh, jmenný rod, číslo, pád, čas, slovesný rod, osobu, vid…) a (ii) reprezentativní tvary, tj. lemmata, a připisujeme je jednotlivým slovům v textech shromážděných v korpusech současné češtiny. Například ve větě

(1) *Bál jsem se ho.*

by disambiguační pravidla měla správně anotovat jednotlivá slova takto:

bál: lemma = „bát (se)“, sloveso, singulár, maskulinum, l-ové příčestí, nedokonavý vid…
jsem: lemma = „být“, sloveso, 1. osoba, singulár, nedokonavý vid…
se: lemma = „se“, zájmeno (přesněji reflexivní částice)
ho: lemma = „on“, zájmeno osobní, maskulinum/neutrum, singulár, genitiv…

Morfologická anotace je úkol obecně nesmírně obtížný a věta (1) to také dokládá, neboť:

- tvar *bál* může být v jiných kontextech tvarem nominativu/akuzativu singuláru substantivního lexému *bál*
- tvar *se* může být v jiných kontextech také tvarem vokalizované předložky *s*
- *ho* může být v jiných kontextech tvarem akuzativu singuláru maskulina/neutra.

Snad nejobtížnější při anotaci flektivních jazyků, které mají pádovou deklinaci, je právě určení pádu. Týká se to pochopitelně i češtiny s jejími sedmi pády. V této studii se budu zabývat dílčím problémem určení pádu, a to určením pádu syntaktického substantiva, který je určen rekcí předložky „za“.

V kapitole 2 nejprve na příkladech představím trojí pádovou rekci (genitiv, akuzativ, instrumentál) předložky „za“ a vymezím repertoár funkcí těchto rekčních pádů. V kapitole 3 stručně pojednám o typech pádového synkretismu substantiv, tj. o pádové homonymii mezi těmito třemi pády. V kapitole 4 pak naznačím hlavní ideje, na jejichž základě lze zformulovat pravidla automatické morfologické disambiguace, jež podle kontextu stanoví náležitou rekci předložky „za“. V závěru (kapitola 5) se zmíním o praktickém využití implementovaných disambiguačních pravidel.

2 Pádová rekce předložky „za“

Předložka „za“ může jménům (tj. substantivům, adjektivům, zájmenům a číslovkám) udělovat tři pády: genitiv, akuzativ a instrumentál. Mimo předložku „za“ má trojí rekci už jen předložka „napříč“: genitiv, dativ a instrumentál, přičemž výrazně převažuje instrumentál. Různá rekce předložky „napříč“ nerozlišuje – zhruba řečeno – na rozdíl od předložky „za“ význam. Trojí rekci předložky „za“ zhuštěně dokládá následující věta:

(2) *Za*-Gen *Husáka*-Gen *jsme za*-Acc *socialismus*-Acc *bojovali za*-Ins *pecí*-Ins.

Indexy vyznačují u předložky „za“ její rekci, u řízených substantiv jejich pád. Na první pohled snadno stanovíme příslušný pád, a to především proto, že větě rozumíme, tj. na základě sémantiky a znalosti světa. Pro úkol automatické softwarové anotace (značkování) českých textů je úloha správně určit pád jmen obecně velmi obtížná; týká se to jak anotace pomocí gramatických pravidel, tak anotace realizované na bázi statistiky, dnes především na bázi neuronových sítí (např. Morphodita[1]). Anotační programy dosud neumějí využívat znalostí světa a jen v omezené míře dokážou mode-

[1] http://ufal.mff.cuni.cz/morphodita.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

lovat sémantiku. Jádrem takových programů je využívání a zpracování syntaktických konfigurací.

Níže se budu zabývat lingvistickými úvahami, které by mohly napomoci formulaci vhodných syntaktických pravidel pro určení (disambiguaci) předložkové rekce předložky „za“ a syntaktického substantiva, které předložka řídí. Nejprve bude třeba si ujasnit povrchověsyntaktické funkce různých rekčních pádů předložky „za“.

2.1 Povrchověsyntaktické funkce předložky „za“

2.1.1 Genitiv (v korpusu SYN2015: 33653 výskyty)

Na rozdíl od ostatních dvou pádů má genitivní rekce předložky „za“ pouze dvě vyhraněné funkce, což je pro automatické značkování pomocí pravidel vítané:

(a) Časové určení – během nějaké doby:

(3) *Když si vyjdu* za krásného počasí, *lezu po ulici jako želva, otírám se o domovní zdi, v ruce hůl.*

(4) *V létě 1949 američtí kryptoanalytici rozluštili šifru, kterou Sověti používali* za války.

(b) Okolnostní určení:

(5) *Tělo se* za normálních okolností *brání dalšímu jídlu, jestliže má vytvořené dostatečné zásoby na „horší časy“.*

(6) *Bylo to šílený, nahrávalo se v hrozně krátkém čase* za hrozných podmínek.

2.1.2 Akuzativ (v korpusu SYN2015: 264109 výskytů)

Akuzativní rekce předložky „za“ výrazně převažuje na ostatními dvěma pády, a to vzhledem k následujícím hlavním funkcím – ty je dost nesnadné ostře vymezit vzhledem ke značnému stupni gramatikalizace předložky „za“:

(a) Místo: vyjádření směru – za určitou mez:

(7) *Brány českých nemocnic se kvůli chystanému exodu tuzemských zdravotníků* za hranice *otevírají pro lékaře-cizince.*

(8) *Interiér nabízí uspořádání sedadel 2+2, ale pokud měříte ke dvěma metrům, nikdo už si* za vás *nesedne.*

(b) Pro co, ve prospěch čeho, kvůli čemu (často o ceně), z jistého důvodu, v jisté kompetenci:

(9) *V životě nic podobného neudělala, ale nabídky* za tři sta *dolarů by si měl člověk vážit.*

(10) *Děkuji* za krásné motivace *… a už teď se těším na Coopera.*

(11) *A ve chvíli, kdy se stal mnou a já prozřel, že se tento pekelný útok opakuje jako vzpoura a zároveň pomsta* za všechno minulé pokoření *trýzněné loutky…*

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

(12) *Zmiňovat tak samozřejmou věc, že bez dovolení nesmíte používat cizí mobil a utrácet cizí peníze* za svoje zprávy *nebo hovory, to snad ani nemá smysl.*

(13) *V nově vzniklé pozici ředitele rozvoje mobilních plateb a inovací bude zodpovědný* za oblast *nových platebních technologií.*

(14) *Slavná kniha nositele Pulitzerovy ceny* za literaturu *a jednoho z nejznámějších amerických spisovatelů Michaela Cunninghama znají Češi například díky románu Hodiny.*

(c) Časové vymezující určení – v nějakém období:

(15) *Odpoledne jí pak doručili šaty, jaké v životě neviděla, spolu s párem střevíčků, které musely stát tolik, co vydělala* za celý rok.

(16) Za pár *vteřin nafoukl balonek.*[2]

(d) Časové určení – po jak dlouhé době:

(17) *Když Clare s Rebeccou odjely, začala se mi hrubě nelíbit představa, že se budou muset* za pár *dní zase vrátit do baráku, kde nahoře skomírá chudák Erich.*

(18) *Zavolejte jí* za půl *roku, pokud ještě budete v Evropě.*

(e) Místo koho/čeho včetně vyjádření zástupnosti:

(19) *Bezpochyby ani on nebyl tím,* za koho *se vydával, ale právě proto bychom mu měli uvěřit.*

(20) *A někdo jim už velice brzo řekne o doktorce Anně Kingový a o tom, že vyměňuje léky* za zachovalý těla *emzáků.*

(21) Za všechny projekty *lze jmenovat třeba rekonstrukci frekventované křižovatky na ulici České mládeže v Liberci.*

(f) Vyjádření způsobu/prostředku/cíle/doteku v širokém smyslu:

(22) *Uvnitř svatého Petra zatahal zvoník* za provaz, *ale neozval se žádný zvuk.*

(23) *Vzala ho* za ruku *a vraceli se do města, povídali si, jako by se nic nestalo.*

(g) Vyjádření ceny, hodnoty (i přeneseně):

(24) *Pak by jedna taková LMB jednotka zásobovala dům energií* za méně *než 75 dolarů za měsíc.*

(25) *Pokud provozujete restauraci a nemůžete si účtovat* za vodu, *jednoduše si necháte zaplatit víc* za hamburger.

(h) Typ pokládat / považovat za:

(26) *A najednou se začnou objevovat elektronické dokumenty v různých formátech, které by také měly být pokládány* za oficiální dokumenty *a účastnit se standardních procesů.*

[2] Věta je zjevně dvojznačná, srov. typ (d) níže.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

(i) Vyjadřování položek ve výčtu:

(27) *Prosím vás, neříkejte mi mistře, to* za prvé.

(j) Vyjádření, že někdo/něco je nějaké nebo je jako něco jiného:

(28) *Přesto ještě dnes může takovou vizi někdo označit* za donquijotskou.

(29) *Od ministerstva kultury vedeného ještě Jiřím Besserem dostal* za úkol, *že musí z úsporných důvodů spojit v jeden celek Národní divadlo a Státní operu.*

(30) *Kdyby to byl prorok, musel by poznat, co je to* za ženu, *která se ho dotýká, že je to hříšnice.*

(k) Zbývající nespecifikovaná skupina významů předložky „za“ – plná gramatikalizace:

(31) *Kdyby ses provdala* za studenta *medicíny, rovnou bych řekl…*

(32) *Mám* za to, *že to bude ten, kterému odpustil nejvíc.*

2.1.3 Instrumentál (v korpusu SYN2015: 84210 výskytů)

Instrumentálová rekce předložky „za“ je poměrně častá; její funkce vykazují menší stupeň gramatikalizace než její akuzativní rekce:

(a) Určení místa kde – za někým/něčím:

(33) *Vysoce postavený manažer, který má* za sebou *úspěšnou kariéru, vydělává hromadu peněz, těší se důvěře nadřízených i podřízených, má rodinu, s ní tráví dovolenou.*

(34) *Půl hodiny nato dojeli do vesničky* za Ženevou, *jmenovala se Cologny; kostelík, pekárna, radnice, vše, jak má být.*

(35) *Seděl* za stolem *a podřimoval.*

(b) Následování:

(36) A *kromě toho, nikdo z těch, koho cestou potkala, nešel* za svým přáním.

(37) *My všichni: ředitel, hudebník, zubař, spisovatel, pokladní, herečka i žena v domácnosti, modelka jsme se na cestě* za štěstím *dostali do slepé uličky.*

(38) *Děti zůstaly u otce a matka odešla* „za hlasem *svého srdce*“.

(39) *A budete pohromadě se spolupracovníky, se kterými půjdete společně* za užitečným cílem – dobrým zdravím.

(c) Určení místa: směru kam, k čemu – ke komu:

(40) *Matka se někde zdržela, otec šel* za kamarádem, *v televizi nedávali nic zajímavého…*

(41) *Při odchodu řekl, že* za ní *večer přijde do klubu.*

(42) *Žena vyšla na dvůr a dívala se* za mužem, *který se sekyrou na rameni mířil do lesa.*

(d) Vyjádření účelu, cíle, důvodu – úplná gramatikalizace:

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

(43) *Pro sbírku, kterou vyhlásil nadační fond* za účelem *obnovy kostela sv. Mikuláše, byl zřízen účet číslo…*

(44) *Přijela sem* za dobrodružstvím, *penězi, snad i najít si manžela, a věděla, že bude dostávat takovéhle návrhy.*

3 Pádový synkretismus

Disambiguační pravidla musí řešit problém pádového synkretismu mezi genitivem, akuzativem a instrumentálem syntaktického substantiva pádově řízeného předložkou „za“. Jde o tyto druhy synkretismu uvedených tří pádů:

Maskulina životná a neživotná:

vzor pán – tvar *pána*: Gsg / Asg[3]; *pány*: Apl / Ipl
vzor hrad – tvar *hrady*: Apl / Ipl
vzor muž – tvar *muže*: Gsg / Asg / Apl
vzor stroj – tvar *stroje*: Gsg / Apl
vzor předseda – tvar *předsedy*: Gsg / Apl / Ipl
vzor soudce – tvar *soudce*: Gsg / Asg / Apl
vzor mluvčí – tvar *mluvčí*: Gsg / Asg / Apl

Feminina:

vzor žena – tvar *ženy*: Gsg / Apl
vzor nůše – tvar *nůše*: Gsg / Apl; *nůší*: Isg / Gpl
vzor píseň – tvar *písně*: Gsg / Apl; *písní*: Isg / Gpl
vzor kost – tvar *kosti*: Gsg / Apl; *kostí*: Isg / Gpl
vzor mluvčí – tvar *mluvčí*: Gsg / Asg / Isg / Apl

Neutra:

vzor město – tvar *města*: Gsg / Apl
vzor moře – tvar *moře*: Gsg / Asg / Apl
vzor stavení – tvar *stavení*: Gsg / Asg / Gpl / Apl

Poznámka. Velmi často pomůže určit rekci předložky „za“, je-li řízené substantivum rozvito syntakticky adjektivním přívlastkem. Například v pádově homonymním spojení

za znamení

není jasný pád ani číslo substantiva „znamení“: může to být buď Gsg / Asg / Apl, tedy předložka „za“ může mít genitivní nebo akuzativní rekci. Je-li však substantivum „znamení“ rozvito syntakticky adjektivním přívlastkem:

za dobré znamení

je jasné, že „znamení“ je v Asg a předložka „za“ má akuzativní rekci.

[3] Genitiv singuláru vs. akuzativ singuláru. Podobně dále.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

4 Stanovení rekčního pádu

Naznačme tu alespoň základní myšlenky (na podrobnější rozklad tu není prostor), jak určit pádovou rekci předložky „za“.

4.1 Genitiv

Vzhledem k absenci možností sémantického zpracování textu, které by podobně jako člověk jednoznačně určilo rekci předložky „za“, se musíme uchýlit k formě a stanovit – na základě korpusových dat – typické kolokace genitivního časového/okolnostního určení. Následující spojení jsou v korpusu SYN2015 takřka se 100% spolehlivostí typu *za*-Gen – Subst-Gen:

za pomoci, za předpokladu, za války, za použití, za účasti, za stavu, za přítomnosti, za doprovodu, za tmy, za vlády, za soumraku, za jízdy, za svítání, za podpory…[4]

Lze tedy formulovat jednoduché disambiguační pravidlo, které s téměř 100% spolehlivostí určí v uvedených a některých dalších případech rekci předložky „za“ jako genitivní a v ostatních ji vyloučí.

4.2 Akuzativ

Vzhledem k množství funkcí akuzativní rekce předložky „za“, a tedy vysoké gramatikalizaci této rekce, naznačím opět jen nejfrekventovanější případy akuzativní rekce předložky na základě korpusu SYN2015:[5]

za porušení, za mříže, za zápěstí, za vítězství, za práva, za léta, za vedení, za energie, za hosty…

Pro takřka bezchybné určení akuzativní rekce předložky „za“ jsou klíčová slovesa-/substantiva umožňující/vyžadující předmětové nebo adverbiální doplnění ve formě „za“ + substantivum v akuzativu; následují typické struktury a konkrétní příklady:

považovat za-Acc Subst-Acc, *pokládat za*-Acc Subst-Acc, *vdát/provdat se za*-Acc Subst-Acc, *uvést za*-Acc Subst-Acc, *přijmout za*-Acc Subst-Acc, *označit za*-Acc Subst-Acc, *zvolit za*-Acc Subst-Acc…

považován za zakladatele, vdala se za muže, uvedla za pořadatele, šel za svědka, označen za zrádce, pojištění za zaměstnance, zvolen za člena, odehrál za hosty, prohlášen za kacíře, stojí za povšimnutí, děkuji za optání, chytne za srdce, ručí za závazky…

Z korpusu je možné snadno vytáhnout konkrétní spojení a zformulovat pravidlo, které v těchto spojeních určí pádovou rekci předložky „za“ jako akuzativní a pád řízeného substantiva jako akuzativ.

4.3 Instrumentál

Spojení „za“ + substantivum v instrumentálu má typicky místní význam, výjimečně význam jiný (*za účelem…*), gramatikalizace instrumentálové rekce předložky „za“ je

[4] Uvádím jen pádově homonymní případy substantiv, nikoli substantiva v jednoznačném pádě.
[5] Opět uvádím jen pádově homonymní případy substantiv.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

ve srovnání s rekcí akuzativní mnohem nižší. Následují typické příklady citované z korpusu SYN2015:

za zády, za prací, za hranicí, za zdí, za zuby, za nocí, za vesnicí, za minulostí, za kořistí.[6]

Pro takřka bezchybné určení instrumentálové rekce předložky „za“ jsou důležitá zejména slovesa/substantiva umožňující adverbiální doplnění ve formě „za“ + substantivum v instrumentálu; opět uvádím nejčastější příklady:

(*držet*) *jazyk za zuby, ruce za zády, noc za nocí, čára za minulostí, odejít/dojíždět za prací, domek za hrází, kousek za vesnicí, putování za zvířaty, alkohol za řídítky, stojí za kampaní…*

Z korpusu je opět možné snadno vydobýt konkrétní spojení a zformulovat pravidlo, které v těchto spojeních určí pádovou rekci předložky „za“ jako instrumentálovou a pád řízeného substantiva jako instrumentál.

5 Závěr

Představil jsem v nezbytné stručnosti způsob, jak metodou morfologické disambiguace založené na pravidlech co nejsprávněji, tj. takřka se 100% přesností, anotovat předložkové skupiny řízené trojrekční předložkou „za“. Úvahy obsažené v tomto drobném příspěvku, sepsaném na počest významného slavisty a bohemisty profesora Pognana, se promítnou do anotace připravovaných korpusů SYN2020 (122 milionů tokenů: 100 milionů slovních tvarů a 22 milionů interpunkčních znamének) a korpusu SYN v9 (celkem cca 5 miliard tokenů).

Literatura

KŘEN, M. CVRČEK, V., ČAPKA, T., ČERMÁKOVÁ, A., HNÁTKOVÁ, M., CHLUMSKÁ, L., JELÍNEK, T., KOVÁŘÍKOVÁ, D., PETKEVIČ, V., PROCHÁZKA, P., SKOUMALOVÁ, H., ŠKRABAL, M., TRUNEČEK, P., VONDŘIČKA, P. et ZASINA, A. (2015), *SYN2015: reprezentativní korpus psané češtiny.* Praha: Ústav Českého národního korpusu FF UK. Dostupný z WWW: http://www.korpus.cz

KŘEN, M. CVRČEK, V., ČAPKA, T., ČERMÁKOVÁ, A., HNÁTKOVÁ, M., CHLUMSKÁ, L., JELÍNEK, T., KOVÁŘÍKOVÁ, D., PETKEVIČ, V., PROCHÁZKA, P., SKOUMALOVÁ, H., ŠKRABAL, M., TRUNEČEK, P., VONDŘIČKA, P. et ZASINA, A. (2019). *Korpus SYN, verze 8 z 12. 12. 2019.* Praha: Ústav Českého národního korpusu FF UK. Dostupný z WWW: http://www.korpus.cz

[6] Opět uvádím jen pádově homonymní případy substantiv.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN : 9782813004260, doi : 10.17184/eac.9782813004260