Univerzita Karlova v Praze

Filozofická fakulta

Ústav českého jazyka a teorie komunikace

Bakalářská práce

Daniela Bodanská

Homonymie pasivních participií v současné češtině

Homonymy of the passive participles in modern Czech

Praha 2015 Vedoucí práce: Doc. RNDr. Vladimír Petkevič, CSc.

Děkuji vedoucímu práce, panu doc. RNDr. Vladimíru Petkevičovi, CSc., za všechny cenné rady, korekce, připomínky a především za jeho trpělivost a vstřícný přístup.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 13. května 2015

.....

Daniela Bodanská

Abstrakt

Práce popisuje a analyzuje různé typy slovnědruhové homonymie tvarů pasivních participií v současné češtině. Teoretickým základem jsou především mluvnice češtiny a odborné práce týkající se anotování korpusů, analyzovaným materiálem data z psaného synchronního korpusu (*SYN2010*). Ve druhé části práce jsou na základě formálních vlastností formulována disambiguační pravidla vybraných homonymních tvarů.

Klíčová slova

slovnědruhová homonymie, pasivní participia (trpná příčestí), disambiguace, tagování (značkování), tag (značka)

Abstract

The thesis describes and analyses various types of parts-of-speech homonymy of forms of passive participles in current Czech. The theoretical basis is particularly Czech grammar and specialized publications on annotating a corpora, the analysed materials are data from written synchronized corpora (*SYN2010*). In the second part of the thesis, disambiguation rules of selected types of homonymics are formulated based on formal characteristics.

Keywords

parts-of-speech homonymy, passive participles, disambiguation, tagging, tag

Obsah

Použité zkratky

ČM – Česká mluvnice

ČNK – Český národní korpus

FF UK – Filozofická fakulta Univerzity Karlovy v Praze

MFF UK – Matematicko-fyzikální fakulta Univerzity Karlovy v Praze

PMČ – Příruční mluvnice češtiny

SSJČ – Slovník spisovného jazyka českého

1 Úvod

Práce se zabývá disambiguací homonymních tvarů pasivních participií v korpusech Českého národního korpusu. Svým obsahem spadá do oblasti morfologie, syntaxe a počítačového zpracování přirozeného jazyka.

Teoretická část práce je rozdělena na dvě kapitoly. V první z nich jsou v chronologickém postupu představeny pohledy na pasivní participium v českých mluvnicích a dalších publikacích od padesátých let 20. století po počátek 21. století. Kapitola je zaměřena nejen na definice pasivního příčestí, ale především na to, jaké jsou podmínky jeho tvoření, jakou funkci ve větě plní a jaký vliv má přítomnost takového tvaru ve větě na své okolí. Každá podkapitola věnovaná jedné jazykovědné příručce je proto vždy rozdělena na část pojednávající o samotném tvaru pasivního příčestí a část popisující typické konstrukce s tímto tvarem.

Druhá kapitola vychází především z odborných prací členů Ústavu formální a aplikované lingvistiky a Ústavu teoretické a komputační lingvistiky a je věnována popisu anotování korpusů, konkrétně automatického morfologického značkování korpusů Českého národního korpusu. Je rozdělena do několika podkapitol odpovídajících jednotlivým fázím procesu automatického značkování, hlavní důraz je kladen na popis automatické morfologické disambiguace a systému *LanGr*, který je pro pravidly řízenou disambiguaci Českého národního korpusu užíván.

V praktické části je popsán současný stav značkování vybraných homonymních tvarů pasivních participií v korpusu psaného jazyka *SYN2010*, v případech značkování chybného jsou na základě formálních vlastností pasivních příčestí prezentovaných v teoretické části navrženy možné úpravy disambiguačních pravidel.

Cílem práce je nastínit rozsáhlou problematiku automatické disambiguace homonymních tvarů ve značkování korpusů a v konkrétních případech vlastními návrhy úprav přispět ke značkování přesnějšímu.

2 Pasivní participium v českých mluvnicích a jiných publikacích

Aby bylo možné zabývat se pasivními participii a homonymií jejich jednotlivých tvarů, je nejprve třeba podívat se do českých mluvnic (a dalších publikací), jak je pasivní participium definováno, jaká jsou omezení jeho tvoření, jakou roli ve větě plní a jaký je jeho vliv na okolí, v jakém kontextu se v češtině obvykle vyskytuje a v jakém se vyskytovat nemůže.

2.1 Mluvnice spisovné češtiny (Trávníček 1951)

2.1.1 Pasivní participia

Všechna participia, tedy participia ve smyslu užším (*nesl, nesen*) a přechodníky (*nesa, nes*), jsou v *Mluvnici spisovné češtiny* řazena k adjektivům, konkrétně k adjektivům dějovým, jelikož vyjadřují „vlastnost plynoucí z děje tkvícího v tom základě, z kterého jsou tvořena ona sama a příslušné určité tvary slovesné“ (Trávníček 1951, s. 1417) a také vyjadřují tři nebo dva mluvnické rody. Příčestí se však od ostatních adjektiv dějových odlišují, jednak tím, že je lze tvořit „z každého slovesného základu, z kterého jsou odvozeny tvary určité a jiné tvary neurčité“ (Trávníček 1951, s. 1417), jednak tím, že vyjadřují některé významy příznačné pro slovesné tvary určité, konkrétně vid, čas a slovesný rod. Navíc u sebe, na rozdíl od ostatních adjektiv dějových, mnohá příčestí mají zvratné *se* a *si* (*bál se, hrajíc si*). Proto je Trávníček zahrnuje do soustavy slovesných tvarů.

O participiích v užším smyslu (tedy bez přechodníků) mluvnice praví, že slovotvorně jsou odvozeninami pomocí přípon *-l, -n* a *-t* ze slovesných kmenů nebo kořenů. Rozděluje je na činná *(nesl*) a trpná (*nesen, bit*).

Pasivní participia mluvnice označuje za zpravidla dokonavá (vyjadřují vlastnost vyplývající z minulého děje), „nemající“ zvratné zájmeno (*najísti se – najeden, napíti se – napit*) a v hovorové a lidové mluvě ustupující před převažujícím užíváním tvarů složených (*jsem unavený, přišel udýchaný*).

2.1.2 Konstrukce s pasivními participii

Trávníček se zabývá především rolí pasivních participií v opisném pasivu. Rozlišuje dvojí rod slovesný – rod činný (aktivní) a rod trpný (pasivní); slovesa rodu činného nazývá činná (aktivní, aktiva), slovesa rodu trpného trpná (pasivní, pasiva). O trpný rod se jedná, pokud činitel děje není mluvnickým podmětem, jeho výrazovými prostředky může být pasivum zvratné nebo pasivum opisné.

Pokud je v pasivní větě přítomen mluvnický podmět, jedná se o trpný rod podmětový (*žák je chválen*), pokud přítomen není, jde o trpný rod bezpodmětý (*města bylo dobyto*). Zvratné pasivum může vyjadřovat pouze trpný rod podmětový, opisné oba, bezpodmětý však méně často. Trpným rodem se vyjadřuje stejný děj jako rodem činným, ale ze stanoviska pacientu, nikoli agentu (*žák je chválen od učitele*), podmětem trpného rodu podmětového tedy bývá patiens, u vět s pasivními tvary opisnými je „mluvnický podmět nositelem stavu plynoucího z činného slovesa“ (Trávníček 1951, s. 1365).

Činitel děje má v pasivu roli příslovečného určení (původce děje), většinou se ve větě nevyjadřuje, rozumí se z jejího obsahu (*bylo nařízeno* [nařídil úřad, policie…]). Vyjadřuje-li se, pak instrumentálem nebo (především pokud je činitel děje životný) genitivem s předložkou *od* (*žák byl pochválen učitelem*/*od učitele*). Trávníček uvádí i velmi ojedinělou možnost vyjádření činitele děje pomocí akuzativu a předložky *skrze* (*dlužník byl upomenut skrze advokáta*).

Podle mluvnice mohou mít trpný rod slovesa předmětová, především přechodná (přímá), řidčeji nepřechodná (nepřímá), stálá reflexiva většinou trpný rod nemívají, výjimkou je reflexivum *dotazovati se* (*odborník byl dotázán*) a ojedinělý archaismus *byl posmíván*.

2.2 Česká mluvnice (Havránek – Jedlička 1981)

2.2.1 Pasivní participia

Česká mluvnice popisuje tvoření pasivních participií – tvoří se od kmene minulého, příponami *–en, -n* a *-t* (*-ena, -na, -ta,…*); příponou *-en* (*-ena,…*) u sloves, jejichž minulý kmen končí na souhlásku nebo kmenotvornou příponu *-i-* (*nes-l – nesen, pros-i-l – prošen*), příponami *-n* (*-na,…*) a *–t* (*-ta,…*) u sloves, jejichž minulý kmen je zakončen jinou

samohláskou nebo samohláskou *-i-*, pokud ta není kmenotvornou příponou (*bra-l – brán, kry-l – kryt, bi-l – bit*).

Pasivní participia podle *České mluvnice* vyjadřují slovesný rod, číslo, jmenný rod a u rodu mužského životnost či neživotnost. V čísle a jmenném rodě se shodují s podmětem (v množném čísle mužského rodu mají podle podmětu i tvar životný nebo neživotný). Tvořit je lze především od sloves předmětových, od sloves bezpředmětových výjimečně (nemají pak význam trpného rodu a neužívá se jich v opisném pasivu, ale přechází v přídavná jména – *ovoce napadáno zůstalo ležet v sadu*). Výjimečná jsou také participia tvořená od sloves pouze zvratných (*tázán, dotazován, obáván, posmíván, vysmíván*).

2.2.2 Konstrukce s pasivními participii

Příčestí trpného se podle *České mluvnice* užívá v opisných tvarech trpného rodu, které mohou být určité (*je volán, buď volán*) i neurčité (*být volán, jsa volán, byv volán*), v opisném typu perfekta (*mám uvařeno*), samostatně v úloze přechodníku nebo přecházejí v přídavná jména slovesná (se ztrátou dějovosti i v přídavná jména).

I tato mluvnice uvádí, že se rod trpný vyjadřuje zvratnou podobou slovesa nebo opisným tvarem trpným, složeným z pasivního participia a tvarů slovesa *být* (řidčeji *bývat*). Trpným rodem bývá vyjádřen děj, v němž není agens podmětem, jak tomu bývá obvykle, ale je-li agens vyjádřen, je to buď 7. pádem (*byl pokárán matkou*) nebo (u osobního původce děje) pádem předložkovým (*byl pokárán od matky*). Nejčastěji však původce děje nebývá vyjádřen vůbec. Mluvnickým podmětem, pokud je ve větě vyjádřen, je patiens (*stroje byly vyrobeny*). Takto se předmět v pasivu stává podmětem u sloves přechodných (jde potom o tzv. pasivum osobní), u předmětových sloves nepřechodných zůstává předmět v témž pádě, věta je jednočlenná slovesná (tzv. pasivum neosobní – sloveso je užito bez podmětu, „neosobně“). Rod trpný vyjádřený neosobně užitým slovesem může vzácně být i u sloves bezpředmětových.

Opisný tvar pasivní je podle *České mluvnice* běžný především u sloves dokonavých, jelikož je v něm podstatný význam stavu jako výsledku činnosti (*kniha byla vydána*), není však nemožný ani u sloves nedokonavých (*knihy jsou vydávány*), zejména ve spojení s pomocným slovesem *bývat* (*knihy bývají vydávány*). V neosobním pasivu je opisný tvar zpravidla jen u sloves dokonavých (*bylo oznámeno*).

2.3 Mluvnice češtiny (Komárek – Kořenský – Petr 1986)

2.3.1 Pasivní participia

Popis tvoření pasivních participií v *Mluvnici češtiny* se od popisu v ČM příliš neliší: participia se tvoří z infinitivního kmene sufixy *-t* a *-n-* (*-en-*, je-li infinitivní kmen zakončen na souhlásku nebo jde-li o sloveso s infinitivní kmenotvornou příponou *-i-*/*-∅-*) a jmennými koncovkami. Před sufixem *-(e)n-* často dochází ke kvalitativním či kvantitativním alternacím (*péct*/*péci – pečen*, *zasáhnout – zasažen*).

I zde se uvádí, že se příčestí trpná tvoří zpravidla od sloves předmětových, především přechodných (a také že jen u přechodných sloves má participium trojí rod a dvojí číslo, u sloves nepřechodných má pouze tvar singuláru středního rodu – *tím mu nebylo pomoženo*). Slovesa bezpředmětová a pouze zvratná pasivní participia obvykle netvoří (až na výjimky – např. *tázán*, *obáván*).

2.3.2 Konstrukce s pasivními participii

Podle *Mluvnice češtiny* fungují pasivní participia primárně jako prostředek diateze v konstrukcích s tvary slovesa *být* (případně *bývat*) – v opisném pasivu. Dále bývají součástí rezultativních konstrukcí typu *máš uvařeno* a *dostal jsem vynadáno*. Samostatně, aniž by bylo součástí nějaké konstrukce, se trpné příčestí užívá jen zřídka, většinou je v takovém případě hodnoceno jako tvar trpného přechodníku (v němž je přechodníkový tvar pomocného slovesa vypuštěn – *dojat jeho slovy rozplakal se = jsa*/*byv dojat*). Také se užívá jako jmenná složka verbonominálního predikátu (*obloha je zatažena*, *zůstal zakryt*); v takovém případě konkuruje dlouhému tvaru slovesného adjektiva (*obloha je zatažená*, *zůstal zakrytý*).

Mluvnice češtiny uvádí poměrně ucelený přehled případů, kdy je (resp. není) možné tvořit opisné pasivum. Opisné pasivum, obdobně jako samo pasivní participium, se tvoří především od sloves přechodných (pasivum osobní), méně od předmětových sloves nepřechodných (pasivum neosobní), od nepředmětových sloves se netvoří. I u přechodných sloves však existují faktory, které tvoření pasivních tvarů omezují – opisné pasivum není vůbec možné tvořit od přechodných sloves vyjadřujících tělesné nebo duševní vztahy (*bolí ho hlava*, *hněte ho svědomí*) a od nedokonavých sloves vyjadřujících statické vztahy (*mám chalupu*, *prach pokrývá nábytek* – zde je možná participiální konstrukce u dokonavého

protějšku slovesa: *nábytek je pokryt prachem*). Též není možné tvořit opisné pasivum od procesuálních přechodných sloves vyjadřujících pohyby částí těla, jelikož zde je agens zahrnut i v patientu a nemůže tedy být odsunut (*sklopil oči*, *přehodil nohu přes nohu*); od dokonavých sloves tohoto typu je však možné tvořit participiální konstrukce se stavovým, rezultativním významem: *seděl nohu přehozenu přes nohu*. Od přechodných sloves s volným morfémem *si* (*dal si sklenici vína*, *oblíbil si toto město*) a od reflexiv tantum (*odvděčit se*, *smát se*) se participiální konstrukce také netvoří.

2.4 Příruční mluvnice češtiny (Karlík – Nekula – Rusínová 1995)

2.4.1 Pasivní participia

Příruční mluvnice češtiny upřesňuje tvoření pasivních participií v češtině: tvoří se dvěma způsoby, pokaždé od kmene minulého – jednak morfy *–t, -ta, -to, -ti, -ty, -ta* (ve shodě se jmenným rodem a číslem subjektu) od sloves typu *minout*, *krýt*, *začít*, jednak morfy *-(e)n, -(e)na, -(e)no, -(e)ni, -(e)ny, -(e)na* (ve shodě se jmenným rodem a číslem subjektu) od sloves ostatních typů. Tvoření tvarů první skupinou morfů bývá doprovázeno alternací typu *-nou-/-nu-* (*prominout – prominut*), u typu *krýt* alternací kvantity, tvoření sloves druhou skupinou morfů může být provázeno alternacemi základového konsonantu či skupiny konsonantů (např. *nosit – nošen*). Některá slovesa 2. třídy mohou mít tvary obojího druhu, ty však bývají významově odlišeny: *tisknut* („tlačen“) – *tištěn* (o knize).

I zde se podotýká, že pasivní příčestí lze tvořit od sloves předmětových.

2.4.2 Konstrukce s pasivními participii

Příruční mluvnice češtiny také udává spojování pasivních participií jak s tvary slovesa *být*, tak s tvary sloves *mít* a *dostat* (*mám vymalováno*, *dostal vyhubováno*). O opisném pasivu říká, že se uplatňuje především u sloves dokonavých (zatímco pasivum vyjádřené reflexivní slovesnou formou se užívá zejména u sloves nedokonavých).

Poměrně podrobně se zde popisují konstrukce se slovesy *mít* a *dostat* – konstrukcí *mít* + pasivní participium a *dostat* + pasivní participium se obdobně jako opisného pasiva užívá především k sekundární diatezi; zde ovšem nejde o deagentizaci typu agens – patiens, ale typu agens – recipient. Konstrukce *mít* + participium trpné je možná jen u sloves vyžadujících jako doplnění kromě sémantických rolí agentu a recipientu ještě

sémantickou roli patientu, osoby či věci dějem zasažené, vyjádřenou akuzativní formou. Konstrukce *dostat* + trpné příčestí je zase vázána na vyhraněné sémantické skupiny sloves – na slovesa s recipientem v dativu a předmětem v akuzativu s významem „způsobit, aby někdo něco měl“ (*přidat*, přidělit, *zaplatit*,... *někomu něco*; *přikázat*, *nařídit*,... *někomu něco*) a na slovesa s recipientem v dativu bez předmětu v akuzativu jako *namlátit*, *nařezat*, *nafackovat*,... *někomu* a *vyhubovat*, *vynadat*,... *někomu*. Věty s druhou skupinou sloves mohou mít pouze inkongruentní tvar příčestí, jelikož neobsahují pozici předmětu (*dostal jsem vyhubováno*).

Participiálních konstrukcí lze užívat i jako syntaktických kondenzací (*Kdo na rychtářku něco věděl, umlkl teď náhle, dojat její žalostí*.), příčestí zde má knižní charakter, zvláště zachovává-li jmenný tvar.

2.5 Encyklopedický slovník češtiny (Karlík – Nekula – Pleskalová 2002)

2.5.1 Pasivní participia

Z *Encyklopedického slovníku češtiny* se o způsobu a omezení tvoření pasivních participií dozvíme prakticky totéž, co z výše uvedených mluvnic. Najdeme zde však zmínku o flexi pasivních participií – ve spojení s tvary pomocného slovesa být mají flexi buď jako krátké (jmenné) tvary adjektiv (ta jim umožňuje vyjadřovat shodu s podmětem – *žák je zkoušen*) nebo, vzhledem ke svým adjektivním vlastnostem, i jako dlouhé (složené) tvary adjektiv (*žák je zkoušený*). Ve spojení s tvary sloves *mít* a *dostat* jsou participia vybavena flexí umožňující vyjadřovat kongruenci s předmětem v akuzativu (*Petr má slíbenu*/*slíbenou odměnu*).

2.5.2 Konstrukce s pasivními participii

Užívání pasivních příčestí se zde omezuje pouze na součásti složených tvarů, a to morfologické formy opisného pasiva, konstrukce sponově-jmenné, rezultativní typy *je vyzkoušeno*, *má vyzkoušeno* a složené tvary typu *dostal vyhubováno*.

Většina věcí, které se o opisném pasivu dozvíme z *Encyklopedického slovníku češtiny*, souhlasí s údaji z ostatních publikací. Výjimkou je zejména popření tvrzení uváděných např. v ČM či PMČ (viz výše), že se opisné pasivum užívá hlavně u sloves dokonavých, jelikož vyjadřuje především stav (na rozdíl od pasiva zvratného, vyjadřujícího

spíše děj, a tedy užívaného zvláště u nedokonavých sloves). Podle *Encyklopedického slovníku češtiny* toto není pravda, „vid při komunikativní volbě prostředků deagentizace nehraje v č. žádnou roli“ (Karlík – Nekula – Pleskalová 2002, s. 379) a např. věta s opisným pasivem *Tyto obrazy jsou právě restaurovány*. vyjadřuje děj stejně jako příslušná věta se zvratným pasivem.

3 Anotování Českého národního korpusu

V procesu značkování (anotování) korpusů se jednotkám textu přiřazují lingvistické informace, formálně vyjádřené prostřednictvím značek (tagů). Značkování korpusů je významné pro vědecké poznání jazyka, umožňuje lingvistům zkoumat jazykové jevy z různých pohledů (např. frekvenci výskytu jednotlivých jevů či kontext, v němž se obvykle vyskytují), na různých rovinách (fonetické, morfologické, syntaktické, sémantické). Nejrozšířenějším typem značkování korpusů je značkování morfologické, prováděné většinou automaticky a zahrnující i automatickou disambiguaci. Následující kapitola, vycházející především z odborných prací členů Ústavu formální a aplikované lingvistiky (MFF UK) a Ústavu teoretické a komputační lingvistiky (FF UK), popisuje tento typ anotování Českého národního korpusu (se zaměřením na popis morfologické disambiguace), bez jehož porozumění by nebylo možné věnovat se dále jednotlivým značkovaným tvarům a formulovat konkrétní disambiguační pravidla.

3.1 Tokenizace, větná segmentace

Tomáš Jelínek a Vladimír Petkevič popisují první dvě fáze morfologického značkování korpusů v širším smyslu, tokenizaci a větnou segmentaci (Jelínek – Petkevič 2011, s. 155).

Během tokenizace se identifikují jednotlivé tokeny, tedy veškeré korpusové pozice – slova, zkratky, interpunkční znaménka apod. K této operaci je používán program zvaný tokenizér, který může být dvou druhů – statistický či založený na pravidlech. Na počátku operace má tento program k dispozici posloupnost mezer a řetězců bez mezer a zpracováváním jednotlivých řetězců mezi mezerami vytváří tokeny. V ideálních případech je tokenem celý řetězec, v jiných případech musí program rozpoznávat nesmysluplné celky, které jsou spojením dvou či více lingvistických prvků, např. slovem a interpunkcí těsně za ním. Program musí umět především zpracovávat víceznačná interpunkční znaménka (rozeznávat zkratky končící tečkou od samostatné interpunkce apod.). Při značkování korpusů řady *SYN*... byl použit na pravidlech založený tokenizér Pavla Květoně.

Při větné segmentaci program segmenter (též segmentátor) rozdělí řetěz tokenů do vět. Řídí se přitom především interpunkčními znaménky zakončujícími větu (přičemž se opět musí vypořádat s jejich častou víceznačností) a velkými písmeny, jimiž běžně věta

začíná. Segmentery také mohou být dvou druhů – statistické a na pravidlech založené; pro značkování korpusů řady *SYN*... je, obdobně jako u tokenizace, užíván pravidly řízený modul Pavla Květoně.

3.2 Morfologická analýza

Morfologická analýza, která přichází na řadu poté, co je text rozdělen do vět, přiřazuje každému slovnímu tvaru všechny jeho možné (na kontextu nezávislé) slovnědruhové a morfologické interpretace. Je prováděna automatickým morfologickým analyzátorem (pro značkování korpusů ČNK vytvořeným autory z MFF UK – Jan Hajič a kol.), jenž na základě morfologického slovníku přiřadí každému tokenu alespoň jednu morfologickou interpretaci, tedy dvojici (*lemma*, *značka*).

Lemma je obvykle základní tvar lexému (tedy ten, který zpravidla nalezneme ve slovnících, např. nominativ singuláru u jmen nebo infinitiv u sloves), k němuž token jakožto slovní tvar patří, morfologickou značkou (tagem) je vyjádřen slovní druh a morfologické vlastnosti tokenu. Morfologická značka je zaznamenána v podobě řetězce znaků, konkrétně ve značkování korpusů Českého národního korpusu je každý znak označením hodnoty z jedné určité morfologické kategorie. Při značkování korpusů ČNK se užívá tzv. pozičního značkování – tagy mají konstantní délku, každý je řetězcem 15 znaků (v korpusech *SYN2005* a *SYN2006PUB* je dostupná i 16. pozice), každá morfologická kategorie zaujímá v řetězci neměnnou pozici (pozice se neposouvají, ani pokud nějaký slovní tvar danou morfologickou kategorii nevyjadřuje). V následující tabulce vytvořené podle popisu morfologických značek Jana Hajiče (Hajič 2000) jsou uvedeny významy jednotlivých pozic tagů v korpusech ČNK:

Pozice	Název pozice	Význam
1	POS	Slovní druh[1]
2	SUBPOS	Detailní určení slovního druhu
3	GENDER	Jmenný rod
4	NUMBER	Číslo
5	CASE	Pád
6	POSSGENDER	Přivlastňovací rod
7	POSSNUMBER	Přivlastňovací číslo
8	PERSON	Osoba
9	TENSE	Čas
10	GRADE	Stupeň
11	NEGATION	Negace
12	VOICE	Aktivum/pasívum
13	RESERVE1	Nepoužito
14	RESERVE2	Nepoužito
15	VAR	Varianta, stylový příznak apod.
16	ASPECT	Vid[2]

3.3 Morfologická disambiguace

Narazí-li morfologický analyzátor na slovní tvar, který má v morfologickém slovníku dvě či více morfologických interpretací (takových tvarů je značné množství, přibližně dvě třetiny – Jelínek – Petkevič, s. 158), přiřadí mu odpovídající počet dvojic (*lemma, tag*). Je na další fázi, tedy na morfologické disambiguaci (neboli zjednoznačnění), aby z poskytnuté nabídky podle kontextu vybrala u homonymního tvaru správné lemma a správnou morfologickou značku.

[1] „Označuje hlavní slovní druh, víceméně podle obvyklého schématu známého z českých gramatik včetně školních. Přiřazení i těchto hlavních slovních druhů je však řízeno především potřebami konzistentnosti další analýzy přirozeného jazyka. Proto je možné, že v některých případech (zejména tehdy, kdy se gramatiky a slovníky v určení slovního druhu neshodují nebo uvádějí jiné rozdělení na významy slova nebo tam, kde ve slovníku najdeme slovnědruhové perly typu "zájmenné příslovce") nemusí být zařazení zcela "tradiční".“ (Hajič 2000)

[2] „Tato pozice byla k původní sadě doplněna Miroslavem Spoustou na základě slovníku morfologické analýzy. Je dostupná pouze v korpusech SYN2005 a SYN2006PUB.“ (Hajič 2000)

Morfologická disambiguace se může provádět následujícími metodami:

(a) statisticky (stochasticky),

(b) pomocí pravidel,

(c) tzv. hybridní metodou.

3.3.1 Statistická metoda

Statistická metoda je založena na tzv. strojovém učení disambiguačních programů – ty se „natrénují“ na „trénovacích datech“, textech, jež byly označkovány ručně (a předpokládá se tedy, že správně). To, co si takto program „natrénuje“ na malém množství označkovaných dat, poté aplikuje na data dosud neoznačkovaná, zpravidla mnohem rozsáhlejší.

3.3.2 Metoda založená na pravidlech

Metoda založená na pravidlech vychází ze zapojení lingvistických pravidel formulovaných na základě intuice a jazykových znalostí lingvistů a ověřovaných na korpusových datech (případně z korpusových dat rovnou vyvozovaných). Pro ČNK je užíván systém pravidly řízené redukční automatické morfologické disambiguace *LanGr* (srov. Květoň 2006) obsahující gramatická pravidla formulovaná v podobě programů.

Morfologická disambiguace založená na pravidlech užívá z velké části negativního přístupu k jazykovému systému, vychází tedy z toho, co jazykový systém neumožňuje, „zkoumá zejména negramatické strukturní a slovosledné kombinace lemmat a tagů, a ty pak u jednotlivých značkovaných slovních tvarů odstraňuje – proto tento typ disambiguace nazýváme redukční“ (Jelínek – Petkevič 2011, s. 160).

Podle Karla Olivy (Oliva 2005, s. 232) je negramatičnost „vždy důsledkem porušení určitého jazykového jevu (tj. porušení zákonitostí, které výskyt tohoto jevu s sebou přináší: např. nutnost shody podmětu s přísudkem v osobě a čísle apod.)“. Dále autor vymezuje následující třídy jevů běžně se vyskytujících ve větách (Oliva 2005, s. 232):

- Jevy selekční, kde selekci (jakožto zobecnění pojmu valence) chápe jako „požadavek nutného výskytu určitého prvku (slova) E1 ve větě, pokud se v téže větě vyskytuje slovo E2 (příp. skupina slov {E2, E3, …, En})“ (Oliva 2005, s. 232). Pokud se tedy ve větě vyskytuje slovo E2 (resp. skupina slov {E2, E3, …, En}), ale slovo E1 se v ní

nevyskytuje, je porušena zákonitost vyžadovaná selekčním jevem a věta je považována za negramatickou.

- Jevy slovosledné, u nichž jsou podstatné jazykové zákonitosti definující vzájemný slovosled dvou či více slov; pokud požadovaný slovosled není dodržen, je opět porušena příslušná zákonitost a věta je negramatická.

- Jevy souvýskytu rysů (souvýskytem rysů míní zobecnění pojmu shoda), u nichž je podstatná zákonitost, že při společném výskytu dvou či více slov v kontextu jedné věty musí být jisté morfologické charakteristiky těchto slov v určitém vzájemném vztahu (kterým je obvykle identita, není to ale podmínkou, vztahy mohou být i mnohem složitější). Pokud toto není splněno, věta je taktéž považována za negramatickou.

Na základě zákonitostí o porušování (nejen) těchto jevů je možné sestavit soubor pravidel a negativních omezení na úrovni morfonologické, morfologické, syntaktické i sémantické, který tvoří tzv. disambiguační gramatiku. Tato gramatika se značně liší od tradičních gramatik dosavadních, jelikož je určena především pro počítačové zpracování, a nemůže se tedy spoléhat na intuici, jazykové povědomí či znalosti člověka, ale musí zcela objektivně a detailně postihovat jazykový systém.

Při tvorbě negativních pravidel v tomto přístupu se částečně vychází z tzv. negativních n-gramů (pro n>1), několika (těsněji i volněji) sousedících slov a jejich gramatických a slovosledných vlastností, jež jsou v jazykovém systému češtiny vyloučené. Nejjednodušší pravidla jsou založena na negativních bigramech (n=2) tvořených bezprostředně sousedícími slovními tvary v daném pořadí.

3.3.2.1 Systém LanGr

Podstatným požadavkem na pravidly řízenou morfologickou disambiguaci je možnost všechna pravidla (ať už ta n-gramového rázu či jiná) naprosto přesně a jednoznačně formulovat pomocí příkazů v automatickém počítačovém programu. Pro disambiguaci ČNK k tomu slouží systém *LanGr* vytvořený Pavlem Květoněm (Květoň 2006). Důležitou vlastností tohoto systému je, že spolu jednotlivá pravidla kooperují – jakmile nějaké pravidlo odstraní jednu či více nesprávných morfologických interpretací, pracuje již každé pravidlo aplikované později s textem disambiguovanějším.

Pravidla systému *LanGr* se dají rozlišovat podle různých vlastností, významným dělením je odlišení pravidel bezpečných, jejichž platnost je téměř stoprocentní, a pravidel heuristických, jejichž platnost je nižší a mohou někdy chybovat. Systém vždy využívá nejdříve všech pravidel bezpečných, a teprve když ta už nemohou ve větě nic disambiguovat, použije pravidla heuristická (po jejichž uplatnění opět přicházejí na řadu pravidla bezpečná a systém takto pracuje až do doby, kdy už není schopen nic disambiguovat ani jednou skupinou pravidel).

Každé pravidlo formulované v systému *LanGr* se skládá ze dvou částí – disambiguačního místa (na němž dochází ke změnám dat, tedy k vlastní disambiguaci) a kontextu. Vladimír Petkevič (Petkevič 2004, s. 144) zjednodušeně popisuje, jak systém jednotlivá pravidla zpracovává: „Kontext musí být vždy jednoznačně specifikován na základě speciální klauzule *IsSafe*, která stanoví, že každá interpretace příslušného slovního tvaru/tvarů musí mít specifikované vlastnosti. Na disambiguačním místě dochází ke změnám; je to oblast slovních tvarů, které mohou být slovnědruhově a morfologicky víceznačné, a specifikuje se pomocí klauzule *Possible*, která stanoví, že aspoň jedna z morfologických interpretací daných slovních tvarů musí mít specifikované vlastnosti.“ Výkonná část každého pravidla má v systému dva základní příkazy – *DELETE*, jenž se uplatňuje při negativním přístupu a odstraňuje některé (ne nutně všechny) nesprávné interpretace, a *LEAVE ONLY*, uplatňující se při pozitivním přístupu ponecháním správných interpretací.

Důležitou součástí systému pravidly řízené disambiguace je také kolokační komponent, modul pro zpracování kolokací, frazémů a idiomů, který se zabývá typickými slovními spojeními a různými výjimkami. Modul spolupracuje se systémem *LanGr* tím, že vyhledaná slova tvořící kolokace či frazémy rovnou disambiguuje; zároveň pro svou práci také využívá výsledků systému *LanGr*.

3.3.3 Hybridní metoda

Hybridní metoda je založena na spojení metody statistické a metody založené na pravidlech. Ukázalo se (Jelínek – Petkevič 2011, s. 169, s odkazem na disertační práci – Spoustová 2007), že nejlepší metodou (tedy s co nejmenším počtem chybných značek) pro konečnou disambiguaci českých korpusů je právě tato metoda (při co největším zapojení pravidly řízeného taggeru). Konkrétně korpusy *SYN2010* a *SYN* jsou

disambiguovány spojením statistického taggeru *Morče* a na pravidlech založeného systému *LanGr* touto posloupností kroků (Jelínek – Petkevič 2011, s. 169):

1. Na výstup morfologické analýzy se spustí bezpečná pravidla systému *LanGr*, která postupnou disambiguací redukují počet nesprávných tagů v jednotlivých větách. („Pravidla jsou formulována tak, aby pokud možno nedošlo k vymazání správné interpretace slovní formy; v případě, že je věta příliš složitá, ponechají se všechny tagy“ – Skoumalová 2011, s. 201.)

2. Poté, co pravidla vyčerpají své možnosti disambiguovat, spustí se na výstup z prvního kroku kolokační/frazémový modul, který ve větách identifikuje a disambiguuje kolokace a frazémy.

3. Na výstup z druhého kroku se opět spustí pravidla systému *LanGr*, tentokrát všechna, tj. bezpečná i heuristická, která opět v opakovaných cyklech probíhají, dokud je co disambiguovat.

4. Zbylé nesprávné značky jsou odstraněny statistickým taggerem *MorČe*, který u každého tvaru ponechá právě jedno lemma a jeden tag.

4 Pasivní participia v korpusu

4.1 Typy morfologické homonymie

Vladimír Petkevič (Petkevič 2005, s. 250) rozlišuje typy homonymie, které nás při pravidly řízené morfologické disambiguaci zajímají:

(a) homonymie náhodná, tedy „možnost vícenásobně morfologicky interpretovat slovní tvar, přičemž možnost vícenásobné interpretace není dána nějakým vztahem mezi interpretacemi“ (Petkevič 2005, s. 250). Rozlišují se zde případy, kdy slovní druh patří:

a. k více lexémům různých slovních druhů (*bez, hořce, brána*) nebo
b. k více lexémům téhož slovního druhu (*hrách, spal*) nebo
c. se kombinuje a. a b. (*děl*)

(b) homonymie systémová, která se také dále rozlišuje na:

a. homonymii v deklinačním či konjugačním paradigmatu téhož lexému – tzv. vlastní homonymie daná kombinací čísla a pádu (*město*) a
b. homonymii mezi různými slovními druhy (*nalezení*).

Tato práce se zabývá homonymií náhodnou, konkrétně jejím prvním podtypem, homonymií slovnědruhovou.

4.2 Slovnědruhová homonymie pasivních participií

Můžeme vymezit několik skupin slovnědruhové homonymie pasivních participií podle toho, k jakému slovnímu druhu tvar homonymní s pasivním příčestím patří – může jít o tvary patřící k lemmatům substantiv, adjektiv, numerálií, adverbií, konjunkcí a interjekcí (případně ještě k jiným tvarům sloves, pak už ale nejde o homonymii slovnědruhovou, která je předmětem práce).

Následující kapitola se bude postupně zabývat těmito skupinami slovnědruhové homonymie. Z každé z nich[3] budou vybrány tvary, u nichž se podíváme na stav značkování v současném korpusu *SYN2010*, a objeví-li se tvary značkované chybně, pokusíme se

[3] Homonymií pouze mezi pasivními participii a adjektivy se zde zabývat nebudeme, jelikož hranice mezi příslušností jednotlivých tvarů k pasivním participiím či jmenným tvarům adjektiv je často velmi nejasná, jde o problematiku hodnou značně podrobnějšího průzkumu, překračujícího rozsah práce,

navrhnout možnou úpravu pro značkování korpusů příštích[4]. Data jsou analyzována na základě korpusu *SYN2010*, jelikož se jedná o korpus referenční – data tedy bude možné kdykoliv zpětně dohledat – a žánrově vyvážený. Korpus *SYN2010* je také dostatečně rozsáhlý, aby výsledky mohly být relevantní, ale na rozdíl od celého korpusu *SYN* není zase natolik rozsáhlý, aby podrobnější práce s ním byla prakticky nemožná, nebo alespoň velmi zdlouhavá.

Výběr konkrétních analyzovaných tvarů se řídí několika kritérii – vybírány byly jednak takové tvary, které se reálně řadí k lemmatům obou slovních druhů[5] a jsou v korpusu opravdu značkovány jako tvary pasivního participia i lemmatu patřícího k druhému slovnímu druhu, jednak takové, u nichž bychom výskyt u obou slovních druhů předpokládali, ale v korpusu jsou všechny řazeny jen pod jedno lemma (ať už správně či chybně), nebo naopak ty, které jsou jen jednoho slovního druhu, ale tagger je řadí k oběma. Dále se výběr tvarů řídil četností jejich užití v rámci korpusu, vybírány byly tvary často užívané (z nich zejména ty, v jejichž tagování byly nalezeny chyby), případně pak i další tvary jiných čísel a jmenných rodů téhož slovesa, i když ty už třeba tolik výskytů v korpusu nemají. Zvoleno však bylo i několik výrazů z jazykové periferie, které jsou zajímavé především tím, jaké tagy jim jsou v korpusu přiděleny.

4.2.1 Substantiva

4.2.1.1 Tvar *brány*

Tvar *brány* se v korpusu *SYN2010* vyskytuje 3936krát, z čehož 3826krát je označkován jako substantivum lemmatu *brána* a jen 110krát jako verbum lemmatu *brát*.

Tvary označkované jako pasivní příčestí se v naprosté většině případů vyskytují ve spojeních s tvary slovesa *být* a výrazy jako *v potaz*, *v úvahu*/*do úvahy*, *doslova*, *ohledy*, *vážně*, *útokem*, *na*/*v zřetel*, *v pochybnost*, *za zaručené*/*jisté*/*pravdivé*…, *na lehkou*/*těžkou váhu*, *jako urážka*/*samozřejmost*/*úspěch*… aj. Nalezneme zde jen dva případy chybného přiřazení tvaru *brány* ke slovesnému lemmatu. První z nich je ve spojení *kolem Zelené brány*; zde by tag snad mohl být opraven přidáním toponyma *Zelená brána* do slovníku kolokací, frazémů a idiomů, díky čemuž by kolokační modul mohl tvar *brány* rozpoznat

[4] Pravidla pro možnou úpravu některých tagů navrhovaná v této kapitole budou formulována pouze lingvisticky, nebudou zde vyjadřována v takových podobách, v nichž by bylo možné je okamžitě aplikovat v systému *LanGr*. Předpokládáme však, že všechna je možné v systému *LanGr* počítačově implementovat, jak bylo popsáno výše.

[5] Tedy ne tvary jako např. *Pluto* či *Rosen*, jejichž reálné užití jako pasivních participií, navíc ještě při zápisu s velkým počátečním písmenem, je velmi nepravděpodobné (a automatický tagger je také značkuje jednoznačně a bez chyb).

a správně zařadit jako substantivum (tvar by mohl být určen jako substantivum i díky tomu, že před ním stojí shodné adjektivum). Druhý nesprávně označkovaný tvar je ve větě *…Šavol jeho pokus vyškrábl z růžku brány.* V tomto případě by mohlo být možné odstranit chybnou značku díky předpokladu, že se ve větě již vyskytuje jiné sloveso, které je jednoznačně určitelné jako sloveso (není s ničím homonymní) a zároveň se (už jen kvůli tomu, že je ve tvaru singuláru maskulina) nemůže pojit s příčestím *brány*.

Mezi tvary otagovanými jako substantiva také najdeme několik tvarů otagovaných chybně (i když, vzhledem k jejich celkovému množství, je těch s chybnou značkou poměrně málo). Ty z nich, které mají tag nominativu, by bylo možné vyloučit podle předpokladu, že je-li ve větě (resp. klauzi) jiné substantivum než tvar *brány* označkované jako femininum nebo neživotné maskulinum v plurálu nominativu, tvar *brány* nebude pravděpodobně substantivem (nejde-li o několikanásobný podmět), jelikož v jedné klauzi obvykle bývá jen jedno substantivum v nominativu. Takto můžeme upřesnit tag například v těchto větách: *S přirozeným citem rodičů se kalkuluje tam, kde děti jsou brány za rukojmí soukromými nebo veřejnými teroristy…, Existující nerovnosti jsou brány s přírodní nutností., I v takovém případě jsou molekuly brány za oxidované., …lékařské důkazy byly opět brány mezi prvními.* Ve třetím případě by mohlo být správnému otagování nápomocné i typické spojení pasivního participia, slovesa *brát*, předložky *za* a akuzativního tvaru, v prvním případě už bohužel nikoliv, jelikož tvar *rukojmí* není označkován správně jako akuzativ, ale jako instrumentál. I v dalších případech by tag bylo možné opravit nejen díky přítomnosti jiného substantiva v odpovídajícím tvaru ve větě, ale také na základě toho, že se tvar *brány* nachází v klauzi společně s odpovídajícím tvarem slovesa *být* a nějakým z výrazů typicky se pojících s pasivními participii slovesa *brát – v úvahu* (*V úvahu jsou brány všechny události, které mohou nastat.*, *…které musí být brány při konstrukci a montáži v úvahu.*), *do úvahy* (*Ekologie krátkých letů a centrální pozice v Evropě budou brány více a více do úvahy…*), *vážně* (*…informace nebyly brány ÚOOZ a případně BIS vážně…*, *…jak vážně byly u nás brány myšlenky tohoto učení.*), *ohledy* (*…aby byly brány ohledy na jeho soukromí a stud.*), *jako* + tvar v nominativu (*…přičemž poslech hudby jakou byla Mejtusova Mladá garda či kompozice Čulakiho a dalších autorů toho rangu byly brány až jako recese., …protože děti jsou brány mnohými rodiči jako přítěž…*) a *za* + tvar v akuzativu (*…za základ byly brány nulové hladiny v nestejné výšce…*, *Ženy jsou v politice brány spíše za ozdobu…*).

4.2.1.2 Tvar *brána*

Tvar *brána* má v korpusu 3261 výskytů, z nichž 108 je označkováno jako pasivní participium. Většinou jde o podobná spojení jako u tvaru *brány* (*v potaz, doslova, v úvahu*/*do úvahy,…*), nalezneme zde však více chyb než u tvaru předchozího – přibližně pětina tvarů označkovaných jako pasivní participia je označkována chybně. Několikrát se tak děje ve spojeních, kde je tvar *brána* součástí toponym: *Česká brána*, *Lesní brána*, *Mariánská brána*, *Matyášova brána*, *Nebeská brána*, *Poděbradská brána* a *Sedlecká brána*. Zde by bylo možné chybné tagy odstranit buď přidáním celých výrazů do slovníku kolokací, nebo za předpokladu, že adjektiva začínající uprostřed věty (jen v jednom případě adjektivem věta začíná) velkým písmenem budou pravděpodobně částí nějakého víceslovného vlastního jména, jehož součástí bude i tvar *brána* (i v těchto případech, podobně jako u spojení *Zelené brány*, by navíc tvary mohly být správně určeny jako substantiva i díky tomu, že těsně před nimi stojí shodný přívlastek). Se zapojením kolokačního komponentu by také mohlo být možné opravit značky v ustálených spojeních *světelná brána*, *Brána Firewal* a *mýtní brána*, případně i ve spojení *není ale brána jako brána*, v němž byl pravděpodobně první z tvarů *brána* určen jako pasivní participium kvůli tomu, že je následován výrazem *jako*, který s tvary slovesa *být* a pasivními participii od slovesa *brát* často tvoří různé fráze (viz výše).

Další případy, v nichž by se tagy daly opravit, jsou takové věty, které již obsahují jiný přísudek, a navíc v nich není jiné substantivum v nominativu, které by mohlo být podmětem: *Podle Dismanovy veduty města z doby kolem roku 1730 vypadala brána jako zaklenutý průjezd.*, *Brána na své vnější straně obsahovala tři půlkruhové stejně veliké portály.*, *Brána podobně jako opevnění města zanikla ve třicetileté válce.*, *Brána jako součást sousedního domu je veřejnosti nepřístupná.* (toto kritérium by ovšem bylo stoprocentně spolehlivé pouze tehdy, kdyby byl automatický tagger schopný brát v potaz širší kontext než jednu větu a mohl tak rozpoznat, zda tvar *brána* skutečně musí být podmětem – a tedy substantivem –, nebo jde-li o elipsu a podmět se nachází v jiné větě. Obdobně by se dal tvar správně určit i ve větách *Podle plánu z 19. století měla brána jako součást průjezdu do města vnitřní lomený oblouk…* a *My jsme měli mnohem víc ještě vyloženějších příležitostí, ale brána jako by pro nás byla zakletá.*, v nichž také při určení tvaru brána jako pasivního participia v klauzi chybí tvar, který by mohl být podmětem (shodujícím se s přísudkem), ve druhém případě se navíc přísudek nemůže vztahovat ani k podmětu klauze první, jelikož s ním není ve shodě.

Mezi tvary označkovanými jako substantiva také nalezneme několik (i když velmi málo) případů značkování chybného. Jeden z nich je ve větě *...má národopisná studia jsou brána tak vážně...*; zde se nachází jak neutrum v plurálu nominativu, tak typické spojení tvaru slovesa *být*, pasivního participia od slovesa *brát* a výrazu *vážně*, není tedy žádný zjevný důvod, proč by tagger neměl tvar *brána* vyhodnotit jako pasivní participium. Další tři chybné tagy se nacházejí u tvarů následovaných předložkou *za* a akuzativem substantiva nebo adjektiva: *Společně objektivizovaná realita je nakonec natolik brána za jistou, že…*, *Verze je spíše – hovoříme-li nezávazně a nesnažíme-li se zodpovědět ani Pilátovu ani Tarského otázku – brána za pravdivou, není-li v rozporu s žádnými z přesvědčení, která…*, *Tenhle svět „malého státu" a nedotknutelné privátní iniciativy, jež je dnes v USA brána za ucho a fakticky znárodňována (a její „ toxické dluhy" mají být uhrazeny z astronomicky deficitního eráru), kontaminuje přece u nás programy všech stran*. Tato spojení jsou také typická pro věty s pasivním participiem slovesa *brát*, mohl by je tedy tagger (případně jeho kolokační komponent) brát v úvahu, ve všech případech jsou navíc přítomny tvary slovesa *být* i substantiva ženského rodu v nominativu shodující se s participiem *brána*. Poslední nalezený tvar pasivního participia otagovaného jako substantivum je v souvětí *Máte za sebou v posledních týdnech testování ve formuli GP2, která je brána pro piloty jako příprava pro přechod do formule 1*. Tento případ je poněkud složitější – kdyby druhá věta zněla *...která je pro piloty brána jako příprava pro přechod do formule 1.*, bylo by pro tagger pravděpodobně jednodušší vyhodnotit tvar *brána* ve spojení *brána jako* správně jako participium. Takto je však nutné, aby tagger spojení vyhledal a vyhodnotil i v širším kontextu, v němž už není příliš typické (jen v jednom případě je v korpusu tvar *jako* až na třetí pozici za pasivním participiem *brána*, navíc ve spojení *brána v potaz jako*, kde je velmi pravděpodobně určující spíš výraz *v potaz* než *jako*).

4.2.1.3 Tvar *bránu*

U tvaru *bránu* je v korpusu *SYN2010* nalezeno 1348 výskytů, z nichž jsou až na dva všechny zařazeny jako substantiva pod lemma *brána*. V obou případech, kdy je tvar označen za pasivní participium, se jedná o značkování chybné. V prvním případě, v němž jde o spojení *Měnínskou bránu*, by opět bylo možné chybu odstranit přidáním toponyma *Měnínská brána* do kolokačního slovníku, v případě druhém – *...viděl jenom okno a auto, které táhlo startovací bránu na startovní čáru.* – je to složitější, jelikož i tvar *táhlo* je určen chybně jako substantivum neutra v 1. pádě singuláru. Při správném určení výrazu *táhlo* jako určitého slovesného tvaru by neměl být tvar *bránu* otagován jako pasivní příčestí,

jelikož sloveso *táhnout* (užité ve významu „přemisťovat něco tahem“) obligatorně vyžaduje valenční doplnění ve 4. pádě (tím by teoreticky mohl být i tvar *čáru*, sloveso *táhnout* však zároveň – ne obligatorně, ale typicky – vyžaduje ještě volné doplnění směrové, jímž je právě spojení *na čáru*). Navíc, pokud by byl výraz *táhlo* určen správně jako sloveso, nemohlo by případné pasivní participium *bránu* stát mezi jím a předmětem v akuzativu.

Tvar *bránu* se tak v korpusu *SYN2010* s nejvyšší pravděpodobností vůbec nevyskytuje jako tvar slovesný, ani po prohledání tvarů zařazených pod lemmatem *brána* nebyl žádný nalezen (např. po odfiltrování všech tvarů, v jejichž bližším okolí se nenachází žádné substantivum ženského rodu v akuzativu singuláru, jež by se mohlo s participiem *bránu* shodovat, byly stejně všechny zbylé výskyty substantivy).

4.2.1.4 Tvar *bráno*

Tvar *bráno* je (na rozdíl od předchozích tří tvarů) mnohem častěji tvarem pasivního participia než substantiva, všechny výskyty jsou označkovány jako pasivní participia a až na dvě výjimky jimi opravdu jsou. Jediné dva tvary, které jsou ve skutečnosti vokativy podstatného jména, jsou ve spojeních *bráno nebes* a *smrti bráno*. V prvním případě by správnému otagování mohlo pomoci přidat do slovníku kolokací ustálené spojení *brána nebes* (bylo nalezeno 14 dalších případů, kde různé tvary lemmat *brána* a *nebesa* tvoří spojení), nebo by tagger mohl značku opravit na základě předpokladu, že genitiv substantiva v plurálu nemůže stát bezprostředně za pasivním participiem *bráno*. Druhé spojení se nachází v polovětné konstrukci *Ha, ty hrozná smrti bráno!*; zde je jednak tvar *bráno* také součástí (i když méně frekventovaného) frazému, jednak se v jeho okolí nevyskytuje žádný tvar slovesa *být* ani jiného slovesa, které by mohlo tvořit jeden analytický přísudek s pasivním participiem, nenajdeme zde ani žádný z výrazů, které v naprosté většině u pasivních participií od slovesa *brát* nacházíme (např. *v potaz*, *v úvahu*, *jako …*, *za …* atd.), mělo by tedy také být možné označkovat tvar jako vokativ substantiva.

4.2.1.5 Tvar *buzen*

Výraz *buzen* ve zkoumaném korpusu najdeme desetkrát, z čehož jen ve třech případech je označkován jako pasivní participium slovesa *budit*, sedmkrát má tag plurálu genitivu substantiva *buzna*. Ve skutečnosti jsou všechny tvary pasivními participii. Věty, v nichž automatický tagger slovesné tvary nerozpoznal, zpravidla nemají typickou strukturu pasivní věty, kdy podmětem je patiens a agens je vyjádřen instrumentálem nebo

genitivem s předložkou *od* (všechny tři správně určené tvary naopak toto schéma splňují – např. *...estetický zážitek je buzen už výstavbou básně*...). Agens vyjádřený substantivem v instrumentálu se však nachází ve větě *I já bych byl nerad buzen dlouho před svítáním hromovým hlasem z amplionu k modlitbě*... Tvar *buzen* je zde určen jako substantivum, ačkoliv ve větě není žádné slovo, které by si žádalo genitivní doplnění, a zároveň je v ní substantivum v instrumentálu, jímž se typicky vyjadřuje agens v pasivní větě, i tvar slovesa *být*, který se v čísle a jmenném rodě shoduje s tvarem *buzen*, označíme-li tento tvar za pasivní participium. Důvodem, proč automatický tagger nerozpoznal tuto strukturu a neurčil tvar *buzen* správně, může být, že sloveso *byl* nestojí těsně před tvarem *buzen* a agens vyjádřený substantivem *hlasem* stojí dokonce až pět pozic za ním (ve všech třech případech, v nichž bylo slovo *buzen* rozpoznáno a určeno správně, stojí tvar slovesa *být* těsně před ním a instrumentál substantiva do dvou pozic za ním). Problém by se tak mohl vyřešit, pokud by tagger bral v potaz širší okolí tvaru *buzen* a rozpoznal pasivní konstrukci v rámci celé klauze.

V souvětí *Domluvil se dokonce s kohoutem, že bude ráno buzen o půl hodiny dříve než ostatní a do budíčku vykoná nějakou dobrovolnou práci*. sice ve druhé klauzi chybí instrumentální vyjádření činitele děje, který je vyjádřen již v klauzi první, ale tvar genitivního plurálu substantiva zde opět není ničím „žádán“, zároveň se ve větě nachází sloveso *být* ve tvaru umožňujícím utvoření pasivní slovesné konstrukce s výrazem *buzen*. Resp. taková situace by nastala, kdyby nebylo chybně určeno i slovo *ráno*, které není označkováno jako adverbium, ale jako substantivum v nominativu. Takto je slovo *ráno* chápáno jako podmět a tvar *buzen* proto nemůže být považován za participium, jelikož není ve shodě s „podmětem“ ve středním rodě; bylo by tedy třeba opravit nejprve značku u tvaru *ráno*.

Naproti tomu v souvětí *Řekl, že nemůže skákat radostí, když je tak časně ráno buzen kvůli tak bezvýznamným věcem.* je tvar *ráno* správně určen jako adverbium, a syntakticky tedy nic nebrání tomu, aby spojení *je buzen* mohlo být opisným pasivem. Zároveň se ve větě opět nenachází žádný tvar, který by vyžadoval doplnění genitivem, není tedy důvod, aby tvar *buzen* byl určen jako substantivum. Také by (v tomto i předchozím případě) kolokační komponent mohl brát v úvahu poměrně typické spojení slov *budit* a *ráno* (téměř ve sto případech se v korpusu *SYN2010* v blízkém okolí tvarů slovesa *budit* nachází výraz *ráno*).

4.2.1.6 Tvar *bit*

Tvarů *bit* je v korpusu nalezeno 355, z nich 14 je označkováno jako pasivní participia a 341 jako substantiva. V obou skupinách se nacházejí i tvary označkované chybně. Pomineme-li zápisy, v nichž sám tvar *bit* není uveden správně (např. *Hrozně ho začali bit…*) nebo je částí textu psaného v jiném jazyce než českém (*Charlie bit my finger – again!*), a automatický tagger ho tedy jen těžko může adekvátně zpracovat, jde o 5 výskytů nesprávně označkovaných jako pasivní příčestí a přibližně 30 jako podstatné jméno.

Jeden z tvarů označkovaný chybně jako participium se nachází ve větě *Naproti tomu bit TXE se nahodí teprve tehdy…*; zde se již jedno sloveso určitého tvaru, které se nemůže pojit s pasivním participiem, nachází, není tedy důvod, aby byl tvar *bit* značen jako pasivní participium. Podobná situace je i ve větě *…objednala své čivavě Bit Bit stejk v přepočtu za 4000 Kč.*, kde se nejenže nachází jednoznačné určité sloveso neshodující se s příčestím *bit*, ale určovaný tvar stojí uprostřed věty s velkým počátečním písmenem, bude tedy velmi pravděpodobně nějakým propriem (či jeho částí). Ve větě *Kromě nejvyšších bitů adresy je do komparátoru zaveden ještě bit nejnižší…* je jiné pasivní příčestí, které je jednoznačné a tvoří jeden pasivní slovesný tvar se slovesem *je*, navíc, označí-li se tvar *bit* také jako participium, chybí ve větě mluvnický podmět. Poslední dva případy, v nichž je tvar *bit* otagován jako pasivní participium, ač je substantivem, jsou ve spojení *byt není bit*. Zde jde bohužel též o chybný zápis, první slovo by mělo být *byte*, nikoliv *byt*, možným řešením by snad mohlo být přidat celé toto spojení do slovníku kolokací s označením tvaru *bit* jako substantiva; je totiž velmi nepravděpodobné, že by se ve stejném spojení mohlo slovo *bit* opravdu vyskytovat jako součást slovesného tvaru, už jen proto, že vzhledem ke svému významu v naprosté většině případů vyžaduje podmět životného rodu, jímž tvar *byt* (resp. *byte*) není.

Přibližně čtvrtina výskytů tvarů chybně označkovaných jako substantiva je v případech, kde se tvar *bit* nachází v klauzi společně se zájmenem vztažným (resp. tázacím) *kdo* nebo neurčitým *někdo*, např.: *Čas od času se naše diplomacie rozhodne bít se za někoho, kdo je bit.*, *Proč bych to měl býti zrovna já, kdo musí být bit?*, *A kdo je na tom bit už teď?*, *V naší zemi nepřipustíme, aby byl někdo bit či dokonce vražděn kvůli barvě pleti, náboženství nebo politickému vyznání…* V těchto případech, je-li v jedné klauzi takové zájmeno, tvar slovesa *být* ve 3. osobě singuláru maskulina (případně v neurčitém tvaru) a zároveň v ní není žádné jiné pasivní příčestí, které by mohlo tvořit s tvarem slovesa *být* opisné pasivum, bude velice pravděpodobně tvar *bit* pasivním participiem.

Zájmena *kdo, někdo* jsou ve větě podmětem, tvar *bit* tedy nebude také substantivem v nominativu, nemůže být ani částí přísudku jmenného se sponou, jelikož zájmena *kdo* a *někdo*, nejde-li o metaforu či jinou figuru, se vztahují ke jménům životným, jímž substantivum *bit* není.

Obdobné případy nastávají, je-li tvar *bit* v jedné klauzi s tvarem slovesa *být* (v takovém rodě a čísle, aby mohlo být ve shodě s participiem *bit*) a substantivem mužského životného rodu v nominativu singuláru: *Copak chrápání ve společné ložnici je přestupek, za který je člověk bit?*, *...pán se bojí, pán se ukrývá, pán je bit a surově pobízen klackem…*, *...člověk byl bit, nevěděl od koho ani čím.*, *Dubček je moskevskými bratry současně bit a líbán…*, *Nenašly se žádné důkazy, že by byl chlapec zneužíván nebo bit…* apod. I zde, nenachází-li se ve větě jiné určité sloveso než tvary slovesa *být* ani jiné možné tvary pasivního příčestí než tvar *bit* (případně jsou-li takové tvary součástí několikanásobného větného členu), by měl být tvar *bit* rozpoznán jako pasivní participium. Také ve větách, kde se nachází tvar slovesa *být* v první/druhé osobě singuláru, a tedy i vyjádřený či nevyjádřený podmět *já*/*ty* (a jde tedy – vzhledem k funkci mluvčího/adresáta – o podmět rodu životného), bude tvar *bit* velmi pravděpodobně pasivním příčestím: *Před odjezdem jsem byl vězněn a bit…*, *Vždy znovu jsem se probouzel jako přejetý válcem a za celý svůj život jsem nebyl tolik bit…*, *...nejenže jsem nebyl bit, ani plísněn jsem nebyl.*, *Tys chtěl být bit?*

V mnoha případech by se kromě výše zmíněného jistě dalo vzít v potaz též pravidlo o vyjadřování činitele děje v pasivních větách instrumentálem či předložkou *od* a genitivem a předpokládat, že ve větách, kde takový tvar najdeme (a budou splněny i další podmínky, např. přítomnost vhodného tvaru slovesa *být* a naopak nepřítomnost neodpovídajících slovesných tvarů), bude výraz *bit* pasivním participiem: *...dokud poškozený neupadl na zem, kde byl opakovaně bit obuškem…*, *...byl doma v Taganrogu denně bit otcovým řemenem…*, *Byl jsem týrán nejhoršími způsoby, bit elektrickými kabely či železnými tyčemi…*, *Když plakal, byl bit od matky.*, *...a skoro pokaždé je při tom bit od Novočtvrťáků.*

4.2.1.7 Tvary *pěny, pěna, pěnu, pěn*

Jedním z tvarů, u nichž by se dalo předpokládat, že je v korpusu nalezneme jako tvary substantivní i verbální, je tvar *pěny*. Jistě si dovedeme představit věty jako *Písně byly pěny s nadšením.* nebo *V kostele byly pěny chorály.*, v nichž je tvaru *pěny* užito jako pasivního participia. Přesto je všech 555 výskytů tohoto tvaru v korpusu *SYN2010*

označkováno jako substantiva lemmatu *pěna*, a ani po zadání pozitivního filtru [tag="V.[IF]P.*"], který odfiltruje všechny výskyty, v nichž se v okolních pěti pozicích před či za tvarem *pěny* nenachází sloveso v plurálu feminina nebo maskulina inanimata (tedy sloveso, které by mohlo tvořit jeden gramatický celek s pasivním participiem *pěny*) a prohlédnutí zbylých 34 výsledků zde nenajdeme jediný, který by skutečně byl pasivním participiem.

Stav tvaru *pěna* je obdobný, všech 520 nalezených tvarů je označkováno jako substantivních a po vyfiltrování a zkontrolování těch, v jejichž okolí se nachází sloveso v odpovídajícím tvaru, opravdu nenalezneme žádný, který by byl pasivním participiem.

Z 260 nalezených tvarů *pěnu* také není žádný značkován jako tvar slovesný. Užití tohoto tvaru jako pasivního participia je sice mnohem méně pravděpodobné než u tvarů předchozích, možné by přesto bylo (např. *Svou oblíbenou píseň, pěnu tím strašným zpěvákem, nemohl ani slyšet.*). Opět ani po podrobném prohlédnutí všech výskytů žádný tvar pasivního participia nenalezneme.

28 výskytů tvaru *pěn* je na tom obdobně, všechny jsou substantivy.

4.2.1.8 Tvar *pěno*

Jediný tvar, který je v korpusu *SYN2010* značen jako pasivní participium od slovesa *pět*, je tvar *pěno*. Tento tvar se v korpusu vyskytuje třikrát, z toho dva výskyty jsou (správně) označkovány jako substantivum a jeden (chybně) jako pasivní participium – v souvětí *To slovo se jménem Štěpánka Štěpánková nemá vůbec nic společného, ale než si Honza Štěpánku v posteli pořádně oťukal, tak jí nikdo jinak než Pěno v celé vesnici neřekl*. Chybné označkování v tomto případě by se dalo odstranit jednoduše za předpokladu, že slovní tvar nacházející se uprostřed věty s velkým počátečním písmenem bude vlastním jménem.

4.2.1.9 Tvar *opraven*

Výraz *opraven* je v korpusu *SYN2010* označkován 10krát jako substantivum a 109krát jako pasivní participium. Mezi tvary označkovanými jako substantiva se skutečně nacházejí jen genitivy lemmatu *opravna* (v naprosté většině jde o názvy společností – *Energetické opravny Prunéřov*, *Železniční*/*Krnovské opravny a strojírny* – nebo o vazbu s typicky genitivní předložkou *do*).

Ve tvarech označkovaných jako participia najdeme osm případů, v nichž je tvar otagován chybně. V žádném z nich není ve větě zároveň i tvar sloves *být* či *bývat*, který by

mohl společně s tvarem *opraven* tvořit pasivní slovesný tvar, konstrukci, v níž se tvar jako trpné příčestí v naprosté většině případů vyskytuje – různé tvary slovesa *být* se nacházejí ve všech případech, kde je výraz *opraven* skutečně participiem – (až na konstrukce *Rodinný dům, zakoupen v roce 1986, postupně opraven.* a *Plášť raketoplánu opraven, počítače na ISS fungují HOUSTON*...). Podíváme-li se na jednotlivé výskyty, v nichž je tvar otagován chybně, dala by se většinou značka opravit i díky konkrétnějším poznatkům – např. ve spojení *závodníci Opraven obuvi Pecha* musí být výraz *Opraven* díky velkému počátečnímu písmenu částí nějakého vlastního jména, ve větě *Naopak služby kadeřníků a opraven zůstanou zdaněny 19 procenty.* nemůže výraz *opraven* tvořit jeden slovesný tvar s plurálem slovesa *zůstat* a není zde ani žádné maskulinum, k němuž by se participium mohlo vztahovat. Ve větách *Vedle hypermarketů přibude celá řada dalších obchodů, kaváren, restaurací, opraven bot či čistíren.* a *Dole byla řada jakýchsi obchůdků, půjčoven a opraven sportovního nářadí a vybavení, dokonce i informační středisko.* je jednak substantivem *řada* žádáno genitivní doplnění, jednak by se zde k pasivnímu participiu *opraven* ani nenacházela možná shoda.

4.2.1.10 Tvar *minut*

Tvar *minut* má v korpusu *SYN2010* 19242 výskytů a všechny z nich jsou substantivy lemmatu *minuta*. Ačkoliv by se dala vymyslet spousta příkladů na užití slova *minut* jako pasivního příčestí (např. *První památník byl minut a zastavilo se až u druhého.*), žádný (ani chybně značkovaný) se v korpusu nenachází, po zadání negativního filtru, jenž odstraní všechny výskyty tvaru *minut*, před nimiž na třech pozicích stojí číslovka (jiná než číslovka násobná, číslovka řadová a číslovka neurčitá s adjektivním skloňováním, jež by mohly být užity před participiem) nebo číselný výraz s číslicemi, výrazy *několik, několika, pár, desítky* a *desítek*[6] ([tag="C[=adhkjlny\?].*"|word="[Nn]ěkolik"|word="[Nn]ěkolika"|word="[Pp]ár"|word="[Dd]esítky"|word="[Dd]esítek"]), zbývá 302 výskytů, z nichž po důkladnějším prohlédnutí opravdu ani jeden není tvarem slovesa.

4.2.1.11 Tvar *minuta*

Výraz *minuta* je na tom obdobně, má v korpusu sice „pouhých“ 540 výskytů, ale navzdory očekávání jsou také všechny značeny jako substantiva a ani po podrobnějším

[6] Bylo by samozřejmě teoreticky možné vymyslet i příklad, v němž by se některý z těchto výrazu nacházel před tvarem *minut* užitým jako pasivní participium (např. *Během několika dní byl opět minut.*), ale vzhledem k tomu, že tvar *minut* není jako příčestí užit ani v mnoha případech, kde by byl „očekávatelnější“, zdá se velmi nepravděpodobné, že bychom tímto filtrem vyloučili nějaké jeho užití jako slovesného tvaru.

průzkumu jednotlivých tvarů nebylo nalezeno žádné „chybně otagované“ participium (např. všechny tvary, před nimiž se v rozsahu jedné až pěti pozic – v rámci jedné věty – nachází nějaký tvar slovesa *být*, jsou skutečně substantivy, stejně tak jsou substantivy všechny výrazy následované v rámci téže klauze do pěti pozic tvarem slovesa *být*).[7]

4.2.1.12 Tvar *říjen*

Tvar *říjen* by slovotvorně mohl být pasivním participiem od slovesa *říjet*. Významově ani syntakticky to však možné není, podle mluvnic se příčestí trpná (až na výjimky) tvoří od sloves předmětových, jímž sloveso *říjet* není. Přesto v korpusu *SYN2010* najdeme dva výskyty tvaru *říjen* označkovaného jako pasivní příčestí od slovesa *říjet* – jednou je tvar značen jako mužský rod životný (*23. říjen 2001*), jednou dokonce jako mužský rod neživotný (*Nehody na Schipholu Říjen 1992*). Vzhledem k tomu, že tvar prakticky pasivním participiem být nemůže, mohly by tyto chyby být odstraněny vyloučením možnosti označkovat výraz *říjen* jako sloveso a ponecháním pouze tagů pro nominativ a akuzativ singuláru substantiva *říjen*.

4.2.2 Numeralia

4.2.2.1 Tvary *jeden, nejeden, set*

Pasivní participia mohou být homonymní i s několika číslovkami – konkrétně se jedná o tvary *jeden, nejeden, set*. Všechny výskyty všech těchto tvarů jsou v korpusu určeny jako číslovky a ani po důkladném hledání se nezdařilo nalézt jediný výskyt některého z nich ve významu slovesném. Určitě si jejich užití jako pasivních příčestí představit dovedeme, ale v praxi se evidentně příliš nevyskytují (z 59662 nalezených slovních tvarů *jeden* a *nejeden* jsou – s největší pravděpodobností – všechny číslovky, ze 7630 tvarů *set* všechny číslovky nebo substantiva).

4.2.3 Adverbia

4.2.3.1 Tvar *zataženo*

Všech 176 výskytů tvaru *zataženo* je v korpusu *SYN2010* označkováno jako adverbia. Většina jimi skutečně je, najdou se však i výjimky, které by měly být označkovány jako participia.

[7] I tvary *minuty* a *minutu* se v korpusu vyskytují jen jako substantiva, tvar *minuto* v korpusu *SYN2010* nenalezneme vůbec.

Tvary, které jsou ve skutečnosti pasivními participii, vždy „vyžadují“ podmět, zároveň se k nim často váže i předmět, typicky vyjádřený předložkou *do* a genitivem (tímto bychom mohli opravit tagy ve větách *...aby zůstal Erland naživu a Království nebylo zataženo do války.*, *Přitom by bylo Československo zataženo do války po boku Sovětského svazu.*, *Je do ní zataženo příliš mnoho lidí…*, *...aby mohlo být Království zataženo do války s Říší…*, *...když je do nich dítě zataženo.*) nebo instrumentálem bez předložky (ve větě *Když zrovna nepršelo, nebe bylo zataženo temnými mraky.*). I ve větách, v nichž činitel děje vyjádřen není, ale mluvnický podmět v odpovídajícím tvaru singuláru neutra ano, je tvar *zataženo* spíš participiem: *Přijel totiž až v devět večer a v té době bylo nebe nad Vídní opět zataženo…*, *Odjížděl rychlíkem Norimberk, Frankfurt a Forbach a nebe nad městem bylo zataženo stejně, jako když sem přijel.*

Ve větách, kde je tvar *zataženo* adverbiem, naopak podmět nebývá, věty jsou jednočlenné: *Je zataženo a ve vzduchu je cítit déšť.*, *Bylo nezvykle vlhko a stále zataženo.*, *Na sever od této hranice bylo zataženo…*, *V newyorské oblasti bylo zataženo už celých devět dní z posledních čtyřiceti.* apod.

4.2.4 Konjunkce

4.2.4.1 Tvar *děleno*

Výskyty tvaru *děleno* jsou v korpusu *SYN2010* označkovány buď jako pasivní participia, nebo (i když pouze ve třech případech z téměř sedmdesáti) jako konjunkce. Považovat tvar *děleno* za spojku není příliš ustálené, např. ve *Slovníku spisovného jazyka českého* heslo *děleno* vůbec nenalezneme, ostatní tři základní matematické operace (*plus*, *minus* a *krát*) jsou uvedeny jako příslovce. Budeme-li zde vycházet ze značkování korpusu a výraz *děleno* v určitých případech (zejména při vyjadřování matematických operací) opravdu chápat jako spojku, bude třeba upřesnit disambiguaci tvarů a spoustu tagů opravit. Všechny tři případy určení tvaru jako spojky jsou ve větách, v nichž se nachází sloveso neshodující se s tvarem *děleno* v čísle či osobě, tvar tedy nemůže být součástí pasivního slovesného tvaru: *Přípustná chyba by pak byla 98% děleno druhou odmocninou z 5825043 neboli 0,04%.*, *...která by byla 98% děleno druhou odmocninou ze…*, *Míra vražd ve Francii tedy byla v roce 2000 rovna 1051 vražd děleno 59225683 obyvatel krát…* Většina tvarů označkovaných jako pasivní participia však také bude spíš spojkami, nalezneme mezi nimi i případy, v nichž se tvar *děleno* nachází ve velmi podobném kontextu jako tvary označené za spojky, přesto jako spojka otagován není (*Třicet děleno*

deseti jsou 3., *Vypočítáme-li 98% děleno odmocninou z 24000, dostaneme…*). Přiřadit značku spojky tvarům v těchto a podobných větách by vzhledem k přítomnosti slovesných tvarů (které nemohou tvořit část pasivní konstrukce) nemělo být problematické, bylo by vhodné brát v potaz i klasické matematické výrazy, v jejichž kontextu bude tvar *děleno* spíš spojkou než pasivním participiem – jde zejména o výrazy *rovná se* (resp. *se rovná*) (např. *…vzdálenost 7,57 m vydělená 24 se rovná 31,5 cm, děleno 22 se rovná 34,4 cm.*, *Nemůžeme říct „a krát x děleno y se rovná zx na ay“…*, *Celý bratr plus dvě poloviční sestry děleno dvěma matkami se rovná jednomu otci plus jednomu víkendovému tatínkovi…*) a *je rovno* (*A skutečně, 98% děleno odmocninou z 1397 je rovno 2,622%…*, *Skutečně je také 98% děleno odmocninou z 5254 rovno 1,352%…*). V úvahu by se mohly brát i další výrazy, jejichž výskyt v klauzi společně s tvarem *děleno* naznačuje, že tvar *děleno* bude velmi pravděpodobně spíš spojkou než příčestím – tvary označující další matematické operace (*plus*, *minus*, *krát*, tvary substantiv *odmocnina*, *procento*,…), číslovky (a zápisy psané číslicemi) nebo znak *%*: *„Plus šest procent z hrubýho zisku děleno čtyřma?“*, *…řekněme 72 let, minus váš aktuální věk děleno délkou vašeho důchodu…*, *Pět, sedm, dvě osmdesát, krát osm, děleno dvěma.*, *…se relativní četnost panen nebude lišit od správné hodnoty 50% o více než 98% děleno odmocninou z počtu mincí.* apod.

Největší problém s automatickým rozlišením spojky od participia je v případech, kdy jediným dalším slovesem ve větě je tvar slovesa *být*, který by mohl být ve shodě s příčestím *děleno*. Takové tvary se bohužel často nacházejí jak v konstrukcích s pasivním participiem *děleno* (např. *To je děleno na oblasti Pošta, Kalendář, Kontakty a Úkoly*, *…vzdělávání není dále děleno ani přerušováno.*, *…je bez zapojení vnějších obvodů děleno vždy v poměru odporů na dvě napětí…* apod.), tak ve vyjádření výsledku podílu, kde je tvar *děleno* spojkou (*Protože 196% děleno dvěma je 98%…*, *Pět minus jedna je čtyři, pět děleno jednou je pět.*, *Potom můžeme říci, že relativní podíl žen z této skupiny, které byly ve 40 letech neprovdané, ale v 75 letech provdané, je asi 0,8% děleno 4,9%, tj. 16,3%.*). V těchto případech je možné brát v potaz jednak výše zmíněné výrazy pojící se často s tvarem *děleno* ve funkci spojky, jednak například typické spojení pasivního participia *děleno* s předložkou *na* a akuzativním tvarem: *…patro je děleno na velkou západní místnost a malou místnost jihovýchodní.*, *…triforium bylo děleno na šest dílů stejně jako okna…*, *Bednění je děleno na jednotlivé díly 4,5 m dlouhé.*).

Dalším kritériem pro odlišení užití tvarů může být „binarita“ spojky – tvar *děleno* užitý jako spojka nutně vyžaduje doplnění dvěma členy, jedním zleva a jedním zprava;

oba mohou být vyjádřeny jak číslovkami (nebo číslicemi), tak podstatnými jmény (případně i zájmeny), není-li však druhý člen vyjádřen číslicí, musí být vždy v instrumentálu (a tedy tvar *děleno*, za nímž se v rámci klauze nenachází žádný tvar v instrumentálu, nemůže být spojkou: *...protože podle kalifornských zákonů je bezpodílové spoluvlastnictví včetně příjmů děleno napůl.*, *„Srbsko nechce dělit Kosovo. Stejně tak ale nechceme, aby bylo děleno Srbsko," uvedl…*, *...je sdíleno, aniž ho ubývá, aniž je děleno.*). Samozřejmě ale neplatí, že by každý tvar následovaný instrumentálem substantiva nutně musel být spojkou, může jít i o klasické vyjádření činitele děje v pasivní větě: *...bylo ve své nejstarší části dosud děleno pozůstatky původních hradeb a vodním příkopem…*, *...průčelí věží a budov bran mohlo být děleno římsami.*

4.2.5 Interjekce

4.2.5.1 Tvar *nevídáno*

Tvar *nevídáno* ve zkoumaném korpusu najdeme 15krát, ve všech případech je označkován jako pasivní participium. Ve *Slovníku spisovného jazyka českého* je samostatné heslo *nevídáno* uvedeno jako citoslovce, které vyjadřuje „přikládání malého n. nepřikládání žádného významu něčemu; to je toho, na tom nezáleží" (SSJČ[8]), jako příklady užití tvaru jsou zde uvedeny konstrukce *on se hněvá? Nevídáno!*, *nevídáno, však to nějak dopadne*, *nevídáno pro korunu* a *nevídáno nějaké škrábnutí*.

Vycházeli-li bychom z této charakteristiky výrazu *nevídáno* jako citoslovce a přihlíželi především k funkci tvaru ve větě, většina tvarů v korpusu by měla být také značkována spíš jako citoslovce. Určitě by tomu tak (v paralele s příkladem uvedeným v SSJČ) mělo být u výskytu tvarů ve spojení s předložkou *pro* a tvarem akuzativu: *Nevídáno pro pár snopů!*, *Jenže kupec Milhost se panoše Oty zastal, že pro pár žertů nevídáno a…*, ale i v jiných případech, nikde se totiž např. nenachází tvar slovesa *být* v příslušném tvaru k utvoření pasivní konstrukce s výrazem *nevídáno*. Jako citoslovce by tvar mohl být značen kupříkladu i ve větách, v nichž je oddělen jako samostatný člen (*Hm, přeťatá žíla nad okem, nevídáno…*, *...jestli se potkám s trním, nevídáno, vždyť mě sotva málem trefili.*, *Cože? Ve snu? Nevídáno, fantazie má zase šlágr…*), obzvlášť je-li navíc zdůrazněn vykřičníkem: *Deset měďáků, nevídáno!*, *Pak mezi kozy chtěl mít zamrdáno – nevídáno!*

[8] Dostupné z WWW: http://ssjc.ujc.cas.cz/, heslo *nevídáno*. Citováno 5. 5. 2015.

5 Závěr

Cílem práce bylo představit na tvarech pasivních participií problematiku automatické disambiguace homonymních tvarů ve značkování korpusů a přednést konkrétní návrhy pro značkování přesnější. V první části bylo stručně popsáno, jak je v českých jazykovědných publikacích tvar pasivního participia definován a jaká jsou omezení jeho tvoření a užívání. Další kapitola byla věnována popisu automatického morfologického značkování Českého národního korpusu se zaměřením na popis automatické morfologické disambiguace.

V praktické části bylo vymezeno šest skupin slovnědruhové homonymie pasivních participií podle toho, k jakému slovnímu druhu patří tvar homonymní s pasivním participiem. Z jednotlivých skupin (kromě adjektiv, jimiž se tato práce nezabývala) byly vybrány konkrétní tvary – nejvíce substantiv, s nimiž jsou pasivní participia homonymní nejčastěji (*brány*, *brána*, *bránu*, *bráno*, *buzen*, *bit*, *pěny*, *pěna*, *pěnu*, *pěn*, *pěno*, *opraven*, *minut*, *minuta*, *říjen*), dále tři tvary numeralií (*jeden*, *nejeden*, *set*) a po jednom tvaru adverbia (*zataženo*), konjunkce (*děleno*) a interjekce (*nevídáno*) –, u nichž byl popsán současný stav značkování v korpusu *SYN2010* a v případech, kde se nacházely i značky chybné, byly navrženy konkrétní návrhy na úpravu.

Na základě zkoumaných tvarů a jejich značkování bylo možné vysledovat několik obecných pravidel, která platí v naprosté většině případů a jsou schopna odstranit chybné tagy u mnoha tvarů (pravidla jsou v této práci formulována jen lingvisticky, předpokládáme, že je možné je počítačově implementovat v systému *LanGr*, který byl popsán v kapitole věnované anotování korpusů). Jedno z takových pravidel, uplatňované u tvarů pasivních participií homonymních se substantivy, je založeno na předpokladu, že v typické české větě bývá právě jedno jméno v nominativu (které je podmětem věty), a tedy tvar, který se nachází v klauzi společně s nějakým takovým jednoznačným jménem, a mohl by být jak pasivním participiem, tak substantivem v nominativu, bude pravděpodobně participiem (za předpokladu, že se ve větě nenachází jiný jev, který by to znemožňoval, např. tvar slovesa s participiem se neshodující). V mnoha jiných případech se tvar chybně označkovaný jako participium nacházel v klauzi společně s určitým tvarem slovesa, který s tvarem pasivního příčestí nebyl ve shodě. V těchto případech spolu oba výrazy nemohou tvořit jeden gramatický celek a homonymní výraz tak obvykle není pasivním participiem, ale tvarem příslušejícím k lemmatu druhého slovního druhu. Jiné pravidlo – uplatňující se také u homonymie se substantivy – může upravit chybné tagy

u tvarů, které jsou značkovány jako pasivní participia, ačkoliv se nacházejí uprostřed věty a začínají velkým písmenem. Takové tvary by podle českého pravopisu měly být nějakým propriem (nebo jeho částí), nemohou být pasivním participiem. Další pravidla vychází z toho, jakými prostředky se v pasivní větě typicky vyjadřuje činitel děje – je-li tak např. v blízkém okolí možného pasivního příčestí jméno v instrumentálu, je pravděpodobné, že jde o vyjádření příslušného činitele děje a tvar opravdu bude participiem (toto však rozhodně není pravidlo platné vždy, např. pro tvar *děleno* – a spoustu dalších – bychom ho užít nemohli).

Většina pravidel je bohužel takového rázu, že nejsou obecně platná ve všech případech, a je tedy nutné pro každý tvar formulovat spoustu pravidel zohledňujících konkrétní vazby a valenční „požadavky" jednotlivých tvarů, častá spojení, v nichž se nacházejí (ať jako tvary pasivního participia nebo jiného slovního druhu), a naopak konstrukce, ve kterých figurovat nemohou. To vyžaduje samozřejmě spoustu důkladné a časově náročné práce, doufáme proto, že jednotlivé návrhy úprav ve značkování předložené v této práci budou alespoň drobným přínosem a usnadněním práce při anotování korpusů příštích.

6 Použitá literatura

Český národní korpus – SYN2010. Praha: Ústav Českého národního korpusu FF UK 2010. Dostupné z WWW: http://www.korpus.cz. Citováno 1. 1. 2015 – 13. 5. 2015.

HAJIČ, Jan: *Popis morfologických značek – poziční systém*. Praha: Ústav formální a aplikované lingvistiky MFF UK 2000. Dostupné z WWW: https://ucnk.ff.cuni.cz/doc/popis_znacek.pdf.

HAJIČ, Jan – KRBEC, Pavel – KVĚTOŇ, Pavel – SPOUSTOVÁ, Drahomíra – VOTRUBEC, Jan: The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In Piskorski, Jakub – Tanev, Hristo: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing.* Prague: Association for Computational Linguistics 2007, s. 67–74.

HAVRÁNEK, Bohuslav a kol. (ed.): *Slovník spisovného jazyka českého* [online]. 1960–1971. Dostupné z WWW: http://ssjc.ujc.cas.cz/. Citováno 5. 5. 2015.

HAVRÁNEK, Bohuslav – JEDLIČKA, Alois: *Česká mluvnice.* Praha: Státní pedagogické nakladatelství 1981 (4., přepracované vydání).

HNÁTKOVÁ, Milena – PETKEVIČ, Vladimír: Automatická morfologická disambiguace předložkových skupin v českém národním korpusu. In Hladká, Z. – Karlík, P. (eds.): *Čeština – univerzália a specifika, 4.* Praha: Nakladatelství Lidové noviny 2002, s. 243–252.

JELÍNEK, Tomáš – PETKEVIČ, Vladimír: Systém jazykového značkování korpusů současné psané češtiny. In Petkevič, V. – Rosen, A. (eds.): *Korpusová lingvistika Praha 2011. 3 – Gramatika a značkování korpusů*. Praha: Nakladatelství Lidové noviny 2011, s. 154–170.

KARLÍK, Petr – NEKULA, Marek – PLESKALOVÁ, Jana (eds.): *Encyklopedický slovník češtiny.* Praha: Nakladatelství Lidové noviny 2002.

KARLÍK, Petr – NEKULA, Marek – RUSÍNOVÁ, Zdenka (eds.): *Příruční mluvnice češtiny.* Praha: Nakladatelství Lidové noviny 1995.

KOMÁREK, Miroslav – KOŘENSKÝ, Jan – PETR, Jan: *Mluvnice češtiny 2. Tvarosloví.* Praha: Academia 1986.

KVĚTOŇ, Pavel: *Rule-Based Morphological Disambiguation.* Disertační práce. Praha: MFF UK 2006.

LOPATKOVÁ, Markéta a kol.: *VALLEX 2.6.3 – Valency Lexicon of Czech Verbs*. Praha: Ústav formální a aplikované lingvistiky MFF UK. Dostupné z WWW: http://ufal.mff.cuni.cz/vallex/2.6.3/doc/home.html.

OLIVA, Karel: Úvahy nad teoretickými základy lingvisticky adekvátní disambiguace jazykových korpusů. In Blatná, R. – Petkevič, V. (eds.): *Jazyky a jazykověda. Sborník k 65. narozeninám prof. PhDr. Františka Čermáka, DrSc.* Praha: Filozofická fakulta Univerzity Karlovy – Ústav Českého národního korpusu 2005, s. 229–245.

PETKEVIČ, Vladimír: Využití pravidel pro negaci v automatickém značkování českých korpusů. In Hladká, Z. – Karlík, P. (eds.): *Čeština – univerzália a specifika, 5.* Praha: Nakladatelství Lidové noviny 2004, s. 143–150.

PETKEVIČ, Vladimír: Za češtinu (ne)homonymní, aneb jak odstranit slovnědruhovou a morfologickou homonymii v českých korpusech. In Blatná, R. – Petkevič, V. (eds.): *Jazyky a jazykověda. Sborník k 65. narozeninám prof. PhDr. Františka Čermáka, DrSc.* Praha: Filozofická fakulta Univerzity Karlovy – Ústav Českého národního korpusu 2005, s. 247–266.

SKOUMALOVÁ, Hana: Porovnání úspěšnosti tagování korpusu. In Petkevič, V. – Rosen, A. (eds.): *Korpusová lingvistika Praha 2011. 3 – Gramatika a značkování korpusů*. Praha: Nakladatelství Lidové noviny 2011, s. 199–207.

SPOUSTOVÁ, Drahomíra: *Kombinované statisticko-pravidlové metody značkování češtiny.* Disertační práce. Praha: MFF UK 2007.

TRÁVNÍČEK, František: *Mluvnice spisovné češtiny II.* Praha: Slovanské nakladatelství 1951.