UNIVERZITA PALACKÉHO V OLOMOUCI

FILOZOFICKÁ FAKULTA

Katedra bohemistiky

Morfologické značkování korpusů češtiny – komparace

(Morphological Tagging of Czech Corpuses – a Contrastive Study)

Bakalářská diplomová práce

Adéla Hanová

Česká filologie se zaměřením na editorskou práci ve sdělovacích prostředcích

Vedoucí práce: PhDr. Petr Pořízka, Ph.D.

Olomouc 2014

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod odborným dohledem vedoucího práce a uvedla jsem všechny použité zdroje a literaturu.

V Olomouci dne: …… Podpis: ……

Tímto děkuji PhDr. Petru Pořízkovi, Ph.D., za odborné vedení při vypracovávání bakalářské práce.

Obsah

Úvod 5

1 Korpus a jeho anotace 7

1.1 Korpus 7

1.2 Anotace korpusu 8

1.2.1 Úrovně lingvistické anotace 9

1.2.2 Morfologické značkování korpusu 9

1.3 Problémy při anotaci 12

1.3.1 Chyby v textu a neznámá slova 12

1.3.2 Homonymie 12

1.3.3 Víceslovné výrazy 13

1.4 Anotace a lingvistická teorie 14

2 Systémy morfologického značkování češtiny 16

2.1 Poziční morfologický tagset 16

2.2 Atributivní morfologický tagset 18

2.3 Kompaktní morfologický tagset 19

2.4 Xerox tagset 20

2.5 Kódovník PMK 22

2.6 Příklady tagů – vizuální srovnání 23

3 Odraz morfologie češtiny v morfologických tagsetech 27

3.1 Mluvnice češtiny 2 27

3.2 Komparace tagsetů a teorie 28

3.2.1 Substantiva 29

3.2.2 Adjektiva 33

3.2.3 Zájmena 37

3.2.4 Číslovky 43

3.2.5 Slovesa 46

3.2.6 Příslovce 51
3.2.7 Předložky 52
3.2.8 Spojky 53
3.2.9 Částice 54
3.2.10 Citoslovce 55
3.2.11 Ostatní kategorie na úrovni slovního druhu 56
3.3 Výsledky komparace morfologických tagsetů s teorií prezentovanou v Mluvnici češtiny 2 59
3.3.1 Poziční tagset 59
3.3.2 Atributivní tagset 61
3.3.3 Kompaktní tagset 61
3.3.4 Xerox tagset 62
3.3.5 Kódovník PMK 62
3.3.6 Otázka možné vzájemné kompatibility 63
4 Závěr 64
Anotace 66
Shrnutí 67
Resumé 67
Seznam použité literatury 68
Seznam tabulek 71

Úvod

V této bakalářské práci se budeme zabývat pěti systémy morfologického značkování češtiny (tzv. tagsety), a to systémem pozičním, atributivním, kompaktním, Xerox tagsetem a v neposlední řadě také kódovníkem PMK. Morfologické tagsety jsou konstruovány na základě odlišných (lingvistických) koncepcí a morfologické kategorie odrážejí v různé míře a s různou komplexností. Je důležité uvědomit si do jaké míry (a zda) jsou jednotlivé koncepce v souladu s lingvistickou teorií. Touto problematikou se před námi nikdo komplexně nezabýval.

Naším úkolem je zjistit, jak a do jaké míry koncepce morfologických tagsetů odrážejí lingvistickou teorii popsanou ve výkladových příručkách české gramatiky. Teorie vyložená v příručkách se může mezi jednotlivými publikacemi značně lišit. Pro komparaci jsme zvolili jako teoretickou bázi tzv. *akademickou Mluvnici češtiny*. Na základě porovnání jednotlivých tagsetů s lingvistickou teorií vyloženou v publikaci bychom měli být schopni postihnout základní rysy tagsetů z hlediska obsažených kategorií a na závěr také zvážit možnosti jejich vzájemné kompatibility.

První kapitola přináší stručný obecný výklad o anotaci korpusů. Systematicky se dostáváme od obecných informací o korpusech a jejich druzích přes nejrůznější možnosti anotace až k anotaci lingvistické a nakonec také ke složení tagsetů. Zejména zde pojednáváme o lingvistické anotaci a o problematických momentech, které se s ní pojí. Problémy spjaté s lingvistickou anotací zmiňujeme zejména proto, že mají podstatný vliv na podobu morfologických tagsetů.

V následující, druhé, kapitole představujeme pět systémů morfologického značkování dostupných pro češtinu. U každého tagsetu vždy uvádíme tvůrce, s jakými institucemi se pojí, dále pak v jakých projektech a korpusech daný tagset figuruje. Nechybí základní popis tagů jednotlivých systémů, jejich struktury a na závěr také jejich vizuální srovnání. V této části práce, také uvádíme zdroje informací o jednotlivých tagsetech, z nichž jsme čerpali materiál pro další studium zmíněných systémů.

Poslední kapitola přináší poznatky získané z analýzy kategorií určovaných v jednotlivých tagsetech a jejich komparace s lingvistickou teorií prezentovanou

v *Mluvnici češtiny*. V závěru kapitoly zmiňujeme nejdůležitější rysy tagsetů vypozorované na základě našeho studia dané problematiky. Vyjadřujeme se také ve věci vzájemné kompatibility systémů.

1 Korpus a jeho anotace

Korpusová lingvistika je poměrně mladá disciplína, která se však velmi rychle vyvíjí. Ústav českého národního korpusu (ÚČNK) byl založen v roce 1994 a české korpusy patří k těm největším a nejlépe zpracovaným v Evropě. Dat, která dnes mají lingvisté prostřednictvím korpusů k dispozici, je nesrovnatelně více, než tomu bylo v době manuální excerpce textů. Počítačové zpracování dat navíc umožňuje jejich systematické studium. Zde tkví hlavní přínos korpusové lingvistiky pro poznání jazyka, které je její zásluhou hlubší a přesvědčivější.[1]

1.1 Korpus

V lingvistice je korpus definován jako „soubor dokladů autentického užití přirozeného jazyka. Je to tedy materiálová základna, která slouží k lingvistické analýze a popisu.“[2] V posledních letech se však pojem korpus používá v lingvistických kruzích téměř výhradně pro označení elektronického souboru textů.[3] Tak jej také definuje např. František Čermák, který korpusem rozumí „ucelený soubor textů, který je sestavený s přihlédnutím ke svému cíli reprezentativním způsobem, je obhospodařovaný počítačově a zpracováván souborem korpusových metod.“[4] Zpracovává reálné texty, a to jak mluvené, tak psané. Korpusy psané dnes stále převládají, zpracování mluvených textů je velmi pracné a časově náročné. Reprezentativnost je důležitou, pro lingvistické zkoumání zcela zásadní, vlastností korpusu. Tuto podmínku dnes již mnohé korpusy splňují, jsou tedy koncipovány tak, aby množstvím a druhem zpracovávaných textů odrážely sledovaný žánr (jedná-li se o korpusy specializované) nebo celé univerzum jazyka.[5]

Korpus je pro lingvistiku zdrojem informací, díky počítačovému zpracování je práce s daty v mnoha ohledech usnadněna. Šulc upozorňuje na fakt, že korpus neslouží jako zdroj informací pouze lingvistům, „ale i širokému spektru zájemců z jiných

[1] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 9.
[2] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 9.
[3] Tamtéž, s. 9.
[4] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 10.
[5] Tamtéž, s. 10.

oborů k mnohostrannému poznání jazyka, zákonitostí lidského myšlení a kultury.“[6] Studium korpusů se děje prostřednictvím korpusového manažeru (v českém prostředí např. Bonito[7] anebo novější NoSketch Engine[8]), který po zadání dotazu nabídne konkordance, úplný soupis výskytů hledaného jevu v jeho bezprostředním (levopravém) kontextu.[9] V případě, že je korpus řádně anotován (viz níže), je možné vyhledávací dotazy upravovat a upřesňovat, ale především vyhledávat efektivněji – prostřednictvím lemmat a tagů: při studiu specifických jevů jazyka je to nespornou výhodou.

1.2 Anotace korpusu

Z hlediska značkování korpusů rozlišujeme korpusy anotované a neanotované. Anotované korpusy jsou velmi užitečné, neboť obsahují doprovodné informace, které uživatelům korpusů výrazně usnadňují a urychlují práci.[10] Anotace a její druh (druhy) jsou údaje spadající do základní charakteristiky daného korpusu, na jejímž základě badatel posuzuje, zda je určitý korpus k jeho výzkumu vhodný, či nikoliv. Zjišťuje tak, jaké informace byly do korpusu přidány. Geoffrey Leech definuje značkování korpusu (anotaci) obecně jako: „přidávání *interpretativní* (zvláště lingvistické) informace k existujícímu korpusu mluveného a/nebo psaného jazyka pomocí nějakého typu kódování.“[11] Rozlišujeme anotaci vnější (přinášející informace o zdrojích textů) a vnitřní (zabývající se strukturou či přímo povahou jazykových jednotek).[12]

[6] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 11.
[7] RYCHLÝ, Pavel. Bonito. In: CZPJ. *Centrum zpracování přirozeného jazyka* [online]. [cit. 2014-04--24]. Dostupné z: http://nlp.fi.muni.cz/projects/bonito/bonito.html.cz.
[8] RYCHLÝ, Pavel. NoSketch Engine. In: CZPJ. *Centrum zpracování přirozeného jazyka* [online]. Poslední aktualizace 3. 3. 2013 [cit. 2014-04-24]. Dostupné z: http://nlp.fi.muni.cz/projects/bonito/bonito.html.cz.
[9] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 11.
[10] JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 154.
[11] LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 185.
[12] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 10.

1.2.1 Úrovně lingvistické anotace

Lingvistickou anotaci korpusu lze provádět na několika rovinách. Leech uvádí tyto úrovně značkování: ortografická, fonetická/fonologická, prozodická, slovnědruhová třída (tj. gramatické značkování), syntaktická (tj. syntaktická analýza), sémantická, pragmatická/diskursová. Jako nejrozšířenější druh lingvistické anotace označuje anotaci gramatickou.[13] Ačkoli popisuje stav typický pro korpusy angličtiny, lze toto tvrzení vztáhnout i na situaci v českém prostředí. Korpusy ČNK obsahují totiž zejména značkování morfologické, jelikož „se dosud [r. 2011] v projektu *Český národní korpus* (ČNK) nepřikročilo ke značkování syntaktickému[14] (tedy k tvorbě syntakticky anotovaných struktur), ani sémantickému (tedy ani např. ke značkování lexikálních významů paradigmaticky homonymních slov, k tzv. disambiguaci lexikálních významů).“[15] Podle Leecha se gramatické značkování stalo nejvyužívanějším způsobem anotace ze dvou důvodů: „(a) je dost jednoduché, aby se dalo dělat do značné míry automaticky; (b) má zřejmé využití: např. v lexikografii, kde představuje první krok k automatické lemmatizaci.“[16]

1.2.2 Morfologické značkování korpusu

Morfologické značkování korpusu je komplexní proces, ještě než dojde k samotné morfologické analýze jednotlivých tvarů, je třeba projít několika fázemi. „Nejprve je text rozdělen na slova a grafické znaky […] a dále na věty pomocí tokenizeru a segmenteru. Poté tzv. ‚morfologická analýza‘ přiřadí každému slovnímu tvaru všechny odpovídající lemmata a značky.“[17] Zde však celý proces nekončí, ještě je třeba provést disambiguaci, při které je pro každý tvar vybrána z přiřazených

[13] LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 186.
[14] Dnes již existují syntakticky anotované korpusy, např. Pražské závislostní korpusy vznikající v ÚFAL MFF UK (ÚFAL Corpora. *Institute of Formal and Applied Linguistics* [online]. © 2014 [cit. 2014-04-24]. Dostupné z: http://ufal.mff.cuni.cz/projects/corpora.). Syntaktické značkování však není předmětem této práce.
[15] JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 155.
[16] LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 188.
[17] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 171.

možností jedna kombinace tagu a lemmatu, z hlediska kontextu ta korektní.[18] Morfologické značkování v širším smyslu zahrnuje všechny výše zmíněné postupy, tedy: tokenizaci, větnou segmentaci, morfologickou analýzu i morfologickou disambiguaci. Hovoříme-li o morfologickém značkování v užším slova smyslu, myslíme tím poslední dva zmíněné procesy, morfologickou analýzu a morfologickou disambiguaci.[19]

1.2.2.1 Morfologická analýza

Jak již bylo řečeno výše, proces morfologické analýzy dat v korpusu spočívá v přiřazování morfologických značek (tagů) a lemmat k jednotlivým slovním tvarům (tokenům). To se děje prostřednictvím programu pro jejich přiřazování, guesseru.[20] Uveďme zde příklad[21] analýzy tvarů, které mají více morfologických interpretací:

tvar: *sluje*
lemma = *sluj*, tag = NNFS2-----A-----
lemma = *sluj*, tag = NNFP1-----A-----
lemma = *sluj*, tag = NNFP4-----A-----
lemma = *sluj*, tag = NNFP5-----A-----
lemma = *slout*, tag = VB-S---3P-AA---1

Pracovat s korpusem, ve kterém může mít jeden slovní tvar více interpretací, by bylo značně neefektivní, proto je nutné přistoupit k zjednoznačnění interpretace, provést disambiguaci.[22]

1.2.2.2 Morfologická disambiguace

Disambiguaci je možné provádět pomocí různých metod: statistických, pravidlových, případně kombinací obou typů. Statistická (stochastická) metoda je založená na tzv. strojovém učení, disambiguační program (tagger) se nejdříve učí vybírat správné

[18] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 171.
[19] JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 155.
[20] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 172.
[21] JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 158.
[22] Tamtéž, s. 159.

kombinace na ručně anotovaném vzorku textu, potom aplikuje získané poznatky a vybere pro dané věty nejpravděpodobnější kombinaci značek.[23] Pravidlová metoda je oproti tomu založená na systému lingvistických pravidel, která jsou speciálním programem opakovaně aplikována na danou větu, dokud nejsou odstraněna všechna negramatická čtení.[24] Protože zkoumá negramatické struktury a kombinace, říká se tomuto typu disambiguace redukční. Jelínek a Petkevič uvádějí u redukční disambiguace následující příklad (při aplikaci negativního bigramu: konfigurace *předložka – sloveso*):

„[když] je **předložka** slovnědruhově jednoznačná (není homonymní s jiným slovním druhem), například *do*, a sloveso je slovnědruhově nejednoznačné, kupříkladu *sluje* [tvarovou homonymii viz výše]. Je pak jasné, že tvar *sluje* je tvar neslovesný, tedy – v tomto případě – substantivní, a příslušné pravidlo odstraní u tvaru *sluje* slovesný tag i lemma *slout*, srov. větu:

(4) *To léto jsem se přestěhoval do sluje* (Subst,~~Verb~~) *nad křižovatkou.*“[25]

Pro automatickou disambigaci češtiny se však jeví jako nejvhodnější metoda hybridní, která spojuje dohromady oba modely.[26]

1.2.2.3 Složky anotačního systému

Kompletní anotační systém podle Leecha obsahuje tři základní složky, jsou to: „(a) *soubor značek* (*tagset*), tj. množina označení gramatických slovních druhů; (b) soubor *definic značek* (*tag definitions*) a (c) soubor *směrnic pro značkování* (*tagging guidelines*), které popisují přiřazování značek slovům ve značkovaném textu.“[27]

[23] JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 159.
[24] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 173.
[25] JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 161.
[26] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 170.
[27] LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 189.

1.3 Problémy při anotaci

Při vnitřní anotaci korpusu se musí lingvisté potýkat s mnoha problémy, které s sebou analýza jazyka a snaha o jednoznačné zařazení každého výrazu vyskytnuvšího se v korpusu přináší. Zabýváme-li se morfologickými systémy značkování a značkováním korpusů vůbec, je nutné připomenout alespoň některé z nich. Na morfologickou analýzu tvarů a lemmatizaci mají vliv jak jevy formální (chyby v textu, neznámá slova), tak i jevy lingvistické, např. homonymie, či existence analytických forem gramatických konstrukcí.

1.3.1 Chyby v textu a neznámá slova

Podstatnou komplikaci představují chyby v textu, ať už jsou způsobeny redaktorem, či autorem, automatické programy užívané při značkování je totiž neumí řešit.[28] Tato situace se ovšem netýká pouze slov s chybným pravopisem, Jelínek upozorňuje, že komplikace mohou zapříčinit i cizí slova, vlastní jména, málo frekventovaná slova či slova složená (např. česko-irský). Slova, která morfologická analýza nemůže v textu rozpoznat, protože nejsou zařazena ve slovníku, s nímž program (guesser) pracuje, a bude jim tak přiřazeno označení „neznámé slovo“, představují asi 1 % z celkového souboru.[29] Tato skutečnost se odráží na koncepci morfologického tagsetu existencí tzv. záchranné hodnoty, která je přiřazována výrazům, jež nelze zařadit do žádné jiné skupiny. Jsou to například: *neznámý, neurčený, neurčitelný slovní druh* (pozční tagset), *zbytek* (kompaktní tagset), *jiné (než slovní druhy 1–9, 0, F* (kódovník PMK).

1.3.2 Homonymie

Automatickou analýzu dále komplikuje v českém prostředí značně rozšířená tvarová homonymie. Jelínek její vliv na anotaci popisuje následovně: „Přiřazení morfologických značek a lemmat slovním tvarům není […] jednoduché, protože většina slov v českém textu (přibližně dvě třetiny) je tvarově nebo i slovnědruhově

[28]ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 13.
[29] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 172.

homonymních, nelze tedy určit značku pouze na základě tvaru slova, je třeba brát v úvahu kontext.“[30] Jak již bylo řečeno výše, k výběru správné značky a lemmatu k danému tvaru dochází procesem disambiguace, kterou je možné realizovat pomocí různých metod (statistické, či pravidlové). Ve snaze omezit chyby vzniklé při disambiguaci textu se přikročilo k tzv. kombinovaným metodám, ale „i v takto označkovaných datech se objevují některé typy systematicky chybného značkování,“[31] upozorňuje Jelínek. Materiál lze ještě dále podrobit automatickým opravám, avšak vždy v korpusu zůstane určité procento slov, která budou označkována nesprávně.[32] Homonymie může negativně ovlivnit úspěšnost distribuce korektních tagů u jednotlivých výrazů v korpusu.

1.3.3 Víceslovné výrazy

František Čermák řadí k nejvážnějším problémům, které se zatím korpusové lingvistice nedaří vyřešit, problém identifikace a označení víceslovných výrazů. Myslí tím jak tvary analytické, tak i skutečné víceslovné lexémy jako frazémy a termíny.[33] Z hlediska morfologického se problematika analytických tvarů závažněji projevuje např. u sloves.

Každé grafické slovo je v korpusu označkováno zvlášť, nejsou tedy brány v potaz analytické slovní tvary. V morfologických tagsetech se tato skutečnost projevuje značnou rozdílností ve zpracování gramatických kategorií u určitých slovních druhů. Nejvíce patrné je to u sloves, v každém systému značkování je určování gramatických kategorií u tohoto slovního druhu pojato jinak, a to i co se týče základních gramatických kategorií. Tento disharmonický jev v konečném důsledku způsobuje vzájemnou nekompatibilitu značkovacích systémů.

Z morfologického hlediska je sice závažnější problematika analytických tvarů, avšak víceslovné lexémy (zejm. frazémy), ač se přímo netýkají morfologie, jsou v některých tagsetech zohledňovány také. Váže se na ně tak existence či absence

[30] JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice*. Praha: Academia, 2008, s. 170.
[31] Tamtéž, s. 175.
[32] Tamtéž, s. 177.
[33] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 13.

speciálních hodnot, např. v kódovníku PMK určení slovního druhu jako: *frazém/idiom*.

1.4 Anotace a lingvistická teorie

Morfologické systémy značkování jsou sestavovány podle různých lingvistických koncepcí, každý z nich jiným způsobem a v různé míře odráží lingvistickou teorii. Geoffrey Leech, který zformuloval pro anotační systémy soubor zásad, jež by se měly při anotaci korpusů uplatňovat, se k teorii staví následovně: „Anotační systémy by se tedy měly co nejvíce zakládat na takových analýzách dat, které by byly neutrální ve vztahu k různým teoriím a byly založené na ‚konsensu'."[34] Takové východisko se v praxi jeví jako značně problematické a představuje ideál, jenž je obtížně realizovatelný. Dokládá to i František Čermák ve své studii *Korpusová lingvistika dnešní doby*, kde konstatuje, že značkování je „založené vždy na určité jazykové teorii a tedy i jen jedné interpretaci množství sporných slov a tvarů."[35] Taková situace má za následek „vnesení výkladu té které gramatiky do korpusu a uživateli se tak vlastně vnucuje gramatická filozofie autorů takové gramatiky."[36] I Jan Hajič v komentáři k pozičnímu systému zmiňuje, že dané pojetí jazyka nemusí být (a není) ve shodě se všemi dostupnými lingvistickými teoriemi a názory na ně. Např. v souvislosti s určováním slovních druhů upozorňuje „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'."[37]

Spjatost anotačního systému s teorií dále souvisí také s jeho podrobností čili s tím, s jakou jemností odráží morfologickou charakteristiku daných slov. Leech upozorňuje na fakt, že různé systémy značkování se mohou svou obsáhlostí značně

[34] LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 186.
[35] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 11.
[36] Tamtéž, s. 11.
[37] HAJIČ, Jan. Popis morfologických značek – poziční systém. In: KOPŘIVOVÁ, Marie a Jan KOCEK. *Manuál korpusového manažeru bonito* [online]. [cit. 2014-03-16]. Dostupné z: http://ucnk.ff.cuni.cz/bonito/znacky.php.

lišit, některá značkování jsou elementární, jiná podrobnější,[38] s ohledem na tuto skutečnost lze konstatovat, že čím je „systém jednodušší, tím méně je pravděpodobné, že se dostane do rozporu s předpoklady té či oné teorie.“[39] Ve všech pozorovaných tagsetech je například přítomno tradiční rozdělení slov do deseti slovních druhů, nicméně jde-li o bližší specifikaci slovního druhu, kterou můžeme vnímat už jako jakýsi „nadstandard“, spatřujeme mezi tagsety značné diference.

Čermák dále upozorňuje, že anotace může vést ke zjednodušujícímu pohledu na jazykovou realitu.[40] Takový problém může vzniknout zejména „u jazyků flektivních, které disponují velkým množstvím tvarů a také jejich vysokou homonymií.“[41] O tom, jak tvarová homonymie komplikuje anotaci, viz výše.

[38] LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 194.
[39] Tamtéž, s. 195.
[40] ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 11.
[41] Tamtéž, s. 11.

2 Systémy morfologického značkování češtiny

V současné době má česká korpusová lingvistika k dispozici několik tagsetů neboli systémů pro morfologické značkování českých jazykových korpusů. Tyto tagsety jsou koncipovány podle různých lingvistických koncepcí, liší se v míře komplexnosti reflexe morfologických kategorií.[42] Systémy se od sebe dále liší podobou tagů, která může být více či méně uživatelsky přívětivá, či například užitou terminologií. V této bakalářské práci se budeme podrobněji zabývat pěti gramatickými anotačními systémy, jsou to: poziční morfologický tagset, atributivní morfologický tagset, kompaktní morfologický tagset, Xerox tagset a „kódovník" PMK.

Naším úkolem je porovnat, do jaké míry jednotlivé tagsety reflektují českou morfologii. Na základě tohoto srovnání bychom měli být schopni zvážit možnosti jejich vzájemné kompatibility. V následující části této kapitoly stručně charakterizujeme jednotlivé anotační systémy a uvedeme základní zdroje informací, ze kterých jsme čerpali data pro další studium zmíněných tagsetů.

2.1 Poziční morfologický tagset

Poziční tagset (pražský) byl navržen Janem Hajičem z Ústavu formální a aplikované lingvistiky MFF UK. V současné době se jedná o nejpoužívanější systém pro morfologické značkování českých korpusů. Můžeme říct, že se stal kanonickým tagsetem pro značkování korpusů v projektu Českého národního korpusu (ČNK), jenž je spravován Ústavem pro český národní korpus (ÚČNK).

Název tagsetu – poziční – vychází z pojetí tagů, které jsou přiřazovány jednotlivým slovům (grafickým slovům, izolovaným slovním tvarům) v korpusu, ukázky tagů viz dále v kap. č. 2.6 (Tabulka 1). Každý tag v systému má stejný počet pozic (původně 15, nyní 16), každá pozice odpovídá jedné morfologické kategorii, pořadí těchto kategorií je neměnné. Pozice jsou obsazovány definovanými hodnotami, jež jsou zastoupeny jedním znakem, nejčastěji velkým písmenem abecedy či číslicí, pokud je pro danou kategorii definováno větší množství hodnot (např. u pozice 2, slovní poddruh), může být znakem zastupujícím příslušnou hodnotu kromě písmena či

[42] POŘÍZKA, Petr a Markus SCHÄFER. MorphCon – A software for Conversion of Czech Morphological Tagsets. In: LEVICKÁ, Jana, ed. a GARABÍK, Radovan, ed. *NLP, Corpus Linguistics, Corpus Based Grammar Research*. Brno: Tribun, 2009, s. 292.

číslice také interpunkční znaménko, případně jiný symbol. V případě, že se daná kategorie u určitého slovního tvaru neurčuje, je obsazena znakem „-“.[43] Charakteristiku jednotlivých pozic a plný soupis hodnot, které se na nich mohou vyskytovat, nalezneme v *Manuálu korpusového manažeru Bonito*, odkud pro naši práci čerpáme informace o pozičním systému.[44]

Popis morfologických značek je v *Manuálu* koncipován na základě pozic, postupně jsou představeny pozice tagu a hodnoty na nich definované. Není koncipován na základě slovních druhů jako např. seznam tagů atributivního systému či kompaktního systému, kde je uveden nejprve slovní druh a posléze k němu přidružené kategorie a hodnoty a kde tak uživatel okamžitě ví, jaké kategorie jsou pro daný slovní druh v tagsetu dostupné. U pozičního systému se uživatel s jistotou dozví, které kategorie a hodnoty je možné u slovního druhu určovat, až z *programu pro vytváření tagů*[45]. Z manuálu se to názorně nedozví. *Program pro vytváření tagů* je aplikací, která má pomoci uživatelům ČNK jednodušeji se orientovat v systému pozičních tagů a prostřednictvím této aplikace i vytvářet konkrétní korpusové dotazy.[46] Její princip spočívá v tom, že po zadání hodnoty na první pozici (slovní druh) nám nabídne výběr kategorií (a hodnot), jež jsou pro vybraný slovní druh v systému nadefinovány. Zmíněný princip však platí pouze pro pozice 1–12, pozice 13–16, jak uvádějí tvůrci: „se nemění, ani u nich nelze vybrat hodnotu – jsou zde [v programu pro vytváření tagů] jen pro informaci.“[47] U pozic 13 a 14 daná situace není na překážku, jelikož tyto pozice nejsou použity pro žádnou morfologickou kategorii. U pozic 15 a 16 však již musí uživatel zvážit možnosti sám, případně postupovat metodou pokusu a omylu při zadávání dotazu do korpusového manažeru.

Jak bylo zmíněno výše, v této práci jsme vycházeli z materiálů obsažených v *Manuálu korpusového manažeru Bonito*. Pro komparaci systému s lingvistickou teorií však bylo nezbytně nutné u pozičního tagsetu ke každému slovnímu druhu přiřadit kategorie a hodnoty pro něj v systému definované. Náš postup při

[43] HAJIČ, Jan. Popis morfologických značek – poziční systém. In: KOPŘIVOVÁ, Marie a Jan KOCEK. *Manuál korpusového manažeru bonito* [online]. [cit. 2014-03-16]. Dostupné z: http://ucnk.ff.cuni.cz/bonito/znacky.php.
[44] Tamtéž, [cit. 2014-03-16].
[45] SKOUMALOVÁ, Hana. *Program pro vytváření tagů* [online]. [cit. 2014-03-16]. Dostupné z: http://utkl.ff.cuni.cz/~skoumal/morfo/?lang=cs.
[46] Tamtéž, [cit. 2014-03-16].
[47] Tamtéž, [cit. 2014-03-16].

přiřazování kategorií ke slovním druhům byl následující: kategorie u pozic 1–12 jsme k jednotlivým slovním druhům přiřadili na základě výše představeného programu pro vytváření tagů; pozice 13 a 14 nejsou obsazovány žádnými hodnotami, tudíž nebylo nutné se jimi zabývat; kategorie na pozicích 15 a 16 jsme ke slovním druhům přiřazovali na základě dat získaných z korpusového manažeru Bonito. Do vyhledávacího řádku manažeru Bonito jsme zadali dotaz ve tvaru např. [tag="N……..(B|P|I)"] (to představuje substantiva, u kterých je definována jakákoli kladná hodnota pro vid), pokud byly na základě zadaného dotazu nalezeny výsledky, kategorii jsme k příslušnému slovnímu druhu zařadili. Šetření jsme prováděli v korpusu SYN2005, jelikož právě s jeho uveřejněním byla kategorie vidu do tagsetu přidána. Ještě je nutné podotknout, že *Manuál* upozorňuje na to, že některé hodnoty jsou k dispozici pouze v některých korpusech ČNK (SYN2006PUB, SYN2005, SYN2000, ORWELL). Tyto hodnoty byly v naší práci zohledňovány také.

2.2 Atributivní morfologický tagset

Atributivní morfologický tagset (brněnský) vytvořili Klára Osolsobě a Radek Sedláček, vznik systému úzce souvisí s vývojem morfologického analyzátoru AJKA.[48] Analyzátor AJKA[49] a potažmo i atributivní tagset je spojen s Centrem pro zpracování přirozeného jazyka (CZPJ) na Fakultě informatiky Masarykovy univerzity v Brně (též Natural Language Processing Centre, zkráceně NLPC). Činnost CZJP je zaměřená na výsledky v oblastech informačních technologií a jazykovědy, mimo jiné se také podílelo například na vzniku Internetové jazykové příručky.[50] Brněnský systém je po pozičním systému druhým nejrozšířenějším způsobem morfologického značkování psaných českých korpusů. Jsou jím označkovány např. korpusy CZPJ, dále také Korpus soukromé korespondence (pod označením *KSKdopisy1*).

[48] POŘÍZKA, Petr a Markus SCHÄFER. MorphCon – A software for Conversion of Czech Morphological Tagsets. In: LEVICKÁ, Jana, ed. a GARABÍK, Radovan, ed. *NLP, Corpus Linguistics, Corpus Based Grammar Research*. Brno: Tribun, 2009, s. 292.

[49] V současné době již existuje nová verze nazvaná MAJKA, která i nadále používá stejný tagset. (CZPJ. Free natural language morphology for Czech, Slovak, Polish and English. *Centrum zpracování přirozeného jazyka* [online]. [cit. 2014-04-24]. Dostupné z: http://nlp.fi.muni.cz/ma/.)

[50] CZPJ. Na čem pracujeme v CZPJ? *Centrum zpracování přirozeného jazyka* [online]. © 2001–2013, poslední aktualizace 11. 4. 2013 [cit. 2014--03-16]. Dostupné z: http://nlp.fi.muni.cz/web2/cgi-bin/index.py?page=main_topics&language=cs.

Atributivní tagset je založen na odlišném principu než systém poziční. Na první pohled viditelným rozdílem je nestejná délka jednotlivých tagů, viz kap. č. 2.6, (Tabulka 1). V atributivním systému jsou totiž u každého slovního druhu zohledňovány pouze ty kategorie, které je u něj z lingvistického hlediska možné určovat; pokud se nějaká kategorie u daného slovního druhu neurčuje, chybí. Ve výsledku tak každý slovní druh disponuje specifickým počtem kategorií a jejich unikátní kombinací. Z toho vyplývá, že jedna morfologická kategorie se může v různých tazích vyskytovat na odlišných pozicích. Tato skutečnost představuje zásadní rozdíl oproti pozičnímu systému. Značky také mají specifickou vnitřní výstavbu. Každá určovaná kategorie je zastoupena v tagu dvěma znaky. První znak určuje, o jakou kategorii se jedná, a má podobu malého písmena abecedy, často vychází z názvu dané morfologické kategorie, např. „c“ pro pád (casus), „g“ pro rod (genus) apod., není to však pravidlem, např. „w“ pro stylistický příznak. Druhý znak představuje hodnotu, jež je danému výrazu ve specifikované kategorii přisouzena, jedná se buď o velké písmeno abecedy, např. „F“ pro ženský rod (femininum), „N“ pro střední rod (neutrum) apod., nebo o číslici, např. „1–7“ u pádu. Tento systém může být pro uživatele při dekódování tagů čitelnější, intuitivně pochopí význam některých spojení znaků kategorie a hodnoty, např. „gF“ (ženský rod), „c3“ (třetí pád) ad.

V popisu atributivního tagsetu jsme vycházeli z dokumentu Radka Sedláčka, který obsahuje úplný soupis značek systému a byl aktualizován v březnu 2006.[51] Z tohoto dokumentu čerpáme data i pro další studium atributivního systému.

2.3 Kompaktní morfologický tagset

Další tagsety, které si představíme, již na poli českých korpusů nejsou tak významně zastoupeny. Mnohdy se váží ke specifickým projektům. Kompaktní tagset byl vytvořen Vladimírem Petkevičem z Ústavu teoretické a komputační lingvistiky na FF UK v Praze. V současné době je v rámci ČNK kompaktními značkami anotován pouze jediný korpus, a to korpus ORWELL. Korpus ORWELL je tvořen textem románu George Orwella *1984*, existuje ve dvou podobách: je označkován

[51] SEDLÁČEK, Radek. *AJKA tagset* [online]. © 2006 [cit. 2014-03-19]. Dostupné z: http://nlp.fi.muni.cz/projekty/ajka/tags.pdf.

kompaktním systémem (pod označením *orw-mte*) a rovněž systémem pozičním (pod označením *orwell*).[52]

Pokud jde o strukturu značek kompaktního systému, můžeme zde vypozorovat jisté rysy, které odkazují k systému pozičnímu i atributivnímu. Lze říci, že přístup obou těchto systémů k tagům do jisté míry kombinuje, srov. kap. č. 2.6 (Tabulka 1). Kompaktní tagset se vyznačuje nestejnou délkou tagů napříč slovními druhy, zde se podobá spíše systému atributivnímu, kde je délka v závislosti na slovním druhu variabilní. Co však již společného nemají, je vnitřní struktura značek, v tomto směru je kompaktní systém spíše podobný pozičnímu, na rozdíl od atributivního totiž není jedna kategorie zastoupena dvěma značkami, ale pouze značkou jedinou podobně, jako je tomu právě v pražském tagsetu. Tagy jsou také poziční, pro každou kategorii je v rámci příslušného slovního druhu vyhrazena jistá pozice, která může být obsazena pouze hodnotami definovanými pro danou kategorii.[53]

Podrobné informace o kompaktním tagsetu nalezneme na stránkách ÚČNK, odkud jsme také čerpali data pro další studium.[54] Seznam značek je koncipován podle slovních druhů, u každého z nich jsou uvedeny pouze ty kategorie, které je u něj možné z lingvistického hlediska v tagsetu určovat, v závislosti na slovním druhu se jejich počet mění. Manuál kompaktního systému je přehledný, za každým slovním druhem následuje příklad (či příklady) užití tagu, např. příslovce „lépe“: [tag="Rgc"] (příslovce, obecné, komparativ).[55] Tyto příklady jsou uváděny ve tvaru, který je možné přímo zadat do vyhledávacího řádku korpusového manažeru (zde Bonito), uživatel má tak už při studiu manuálu možnost vidět užití tagů v praxi.

2.4 Xerox tagset

Tento tagset byl vyvinut společností Xerox a je k dispozici na webovém portále Open Xerox.[56] Na těchto webových stránkách Xerox Corporation zpřístupňuje pro externí uživatele technologie, které jsou mnohdy stále v procesu vývoje. Jedním

[52] ÚČNK. Korpus ORWELL. *Český národní korpus* [online]. [cit. 2014-03-19]. Dostupné z: http://ucnk.ff.cuni.cz/orwell.php.
[53] PETKEVIČ, Vladimír. Popis morfologických značek použitých v korpusu orw-mte. In: *Český národní korpus* [online]. [cit. 2014-03-28]. Dostupné z: http://ucnk.ff.cuni.cz/orwell.php.
[54] Tamtéž, [cit. 2014-03-28].
[55] Tamtéž, [cit. 2014-03-28].
[56] XEROX CORPORATION. About. *Open Xerox* [online]. ©1999–2014 [cit. 2014-03-23]. Dostupné z: http://open.xerox.com/Pages/About.

z odvětví technologických výzkumů, kterým se společnost Xerox věnuje, jsou lingvistické nástroje. V současnosti společnost Xerox nabízí veřejnosti řadu lingvistických nástrojů zaměřených na zpracování přirozeného jazyka, mezi tyto nástroje patří také morfologický anotační systém Xerox (Xerox tagset).[57] Kromě češtiny se společnost Xerox ve vývoji lingvistických nástrojů zaměřuje na dalších osm evropských jazyků: francouzštinu, angličtinu, němčinu, řečtinu, maďarštinu, italštinu, polštinu a ruštinu.[58] Nejen, že Xerox tagset nebyl primárně vyvíjen pro potřeby češtiny, ale dokonce nebyl vyvíjen ani pro účely čistě lingvistické. Hlavním cílem společnosti Xerox při vývoji zmíněných lingvistických nástrojů bylo jejich využití v jiných projektech, např. v technologii OCR (optické rozpoznávání znaků), u tvorby digitálních knihoven či překladatelských systémů (tzv. překladačů).[59] S tím souvisí skutečnost, že tímto tagsetem není, pokud je nám známo, anotován žádný existující korpus. Anotovaný text získáme zadáním příslušného textu do on-line nástroje pro anotaci[60], který zadaný text v reálném čase označkuje a nabídne výsledný produkt. Nevýhodou je, že při tomto způsobu anotace je procento chyb v přiřazování tagů k jednotlivým výrazům poměrně vysoké. Pro demonstraci jsme do zmíněného nástroje pro anotaci zadali náhodně vybranou část textu:

„Chodba páchla vařeným zelím a starými hadrovými rohožkami. Na stěně na jednom konci úzkého prostoru byl připíchnut barevný plakát, který se svou velikostí dovnitř nehodil. Byla na něm jen obrovská tvář muže, asi pětačtyřicetiletého, s hustým černým knírem, drsných, ale hezkých rysů."

Úryvek obsahuje 50 položek včetně interpunkce, z toho čtyři byly označeny chybně („hadrovými" jako substantivum místo adjektivum, „který" jako adjektivum místo zájmeno, „svou" jako adjektivum místo zájmeno, „jen" jako částice místo příslovce), tzn., že chybovost v přiřazování tagů je v případě této ukázky 8 %.

[57] XEROX CORPORATION. *Open Xerox* [online]. ©1999–2014 [cit. 2014-03-23]. Dostupné z: http://open.xerox.com/.
[58] XEROX CORPORATION. Part-of-Speech Tagsets. *Open Xerox* [online]. ©1999–2014 [cit. 2014--03-23]. Dostupné z: http://open.xerox.com/Services/fst-nlp-tools/Pages/Part-of-Speech%20Tagsets.
[59] XEROX CORPORATION. Linguistic tools. *Open Xerox* [online]. ©1999–2014 [cit. 2014-03-23]. Dostupné z: http://open.xerox.com/Services/fst-nlp-tools.
[60] XEROX CORPORATION. Part of Speech Tagging (Real Time). *Open Xerox* [online]. ©1999–2014 [cit. 2014-03-23]. Dostupné z: http://open.xerox.com/Services/fst-nlp-tools/Consume/181.

Při studiu tohoto tagsetu vycházíme ze seznamu značek dostupného na stránkách Xerox Corporation.[61] Vzhledem k okolnostem svého vzniku je tento anotační systém výrazně odlišný od ostatních systémů morfologického značkování češtiny. Ze všech popisovaných tagsetů je nejméně podrobný, zřídka se v něm objevuje i jiná morfologická kategorie než slovní druh. Terminologie vychází z angličtiny, jazyka tvůrců tagsetu. Struktura značek se od ostatních morfologických anotačních systémů také značně liší, viz kap. č. 2.6 (Tabulka 1). Značky jsou psány verzálkami, mají podobu celých termínů či jejich zkratek a jsou uvedeny znaménkem „+“, např.: „+NOUN“, substantivum; „+ADJ“, adjektivum apod. Hodnoty definované v Xerox tagsetu pro značkování češtiny jsou v seznamu řazeny abecedně podle značek, jelikož vycházejí z anglických termínů, může být pro uživatele navyklého na termíny české, případně mezinárodní obtížné se v seznamu rychle orientovat. Kladně hodnotíme skutečnost, že jsou v soupisu značek uvedeny kromě popisů jednotlivých tagů také příklady výrazů, kterým mohou být tagy přiřazeny.

2.5 Kódovník PMK

V kódovníku PMK má čeština svůj nejpodrobnější morfologický anotační systém. Zatímco výše uvedené systémy jsou primárně spjaty s korpusy psanými, kódovník byl vyvinut a použit pro označkování korpusu mluveného, konkrétně Pražského mluveného korpusu (PMK). Označkována je však pouze verze PMK vydaná knižně (*Frekvenční slovník mluvené češtiny*).[62] Pražský mluvený korpus je prvním korpusem mluvené češtiny u nás, zachycuje autentickou českou mluvu především z oblasti Prahy a jejího okolí. PMK je kolektivním dílem, pracovali na něm především: Anna Adamovičová, František Čermák, Jiří Pešička, Josef Šimandl, Jitka Šonková, Petr Savický a Zdena Smetanová z UK FF, s nahrávkami zmíněným autorům pomáhala i řada studentů.[63]

Značky užité v kódovníku mají skutečně podobu číselných kódů, viz Tabulka 1 v kap. č. 2.6. Každé určované kategorii je v tagu přisouzena jedna pozice, hodnoty

[61] XEROX CORPORATION. Czech Part-of-Speech Tagset. *Open Xerox* [online]. ©1999–2014 [cit. 2014-02-01]. Dostupné z: http://open.xerox.com/Services/fst-nlp-tools/Pages/Czech%20Part-of-Speech%20Tagset.
[62] ČERMÁK, František et al. *Frekvenční slovník mluvené češtiny*. Praha: Karolinum, 2007.
[63] ČERMÁK, František. Pražský mluvený korpus. ÚČNK. *Český národní korpus* [online]. [cit. 2014--03-30]. Dostupné z: http://ucnk.ff.cuni.cz/pmk.php.

definované pro danou kategorii jsou v tagu zastoupeny číslicemi, pokud číslice nepostačují, jsou užita velká písmena abecedy (užití písmen je však v tomto systému spíše výjimkou). Zastoupení kategorie v tagu jedním znakem má blízko k systému pozičnímu, avšak délka tagu závisí na slovním druhu. Každý slovní druh má specifický počet určovaných kategorií, tedy i specifickou délku značky, z toho vyplývá, že totožná kategorie nemusí být u dvou různých slovních druhů na stejné pozici. Podobně je tomu například u systému kompaktního či atributivního. První pozice je jediná, jejíž význam se nemění, určuje slovní druh. Dále zůstává stejná také poslední pozice kódu označující styl označovaného výrazu, je uváděná za lomítkem odděleně od ostatních kategorií, vzhledem k variabilní délce tagů nelze přesně určit její pořadí.

Kódovník PMK umožňuje dva způsoby kódování, plné a redukované. Při redukovaném kódování jsou v tagu zahrnuty pouze některé kategorie obsažené v kódování plném, na které se v této práci zaměřujeme. Systém plného kódování je detailně vyložen ve *Frekvenčním slovníku mluvené češtiny*[64], jenž nám slouží jako základní literatura pro studium tohoto tagsetu. Značky jsou v seznamu řazeny podle slovních druhů, u každého druhu jsou na příslušných pozicích uvedeny ty kategorie, které jsou u něj určovány. Vedle nadefinovaných hodnot zde také ve většině případů nalezneme příklady výrazů, ke kterým může být příslušná hodnota přiřazena.

2.6 Příklady tagů – vizuální srovnání

Na závěr této kapitoly zde přinášíme přehled tagů z jednotlivých systémů (Tabulka 1). Můžeme tak porovnat jejich konkrétní podobu, kterou jsme se pokusili u daných anotačních systémů charakterizovat výše v textu. V Tabulkách 2–6 popisujeme význam jednotlivých znaků v uvedených tazích, zaměřujeme se na příklady tagů pro substantivum „dům“.

[64] ČERMÁK, František et al. *Frekvenční slovník mluvené češtiny*. Praha: Karolinum, 2007, s. 15–24.

TAGSET	TAG: „dům“	TAG: „a“
Poziční	NNIS1-----A-----	J^----------
Atributivní	k1gInSc1	k8
Kompaktní	Ncmsn	Cc
Xerox	+NOUN	+CONJ
Kódovník PMK	11902111/1	8111/1

Tabulka 1: Ukázky tagů na příkladech substantiva „dům“ a spojky „a“

	KATEGORIE	HODNOTA
N	slovní druh	substantivum
N	slovní poddruh	substantivum, obyčejné
I	rod	maskulinum inanimatum (rod mužský neživotný)
S	číslo	singulár (jednotné číslo)
1	pád	nominativ (1. pád)
-	přivlastňovací rod	neurčuje se
-	přivlastňovací číslo	neurčuje se
-	osoba	neurčuje se
-	čas	neurčuje se
-	stupeň	neurčuje se
A	negace	afirmativ (bez negativní předpony „ne-“)
-	aktivum/pasívum	neurčuje se
-	nepoužito	-
-	nepoužito	-
-	varianta, stylový příznak apod.	neurčuje se
-	vid	neurčuje se

Tabulka 2: Popis kategorií a hodnot pozičního systému na příkladu tagu pro substantivum „dům“

	KATEGORIE	HODNOTA
k1	k = slovní druh	1 = substantivum
gI	g = rod	I = rod mužský neživotný
nS	n = číslo	S = číslo jednotné
c1	c = pád	1 = pád první

Tabulka 3: Popis kategorií a hodnot atributivního systému na příkladu tagu pro substantivum „dům“

	KATEGORIE	HODNOTA
N	slovní druh	podstatné jméno /Noun/
c	typ	obecné
M	rod	femininum (ženský rod)
s	číslo	singulár (jednotné číslo)
n	pád	nominativ (1. pád)

Tabulka 4: Popis kategorií a hodnot kompaktního systému na příkladu tagu pro substantivum „dům“

	KATEGORIE	HODNOTA
+NOUN	slovní druh	noun

Tabulka 5: Popis kategorií a hodnot Xerox tagsetu na příkladu tagu pro substantivum „dům“

	KATEGORIE	HODNOTA
1	slovní druh	substantivum
1	druh	běžné
9	třída	jiné/nejasné
0	valence	bez valence
2	rod	maskulinum inanimatum
1	číslo	singulár
1	pád	nominativ
1	funkce	subjekt
1	styl	základní styl neformálních ústních projevů

Tabulka 6: Popis kategorií a hodnot kódovníku PMK na příkladu tagu pro substantivum „dům“

Na uvedených příkladech můžeme pozorovat charakteristické rysy jednotlivých tagsetů. Poziční systém se na rozdíl od ostatních vyznačuje neměnnou délkou tagů. Pro systém atributivní je typické sdružování znaků do dvojic, v nichž první člen představuje kategorii (např. „k“, slovní druh) a druhý vybranou hodnotu (pro „k“: „1“, substantivum; „8“, spojka). Kompaktní systém se v proměnlivé délce podobá systému atributivnímu, na rozdíl od něj je však jedna kategorie v tagu zastoupena jedním znakem, nikoliv dvojicí znaků (bigramem). Charakter tagů Xerox tagsetu se od ostatních systémů liší již na první pohled, oba uvedené tagy, ačkoli obsahují pět znaků, představují pouze jedinou kategorii, a to slovní druh. Povšimněme si, že tagy kódovníku PMK skutečně mají podobu číselných kódů, bez manuálu jsou zřejmě nejobtížněji dešifrovatelné. Nestejnou délkou tagů vzhledem k slovním druhům se pak kódovník podobá systému atributivnímu a kompaktnímu. Není pochyb o tom, že se mezi sebou jednotlivé tagsety velmi liší, a to nejen podobou svých tagů. I když již

ta nám napovídá, že srovnání těchto anotačních systémů z hlediska jejich vzájemné kompatibility může být velmi problematické.

3 Odraz morfologie češtiny v morfologických tagsetech

Tagsety jsou založeny na různých koncepcích a i česká morfologie je popsána v několika příručkách, jež přinášejí teoretický výklad gramatiky češtiny. Pojetí morfologické problematiky se v nich v mnoha ohledech liší. Nicméně naším úkolem je porovnat, do jaké míry jednotlivé tagsety odrážejí českou morfologii. Ke komparaci systémů značkování s českou gramatikou jsme vybrali *Mluvnici češtiny 2* neboli tzv. *akademickou Mluvnici češtiny 2* (dále MČ2).

3.1 Mluvnice češtiny 2

Tzv. *akademická Mluvnice češtiny* přináší ve třech svazcích komplexní synchronní výklad gramatiky českého jazyka. Je nejobsáhlejší gramatikou současné spisovné češtiny, a ačkoli od vydání prvního svazku uplynulo již více než dvacet let, je stále jedním ze stěžejních děl moderní české lingvistiky. *Mluvnice češtiny* je stejně jako ostatní gramatiky českého jazyka založena na popisu jazykového systému a výběrově informuje o způsobech užívání jazykových prostředků.[65] Z hlediska obsahového je členěna do oddílů tematicky se věnujících jednotlivým jazykovým rovinám. Tradiční řazení oddílů je následující: hláskosloví, tvarosloví, slovotvorba, větná a nadvětná syntax, případně syntax textu, eventuálně i stylistika.[66] *Mluvnice češtiny* toto řazení respektuje, avšak v rozporu s tradicí zaměnila pořadí výkladů o slovotvorbě a tvarosloví.

Vzhledem k charakteru bakalářské práce nás budou zajímat ty části MČ2, které jsou věnovány morfologii, jedná se o druhý díl, pojednávající o tvarosloví, a závěrečné kapitoly dílu prvního, které se zabývají slovotvorbou. Druhý svazek je rozdělen na dva oddíly: funkční a formální tvarosloví. Pro naši práci je stěžejní oddíl funkčního tvarosloví, jelikož obsahuje výklad o jednotlivých slovních druzích a jejich gramatických kategoriích. Určení slovního druhu daného tvaru a jeho gramatická charakteristika, tedy určení gramatických kategorií, které u něj lze pozorovat, jsou jádrem morfologického tagsetu. Oddíl formálního tvarosloví je věnován ohebným

[65] KOŘENSKÝ, Jan. Jaké gramatiky češtiny dnes a zítra? *Naše řeč*. 2007, č. 4, s. 170.
[66] Tamtéž, s. 170–171.

slovním druhům, přináší popis systému tvarů (paradigmatiky) současného spisovného jazyka a zabývá se mimo jiné také morfematickou strukturou slovních tvarů. Slovotvorba, která rovněž spadá do oblasti morfologie, je, jak již bylo řečeno výše, součástí dílu prvního. Nevěnujeme jí příliš pozornosti, jelikož otázky slovotvorby jsou v morfologických tagsetech reflektovány pouze okrajově, jsou-li vůbec reflektovány.[67]

Mluvnici češtiny 2 jsme ke srovnání morfologických tagsetů s lingvistickou teorií vybrali pro komplexní a erudované zpracování dané problematiky. V její prospěch také mluví to, že je stále respektovaná v odborných kruzích. Již po svém vydání byl druhý díl *Mluvnice češtiny* odbornou veřejností přijat kladně, dokládá to posudek druhého svazku publikovaný v časopise *Naše řeč*, kde jej Alois Jedlička hodnotí následovně: „Druhý díl Mluvnice češtiny představuje důkladné a všestranné, metodologicky tvůrčím způsobem promyšlené a materiálově bohatě dokumentované zpracování českého tvarosloví v rovině funkční a formální.“[68] Stále také funguje, i přes různé výtky, např. ohledně popisu adjektivní flexe[69], jako základní a reprezentativní příručka pro studium morfologie češtiny.

3.2 Komparace tagsetů a teorie

V souladu s koncepcí MČ2 jsme se rozhodli při komparaci morfologických tagsetů s lingvistickou teorií postupovat podle jednotlivých slovních druhů. Pozornost upřeme především na **základní gramatické kategorie**, které jsou u jednotlivých slovních druhů určovány. Zde jsou základními kategoriemi míněny ty gramatické kategorie, které jsou u slovních druhů tradičně určovány, např. rod, číslo a pád u substantiv. Nejprve uvedeme výklad těchto základních gramatických kategorií podle MČ2. Dále nás bude zajímat, do jaké míry (a zda vůbec) jsou reflektovány v jednotlivých tagsetech.

Následně porovnáme s lingvistickou teorií také další kategorie tagsetů, zde nazvané **rozšiřující kategorie**, tedy kategorie, které jsou u daného slovního druhu v tagsetu

[67] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986.
[68] JEDLIČKA, Alois. Druhý svazek akademické Mluvnice češtiny. *Naše řeč*. 1989, č. 3, s. 151.
[69] Srov. BEDNAŘÍKOVÁ, Božena. *Slovo a jeho konverze*. Olomouc: Univerzita Palackého v Olomouci, 2009, s. 86–88.

určovány (jsou pro ně v tagsetu definovány hodnoty), a současně neodrážejí základní gramatické kategorie slovního druhu.

3.2.1 Substantiva

3.2.1.1 Základní kategorie

Jako základní gramatické kategorie (prostředky) určované u podstatných jmen označuje MČ2 rod, číslo a pád. U **gramatického rodu** je brán ohled i na životnost, ve výsledku tedy vydělujeme čtyři třídy substantiv: *mužská životná* (*masculina animata*), *mužská neživotná* (*masculina inanimata*), *ženská* (*feminina*) a *střední* (*neutra*). **Gramatické číslo** vyjadřuje protiklad jednosti a nejednosti, kdy jednost je vyjádřena formou *čísla jednotného* (*singuláru*) a nejednost formou *čísla množného* (*plurálu*), v tvarových paradigmatech některých substantiv nalezneme zbytky tzv. *dvojného čísla* (*duálu*), konkrétně se objevuje částí těla vyskytujících se po dvou (oči, uši, ruce apod.). MČ2 také hovoří o zvláštních případech, kdy se některých substantiv užívá pouze (převážně) v singuláru nebo pouze v plurálu, ale tyto případy významně souvisí se sémantikou těchto jmen, nás však zajímá především hledisko formální, tudíž zůstaneme u třech základních určení čísla: singuláru, plurálu a duálu. Čeština rozlišuje sedm **pádů**: *první pád* (*nominativ*), *druhý pád* (*genitiv*), *třetí pád* (*dativ*), *čtvrtý pád* (*akuzativ*), *pátý pád* (*vokativ*), *šestý pád* (*lokál*) a *sedmý pád* (*instrumentál*).[70]

Poziční tagset odráží všechny výše zmíněné základní gramatické kategorie substantiv. V kategorii rodu je zahrnuta v souladu s MČ2 také životnost, k dispozici jsou tedy pro mluvnický ROD čtyři základní hodnoty: *femininum (ženský rod)*, *maskulinum inanimatum (rod mužský neživotný)*, *maskulinum animatum (rod mužský životný)* a *neutrum (střední rod)*, kromě nich je v pozičním systému nadefinována také hodnota pro tzv. *libovolný rod*, jejíž přítomnost v tagsetu je zapříčiněna problematickým určováním gramatických kategorií u některých výrazů, obzvlášť těch cizího původu, více v kapitole č. 1.3. Gramatické ČÍSLO je v tagu představováno hodnotami pro *duál*, *plurál (množné číslo)* či *singulár (jednotné číslo)*, jsou tak zastoupeny všechny varianty, které se odrážejí ve formální podobě tvaru, také je k dispozici hodnota související s problematickým určováním kategorií, tzv. *libovolné*

[70] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 29–66.

číslo (P/S/D). V pozičním systému se určuje tradičně sedm PÁDŮ, jak je avizováno v popisu tohoto tagsetu.

Všechny základní kategorie určované u substantiv rovněž najdeme i v **atributivním systému** značkování. Hodnoty u nich určované jsou vesměs tradiční, pro ROD zde nalezneme hodnoty zahrnující i životnost: *rod mužský životný*, *rod mužský neživotný*, *rod střední*, *rod ženský*; u ČÍSLA najdeme hodnotu pro *číslo jednotné, číslo množné* i *duál*, definováno je všech sedm PÁDŮ. V atributivním systému najdeme u základních kategorií také jedno specifikum, tím je zohledňování zvláštní skupiny substantiv, a to jmen označujících rodinu (příjmení), v MČ2 jsou tato substantiva zařazena do třídy skupinových antroponym, jedná se o antroponymické názvy celků, pro které je množné číslo základní formou užití, kromě zmíněných jmen rodinných (např. Novákovi) zde spadají také jména rodová (Přemyslovci) či jména národní, kmenová a obyvatelská (např. Francouzi, Hanáci, Pražani).[71] Pro tato jména jsou definovány speciální hodnoty v kategorii rodu, *rodina (příjmení)*, a čísla, *hromadné označení členů rodiny (Novákovi)*. Tím se tagset liší v pohledu na tuto problematiku od MČ2, která tyto speciální kategorie v rámci gramatického rodu a čísla nereflektuje. Zároveň však MČ2 přiznává, že propria se v oblasti rodu a čísla vyznačují jistou specifičností.[72]

V **systému kompaktním**, jenž také postihuje všechny základní kategorie, je na rozdíl od výše zmíněných tagsetů životnost vydělena do zvláštní kategorie. Tři základní gramatické prostředky jsou tak v tagsetu zastoupeny čtyřmi kategoriemi, rodem, číslem, pádem a životností. V kompaktním systému najdeme elementární hodnoty typické pro danou základní gramatickou kategorii: ROD (*mužský*, *ženský* a *střední*), ČÍSLO (*singulár*, *plurál* i *duál*), všech sedm PÁDŮ a ŽIVOTNOST (*životné*, *neživotné*).

Xerox tagset oproti výše zmíněným systémům nereflektuje žádnou ze základních gramatických kategorií uvedených v MČ2. V popisu tohoto tagsetu[73] jsou pod

[71] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 348–349.
[72] Tamtéž, s. 350.
[73] XEROX CORPORATION. Czech Part-of-Speech Tagset. *Open Xerox* [online]. ©1999–2014 [cit. 2014-03-2020]. Dostupné z: http://open.xerox.com/Services/fst-nlp-tools/Pages/Czech%20Part-of-Speech%20Tagset.

značku „+NOUN“ zahrnuta substantiva ve všech sedmi pádech, navíc jsou zde také obsaženy zkratky, iniciály a jednotky.

Kódovník PMK je velmi podrobný systém značkování, základní kategorie postihuje kompletně. Životnost je zahrnuta v RODU, opět se zde setkáváme se čtyřmi hodnotami, *maskulinum animatum, maskulinum inanimatum, femininum* a *neutrum*. Pro ČÍSLO jsou definovány tradičně hodnoty *singuláru* a *plurálu*, avšak zcela chybí hodnota pro duál. Naopak zde pozorujeme přesah do sémantické roviny, neboť jsou definována *pluralia tantum* a *kolektiva* čili podstatná jména hromadná, která jsou řazena k singulariím tantum. Singularia tantum spolu s pluralii tantum tvoří skupinu substantiv s omezeným číselným protikladem.[74] Pojetí pádu z teorie nijak nevybočuje. U všech tří kategorií je v systému obsažena možnost *nelze určit*, která obdobně jako u pozičního tagsetu souvisí s problematičností klasifikace některých slovních tvarů.

3.2.1.2 Rozšiřující kategorie

Kromě základních gramatických kategorií mohou být u jednotlivých slovních druhů v tagsetech obsaženy i jiné kategorie, rozšiřující. V této oblasti již nepanuje mezi anotačními systémy taková shoda jako v předešlém případě.

Poziční systém vedle základních kategorií nabízí také určení SLOVNÍHO PODDRUHU zastoupené hodnotami: *substantivum, obyčejné* a *substantivum jako zkratka*. Poziční tagset nabízí na druhé pozici pouze dvě hodnoty, postrádáme zde tedy detailnější popis substantiv. Tato situace představuje nepoměr nejen v porovnání s jinými slovními druhy (zejm. ohebnými), které jsou v systému rozebrány podstatně obsáhleji (např. zájmena, číslovky ad.), ale také vzhledem k prostoru, který substantiva zaujímají v teoretických příručkách, zde MČ2, v níž podstatná jména zaujímají značnou část výkladu. Dále je u substantiv reflektována NEGACE, tedy zda se jedná o *afirmativ (bez negativní předpony „ne-“)*, či *negaci (tvar s negativní předponou „ne-“)*. Negace nijak neovlivňuje gramatickou charakteristiku daného slovního tvaru, změna se odehrává v rovině významové, můžeme proto říci, že tato kategorie je důležitější v oblasti slovotvorby než tvarosloví, které by mělo být v tagsetech reflektováno primárně. U některých podstatných jmen je v pozičním

[74] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 46.

systému označena také kategorie vidu, reflexe vidu u substantiv sice není obvyklá, ale není vyloučená. MČ2 konstatuje, že: „Verbální substantivum zachovává vidový význam.“[75] Stylistický aspekt vnáší do morfologické charakteristiky slovních tvarů 15. kategorie: VARIANTA, STYLOVÝ PŘÍZNAK APOD. Zde se však již jedná o kategorii nadstavbovou v pravém slova smyslu, morfologické teorie se dotýká jen okrajově, hodnoty náležící k této kategorii nemají oporu v MČ2.

V **atributivním systému** není uvedena žádná bližší specifikace slovního druhu kromě SPECIÁLNÍHO VZORU *půl*. Podle MČ2 se však v případě slova „půl“ jedná o číslovku, nikoli o substantivum. Brněnský systém obsahuje, pokud jde o stylistický pohled na slovní tvary, podobnou kategorii jako systém poziční, tzv. STYLISTICKÝ PŘÍZNAK TVARU, hodnoty se mírně liší od hodnot definovaných v pozičním systému, vztah k výkladu MČ je zde obdobný. Dále je v tagsetu v kategorii TYP TVARU určeno, zda se u substantiva jedná o *tvar s příklonným -s*, či nikoli. Tato kategorie reflektuje, zda je k danému slovnímu tvaru připojeno sloveso být ve druhé osobě jednotného čísla „jsi“ v podobě „-s“. MČ2 tvary tohoto typu také zmiňuje.[76] Avšak tato problematika je velmi rozsáhlá, nebudeme se jí podrobněji zabývat. Stačí, když si uvědomíme, že kategorie reflektující příklonné „-s“ souvisí s problematikou analytických tvarů a jedná se tedy o další projev vlivu této problematiky na podobu tagsetů (jiné projevy byly zmíněny v první kapitole).

U **kompaktního systému** nalezneme kromě základních kategorií nadefinované hodnoty pro základní významové členění substantiv (kategorie TYP), a to na podstatná jména *obecná* a *vlastní*. Toto rozdělení se projevuje i v MČ2, dokonce ve strukturním rozvržení výkladu o substantivech.[77]

Pokud jde o určované kategorie, je **Xerox** tagset obsahově velmi chudý. Můžeme zde nalézt pouze jednu hodnotu přinášející rozšiřující informace o substantivu, a to sice, že se jedná o proprium („+PROP“, *proper name*).

Kódovník PMK obsahuje velmi podrobné určení slovního druhu, které je reprezentováno dvěma kategoriemi, pro druh a třídu. Toto rozdělení má opodstatnění z hlediska skutečností, které reflektuje. Kategorie DRUH se zaměřuje na specifikaci

[75] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 51.
[76] Tamtéž, s. 494.
[77] Tamtéž, s. 290, 345.

tvaru po formální stránce, obsahuje hodnoty pro substantivum *běžné*, *adjektivní*, *zájmenné*, *číslovkové*, *slovesné*, *slovesné zvratné*, pro *zkratkové slovo* a *substantivum nesklonné*. Z toho, že druh reflektuje formální stránku tvaru, a z nabídnutých hodnot můžeme vyčíst, že se jedná o kategorii, která úzce souvisí s gramatickou charakteristikou daného slovního tvaru (jež by měla být v centru pozornosti tagsetu). Avšak není tomu tak u druhé z kategorií, TŘÍDY, která je naopak zaměřena na sémantickou stránku substantiv, určuje, zda je denotátem daného podstatného jména *osoba*, *živočich*, *konkrétní věc*, *abstraktní věc*, případně, zda se jedná o jméno, u něhož je takovéto určení *jiné/nejasné*. Další kategorie kódovníku určované u substantiv (valence a funkce) zasahují do syntaktické roviny. VALENCE reflektuje případné vztahy daného substantiva s dalšími jednotkami jazyka a kategorie FUNKCE odráží funkci, kterou plní daný tvar ve větě. Najdeme zde (podobně jako u systému pozičního a atributivního) kategorii přesahující do roviny lexikální, STYL. Hodnoty dostupné pro tuto kategorii byly výrazům přiřazovány na základě charakteru promluvy, respektive kontextu, v jakém se vyskytovaly. Tato kategorie tedy hodnotí tvary z hlediska stylistického, v rámci morfologického tagsetu o ní můžeme hovořit jako o kategorii nadstavbové.

3.2.2 Adjektiva

3.2.2.1 Základní kategorie

Gramatickými kategoriemi určovanými u adjektiv jsou gramatický rod, gramatické číslo, pád a stupeň. *Mluvnice češtiny 2* upozorňuje, že **rod**, **číslo** a **pád** jsou u adjektiv gramatickými prostředky reduplikačními.[78] Reduplikují gramatické prostředky dominujícího substantiva, ty jsou podrobněji rozebrány v kapitole číslo: 3.2.1.1. Specifickou kategorií u adjektiv je **stupeň**. Rozlišujeme tři stupně: *pozitiv* (*první stupeň*, základní tvar, „velký“), *komparativ* (*druhý stupeň*, „větší“) a *superlativ* (*třetí stupeň*, „největší“).[79]

V **pozičním tagsetu** jsou u adjektiv určovány všechny výše zmíněné základní gramatické kategorie. U rodu, čísla a pádu je situace obdobná jako u substantiv (viz kap. č. 3.2.1.1). V kategorii RODU jsou klasicky definovány čtyři základní hodnoty

[78] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 78.
[79] Tamtéž, s. 78–80.

(feminina, neutra, maskulina animata a maskulina inanimata). Kromě nich zde však najdeme také další hodnoty, ty jsou však pouze kombinacemi zmíněných čtyř základních hodnot a z hlediska lingvistické teorie tak postrádají význam, patří zde: *femininum singuláru nebo neutrum plurálu (pouze u příčestí a jmenných adjektiv)*; *masculinum inanimatum nebo femininum (jen plurál u příčestí a jmenných adjektiv)*; *masculinum (animatum nebo inanimatum)*. Tyto tzv. kombinační hodnoty vznikly nejspíše na základě problematického značkování homonymních tvarů v češtině, jsou tedy motivovány technicky; pro vyhledávání ovšem představují jakousi past na uživatele. Pokud jde o ČÍSLO, u adjektiv nalezneme oproti substantivům pouze jednu hodnotu navíc, *„W“ = pouze v kombinaci s jmenným rodem „Q“ (singulár pro feminina, plurál pro neutra)*, podobně jako u rodu, ani zde nemá daná hodnota z hlediska lingvistické teorie význam. Ostatní hodnoty gramatického čísla se shodují s těmi určovanými u substantiv (viz kap. č. 3.2.1.1). Pojetí PÁDU je rovněž shodné jako u substantiv. Pro kategorii STUPNĚ jsou v pozičním systému, v souladu s MČ2, k dispozici tři hodnoty: *1. stupeň*, *2. stupeň* a *3. stupeň*.

Atributivní tagset také odráží všechny základní gramatické kategorie adjektiv. Kategorie jsou obdobné jako u substantiv, na rozdíl od podstatných jmen však není v rámci adjektiv zohledňována žádná specifická skupina. U kategorií rodu, čísla a pádu tedy pozorujeme výlučně elementární hodnoty. ROD zahrnuje i životnost: *rod mužský životný*, *rod mužský neživotný*, *rod střední*, *rod ženský*; u ČÍSLA najdeme hodnotu pro *číslo jednotné*, *množné* i *duál*, definováno je všech sedm PÁDŮ. Atributivní tagset také rozlišuje tradiční tři STUPNĚ adjektiv: *pozitiv*, *komparativ* a *superlativ*.

Všechny základní kategorie adjektiv nalezneme také v **tagsetu kompaktním**. Hodnoty u RODU, ČÍSLA, PÁDU a zvlášť vydělené kategorie pro ŽIVOTNOST se shodují s hodnotami určovanými v kompaktním systému u substantiv, viz kap. č. 3.2.1.1. Navíc je zde ve shodě s MČ2 přítomná kategorie STUPNĚ, která zahrnuje hodnoty pro *pozitiv*, *komparativ* a *superlativ*.

Xerox tagset, podobně jako u substantiv, ani v případě adjektiv nereflektuje žádnou ze základních gramatických kategorií uvedených v MČ2. Disponuje pouze určením

slovního druhu „+ADJ“ (*adjective*, adjektivum). V rozporu s MČ2[80] jsou mezi přídavná jména zahrnuty také číslovky řadové (*ordinal*), které zde byly zařazeny pravděpodobně na základě velké formální podobnosti s adjektivy. Ačkoli je pravda, že z druhů číslovek se po formální stránce s adjektivy kryjí číslovky řadové nejvíce, nejedná se o jedinou třídu číslovek, u které je možné pozorovat adjektivní skloňování. Tvary adjektivní flexe nalezneme i u jiných druhů číslovek, např. u číslovek druhových, velikostních či násobných.[81] Z formálního hlediska je do jisté míry pochopitelné, že se tvůrci tagsetu rozhodli zařadit řadové číslovky mezi adjektiva, ovšem nelze pominout, že se jedná o hrubý a nesystematický zásah do základního pojetí slovních druhů. Zároveň také nelze z počínání tvůrců tagsetu vyvodit snahu o sjednocení tvarů s adjektivní flexí, uvážíme-li, že jiné druhy číslovek, u kterých se také projevuje adjektivní flexe, do adjektiv zahrnuty nebyly. Musíme konstatovat, že za těchto okolností je počínání tvůrců Xerox tagsetu velmi nesystematické.

V **kódovníku PMK** jsou postiženy všechny základní gramatické kategorie (rod, číslo, pád a stupeň). ROD je pojat stejně, jako je tomu u substantiv, viz kap. č. 3.2.1.1. ČÍSLO disponuje pouze hodnotami pro *singulár* a *plurál*, chybí určení duálu. Pojetí kategorie pádu je tradiční. U všech třech kategorií je definována hodnota *nelze určit* (obdobně jako u substantiv viz 3.2.1.1). Nenajdeme ji však u kategorie STUPNĚ, kde kódovník nabízí pouze tři možnosti, stupeň *první*, *druhý* nebo *třetí*.

3.2.2.2 Rozšiřující kategorie

Kategorie SLOVNÍ PODDRUH (na 2. pozici) je v **pozičním tagsetu** u adjektiv obsáhlejší než u substantiv, obsahuje osm hodnot, přičemž je svým obsahem velmi různorodá. U substantiv, kde je na druhé pozici možnost si vybrat pouze ze dvou hodnot, není tato situace tak patrná, proto se jí podrobněji věnujeme až zde. Tato kategorie se právě díky své různorodosti jeví jako velmi problematická. Hodnoty zde definované neodrážejí jedno kritérium, jako je tomu např. u rodu, čísla apod., nýbrž jsou výsledkem působení faktorů různé povahy. Z hlediska formálního lze adjektiva rozdělit na tvary dlouhé (složené) a krátké (jmenné), jako třetí skupina jsou vydělena adjektiva přivlastňovací (posesiva). Jmenná a přivlastňovací adjektiva mají vlastní

[80] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 123–124.
[81] Tamtéž, s. 403.

hodnoty: *adjektivum, jmenný tvar* a *adjektivum přivlastňovací (na „-ův“ i „-in“)*. Dlouhé tvary adjektiv vlastní hodnotu nemají. Pro tvary složené lze v tagsetu nalézt dokonce tři hodnoty: *adjektivum obyčejné*, *přídavné jméno odvozené od slovesného tvaru přítomného přech.* a *přídavné jméno odvozené od slovesného tvaru minulého přech*. Jak je patrné ze zmíněných hodnot, jsou u dlouhých tvarů zvlášť vydělena adjektiva verbální, která jsou výsledkem slovnědruhové transpozice.[82] Označení verbálních adjektiv však není provedeno důsledně. Jsou zohledněna pouze adjektiva vzniklá z přechodníků a nikoli z příčestí, a tudíž nejsou reflektovány všechny druhy verbálních adjektiv. Dále je také určení, zda jde o verbální adjektivum, k dispozici pouze v souvislosti se složenými tvary adjektiv a nikoli s tvary jmennými (které jsou všechny sdružené pod jedinou hodnotu – adjektivum jmenný tvar), i když verbální adjektiva „disponují v závislosti na syntaktické funkci formou krátkou i dlouhou“[83]. Podobně jako u substantiv (viz kap. č. 3.2.1.2) jsou i zde vyděleny zkratky – *adjektivum jako zkratka*. V rozporu s teorií jsou do adjektiv zařazena *samostatně stojící zájmena „svůj“, „nesvůj“, „tentam“*. Naprosto nelingvistickou hodnotou potom zůstává *slovo před pomlčkou*, pro něž není z hlediska lingvistické teorie opodstatnění. Po prozkoumání některých tvarů, kterým byla tato hodnota přiřazena, bylo zjištěno, že se mnohdy nejedná o adjektivum, ba dokonce se daný tvar ani nenachází před pomlčkou.

Další kategorií pozičního tagsetu určovanou u adjektiv je PŘIVLASTŇOVACÍ ROD, který se vztahuje k substantivnímu základu posesiv. Substantivní základy podle MČ2 „označují jedince (osoby, méně často zvířata) rodu mužského a ženského, nikoli středního.“[84] Tomu odpovídají i hodnoty pro přivlastňovací rod definované: *femininum (ženský rod)* a *maskulinum animatum (rod mužský životný)*. Dále je u adjektiv reflektována NEGACE (viz kap. č. 3.2.1.2). U některých adjektiv je definována i kategorie VIDU, její zohlednění není v rozporu s teorií, která připouští, že verbální adjektiva zachovávají významy motivujících sloves.[85] Je zde přítomná i kategorie VARIANTA, STYLOVÝ PŘÍZNAK APOD. spadající spíše do roviny lexikální s důrazem na stylistické faktory, viz kap. č. 3.2.1.2.

[82] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 141.
[83] Tamtéž, s. 143.
[84] Tamtéž, s. 77.
[85] Tamtéž, s. 146.

V **atributivním systému** najdeme kromě základních gramatických kategorií určení NEGACE, která je zastoupená dvěma hodnotami: *afirmace* a *negace*. Negace se týká spíše roviny významové, jak bylo řečeno u substantiv v systému pozičním (viz kap. č. 3.2.1.2). Posledními kategoriemi určovanými u adjektiv v brněnském tagsetu jsou STYLISTICKÝ PŘÍZNAK TVARU a TYP TVARU – více taktéž v kap. č. 3.2.1.2.

Kompaktní tagset rovněž disponuje několika rozšiřujícími kategoriemi, jsou jimi typ a tvar. Tyto dvě kategorie spolu souvisejí, jelikož se v obou projevuje hledisko formální. TYP rozděluje adjektiva na *přivlastňovací* a *kvalifikující*. Kategorie TVAR potom specifikuje, zda se jedná o tvar *jmenný* či *složený*. Ve zmíněných dvou kategoriích se tak promítá formální hledisko přítomné i v MČ2.

Xerox tagset neobsahuje, pokud jde o adjektiva, žádné rozšiřující gramatické kategorie.

Zato **kódovník PMK** v sobě zahrnuje rozšiřujících kategorií hned několik. Pro bližší specifikaci slovního druhu jsou zde definovány tři kategorie druh, poddruh a třída. Kategorie DRUH určuje, zda se jedná o adjektivum *nespecifické*, *slovesné* či *přivlastňovací*. PODDRUH se vztahuje ke kategorii předchozí, podrobněji určuje slovní druh. Dochází zde k určení jmenných tvarů, pro které jsou definovány tři hodnoty: *jmenná forma, singulár neutra*; *jmenná forma jiná* a *zvratné – jmenná forma*. Zvratnost je reflektována i u adjektiv verbálních hodnotami pro adjektivum: *deparcipiální prosté* a *zvratné*. Třetí zmíněná kategorie, TŘÍDA, je založena na sémantické stránce adjektiv a vyděluje tyto hodnoty: adjektivum *deskriptivní*, *deskriptivní propriální*, *evaluativní*, *intenzifikační*, *restriktivní* a adjektivum, které *nelze určit*. Zohledňování významové stránky však není v tagsetu prioritou. I zde jsou podobně jako u substantiv definovány kategorie spjaté se syntaktickou rovinou jazyka: VALENCE a FUNKCE. Také zde najdeme kategorii STYL, viz kap. č. 3.2.1.2.

3.2.3 Zájmena

3.2.3.1 Základní kategorie

Mezi základní gramatické kategorie zájmen patří, obdobně jako u substantiv či adjektiv, **rod**, **číslo** a **pád**. Zároveň však MČ poukazuje na fakt, že zájmena vykazují v rámci základních kategorií četné modifikace, specifické vlastnosti,

související s funkční sémantikou zájmen.[86] Příkladem této modifikace může být kategorie rodu u zájmen bezrodých, která nedisponují diferenčními prostředky pro rozlišení mluvnického rodu,[87] a rod u nich tedy není určován.

V **pozičním tagsetu** jsou reflektovány všechny základní gramatické kategorie zájmen. Pro ROD jsou kromě čtyř tradičních hodnot definovány také jejich kombinace, jak bylo řečeno výše u adjektiv, tyto kombinované hodnoty postrádají z hlediska lingvistické teorie význam, patří zde hodnoty: *femininum nebo neutrum (tedy nikoli maskulinum), masculinum (animatum nebo inanimatum)* a *‚nikoli femininum' (tj. M/I/N; především u zájmen)*. Přítomna je taktéž hodnota *libovolný rod*, viz kap. č. 3.2.1.1. Hodnoty pro gramatické ČÍSLO a PÁD se shodují s hodnotami definovanými u substantiv, viz výše kap. č. 3.2.1.1.

Atributivní systém rovněž odráží výše zmíněné základní gramatické kategorie zájmen. Hodnoty pro tyto kategorie (ROD, ČÍSLO, PÁD) definované se shodují s hodnotami určovanými v atributivním systému u adjektiv. Viz kap. č. 3.2.2.1.

Kompaktní tagset také zahrnuje všechny tyto kategorie. Hodnoty definované pro ROD, ČÍSLO, PÁD a zvlášť vydělenou kategorii ŽIVOTNOSTI se shodují s hodnotami pozorovanými u substantiv a adjektiv. Viz kap. č. 3.2.1.1.

V **Xerox tagsetu** je zohledněna jediná základní gramatická kategorie, a to kategorie PÁDU. Obsahuje hodnoty: *pronoun: accusative*; *pronoun: dative*; *pronoun: genitive*; *pronoun: instrumental*; *pronoun: locative*; *pronoun: nominative* a *pronoun vocative*. Řazení v seznamu značek je poněkud netradiční (řazeno abecedně podle počátečních písmen anglických názvů pádů), ale je zde přítomno všech sedm pádů.

Kódovník PMK opět postihuje všechny základní kategorie, rod, číslo i pád. ROD obsahuje kromě čtyř tradičních hodnot i speciální hodnotu pro zájmeno *bezrodé*. MČ2 o zájmenech bezrodých hovoří jako o specifické skupině zájmen nedisponující diferenčními prostředky pro lišení rodu, tedy jako o zájmenech rodově neutrálních.[88] Nevytváří však pro ně v rámci rodu specifickou hodnotu. U ČÍSLA jsou k dispozici hodnoty pro *singulár* a *plurál*, duál není definován, tím se situace v kódovníku nijak

[86] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 84.
[87] Tamtéž, s. 101.
[88] Tamtéž, s. 101.

neliší od adjektiv. Avšak pokud jde o kategorii čísla, je u zájmen zohledňován *plurál v platnosti singuláru (vykání)*. Tato hodnota má oporu v MČ2, která uvádí: „Zájmena *my, vy* jsou někdy používána – vlastně v rozporu se svým plurálovým významem – k označení jednotlivců. Kodifikováno je tzv. vykání, tj. používání plurálového výrazu *vy* pro komunikaci s jednotlivým adresátem ze zdvořilostních důvodů.“[89] Nicméně se stále jedná spíše o zohlednění sémantické stránky daných zájmen, nikoli formální. Pojetí PÁDU je tradiční, viz kap. č. 3.2.1.1. U všech tří kategorií je definována hodnota *nelze určit* (viz kap. č. 3.2.1.1).

3.2.3.2 Rozšiřující kategorie

Na úvod dalšího rozboru kategorií určovaných v morfologických systémech značkování u zájmen je třeba upozornit na specifické pojetí pronomin v MČ2. Ve výkladu jsou k zájmenům řazena také zájmenná příslovce a zájmenné číslovky, MČ2 je souhrnně nazývá deiktická slova neboli pronominalia.[90] Považujeme za důležité zmínit, že studované tagsety se v tomto ohledu s teorií uvedenou v MČ2 rozcházejí. Zaměřují se výhradně na pronomina. Jelikož jsou zájmena velmi různorodou skupinou slov, je nasnadě se zmínit také o různých třídách, které se v jejich rámci vyskytují. Co se týče klasifikace zájmen z hlediska funkčně-sémantického, MČ2 rozděluje deiktická slova do několika skupin sdružujících se v několika okruzích. V užším smyslu uvádí MČ2 toto dělení: osobní zájmena; zvratné zájmeno osobní; přivlastňovací zájmena osobní; přivlastňovací zájmeno zvratné; demonstrativa – ukazovací zájmena, příslovce a číslovky; vlastní identifikátory; interogativa (výrazy tázací) a relativa (výrazy vztažné); indefinita; totalizátory; negativa (negátory).[91] Jedná se o velmi podrobné třídění, v tagsetech se často shledáme s rozličnými případy nakládání s těmito třídami. Stává se, že je v tagsetu i více ze zmíněných tříd zahrnuto do jedné hodnoty, nebo naopak je v rámci jedné třídy pro tagset nadefinováno více hodnot. Tento druhý případ je patrný především u systému pozičního.

Druhá pozice v **pozičním tagsetu** je, co se týče zájmen, velmi obsáhlá. Kategorie SLOVNÍ PODDRUH zde zahrnuje dvacet hodnot. Pro zájmena osobní jsou v tagsetu nadefinovány čtyři hodnoty: *osobní zájmena (vč. tvaru „tys“)*; *krátké tvary osobních*

[89] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 102.
[90] Tamtéž, s. 81–104.
[91] Tamtéž, s. 88–99.

zájmen („mě“, „mi“, „ti“, „mu“); *zájmeno „on“ ve tvarech po předložce (tj. „n-“: „něj“, „něho“, ...)* a *předložka s připojeným „-ň“ (něj), „proň“, „naň“, atd. (značkováno jako slovní druh: zájmeno – P)*. Zvlášť jsou vyděleny tvary osobních zájmen následujících po předložkách, poziční systém tak odráží existenci specifických tvarových souborů pro bezpředložkové a předložkové pozice.[92] Poziční systém je tak podrobný, že reflektuje dokonce i silně knižní tvar zájmena on „-ň“ nacházející se vždy ve spojení s předložkou.[93] Tvůrci upozorňují, že vzniklý slovní tvar, spojení předložky a zájmena, je označkován jako zájmeno, nikoli předložka. Hojná tvarová variantnost také vybízí k reflexi souborů delších a kratších tvarů, ke které v pozičním systému také došlo, jak je patrné z výše vypsaných hodnot. Delší a kratší tvary najdeme také u zájmen zvratných reprezentovaných dvěma hodnotami: *reflexívní zájmeno „se“, „si“ pouze v těchto tvarech* a dále *„ses“, „sis“ a reflexívní zájmeno „se“ v dlouhých tvarech*. Dvě hodnoty jsou přítomny také u zájmen přivlastňovacích: *zájmeno přivlastňovací „můj“, „tvůj“, „jeho“ (vč. plurálu)* a *přivlastňovací zájmeno „svůj“*. Výjimečně jedinou hodnotou jsou reprezentována zájmena ukazovací: *zájmeno ukazovací („ten“, „onen“, ...)*. Zájmena vztažná a tázací jsou, v důsledku toho, že disponují výrazově totožnými prostředky,[94] propojena v rámci čtyř hodnot: *vztažné nebo tázací zájmeno s adjektivním skloňováním (obou typů: „jaký“, „který“, „čí“, ...)*; *zájmeno tázací nebo vztažné „kdo“, vč. tvarů s „-ž“ a „-s“*; *zájmeno tázací/vztažné „co“, „copak“, „cožpak“*; *zájmeno „co“ spojené s předložkou („oč“, „nač“, „zač“)*. I v této skupině zájmen můžeme pozorovat propojení zájmena a předložky, opět je vzniklý tvar řazen mezi zájmena. Vztažná a tázací zájmena, však nejsou v tagsetu značena pouze propojeně, zájmena vztažná mají také své vlastní hodnoty, avšak zájmena tázací již nikoli. Vztažná zájmena najdeme samostatně ve čtyřech hodnotách: *vztažné zájmeno „což“*; *vztažné přivlastňovací zájmeno „jehož“, „jejíž“, ...*; *vztažné zájmeno „jenž“ („již“, ...) bez předložky*; *vztažné zájmeno „jenž“, „již“, ... po předložce („n-“: „něhož“, „níž“, ...)*. Další reflektovanou skupinou jsou zájmena neurčitá, pro která zde nalezneme dvě hodnoty: *zájmeno neurčité („nějaký“, „některý“, „čikoli“, „cosi“, ...)* a *zájmeno neurčité „všechen“, „sám“*. Upozorněme pouze, že zájmena „všechen“ a „sám“ jsou v MČ2 řazena mezi totalizátory, které jsou dále zahrnovány

[92] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 396.
[93] Tamtéž, s. 397.
[94] Tamtéž, s. 95.

mezi deiktická slova určitá (totalizátory a negativa). MČ2 charakterizuje tyto skupiny deiktických slov následovně: „Jednotky obou těchto tříd označují krajní množství (totálnost nebo nulovost) toho, k čemu ukazují: živých bytostí, věcí atd.“[95] Toto označení krajního množství je pro MČ2 ohledně rozlišování určitosti a neurčitosti evidentně rozhodující. Poslední hodnotou definovanou pro zájmena na druhé pozici jsou *zájmena záporná („nic“, „nikdo“, „nijaký“, „žádný“, ...)*. V MČ2 je najdeme pod označením negativa (negátory).[96]

V rámci pozičního systému dále nalezneme kategorie blíže specifikující vlastníka, a to přivlastňovací rod a přivlastňovací číslo. S pojmem vlastník se u zájmen setkáváme právě v MČ2, která s tímto pojmem operuje např. u zájmen přivlastňovacích, jež: „identifikují v podmínkách konkrétní komunikace buď mluvčího, nebo adresáta, nebo předmět komunikačního aktu jako vlastníka objektu, k němuž ukazují.“[97] PŘIVLASTŇOVACÍ ROD disponuje dvěma hodnotami, jednou tradiční *femininum (ženský rod)* a jednou kombinující v sobě zbylé hodnoty tradičně u mluvnického rodu určované: *‚nikoli femininum‘ (tj. M/I/N; u přivlastňovacích adjektiv)*. PŘIVLASTŇOVACÍ ČÍSLO disponuje hodnotami pro *singulár (jednotné číslo)* a *plurál (množné číslo)*. Určována je zde také kategorie OSOBY s hodnotami: *1. osoba, 2. osoba* a *3. osoba*. To je v souladu s teorií. MČ2 se zmiňuje o kategorii osoby ve výkladu o zájmenech osobních, kde uvádí, že rozlišují „v zásadě tři komunikační osoby: 1. osoba – mluvčí; 2. osoba – adresát; 3. osoba – předmět (téma) komunikace.“[98]

U pronomin se dále objevují i kategorie čas, negace a slovesný rod (kategorie pojmenovaná aktivum/pasívum). U zájmen je v každé z těchto kategorií v nabídce pouze jedna hodnota, *prézens (přítomný čas)* pro kategorii ČASU, *afirmativ (bez negativní předpony „ne-“)* pro NEGACI a *aktivum nebo „nikoli pasívum“* pro AKTIVUM/PASÍVUM. Po zadání tagu obsahujícího pro zájmeno tyto tři hodnoty do korpusového manažeru zjistíme, že jsou tímto způsobem označeny pouze tvary „tys“, tedy spojení zájmena „ty“ se slovesem „jsi“, které je ve spojení se zájmenem (případně jiným slovním druhem) v podobě klitického (příklonného) „-s“ vždy ve

[95] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 98.
[96] Tamtéž, s. 98.
[97] Tamtéž, s. 91.
[98] Tamtéž, s. 89.

tvaru prézentu aktiva, netřeba zdůrazňovat, že se jedná o afirmativ. Poslední kategorií je VARIANTA, STYLOVÝ PŘÍZNAK APOD., více viz kap. č. 3.2.1.2.

Atributivní systém také reflektuje funkčně-sémantické třídy zájmen. Najdeme zde dvě kategorie pro DRUH ZÁJMENA, obsahují hodnoty pro zájmeno: *osobní*, *přivlastňovací*, *ukazovací*, *vymezovací*, *reflexivní*, *tázací*, *vztažné*, *záporné* a *neurčité*. Vidíme, že v podstatě odpovídají třídám prezentovaným v MČ2, ovšem s tím rozdílem, že hodnoty tagsetu nejsou tak podrobné. Dále je zde obsažena kategorie OSOBY, která souvisí s třídou zájmen osobních. Kromě tradičních hodnot *první osoba*, *druhá osoba* a *třetí osoba*, je zde také hodnota pro kombinaci všech tří osob (*první nebo druhá nebo třetí*), která je v atributivním systému ojedinělá. Najdeme zde také kategorii pro STYLISTICKÝ PŘÍZNAK TVARU a TYP TVARU, viz kap. č. 3.2.1.2.

I **kompaktní tagset** obsahuje bližší specifikaci zájmen v podobě kategorie TYP, v níž se nachází rovněž dělení zájmen do tříd funkčně-sémantických. Najdeme zde označení pro zájmeno: *osobní*, *ukazovací*, *neurčité*, *přivlastňov.*, *tázací*, *vztažné*, *reflex. ‚se'*, *záporné* a *totální*. Kromě tohoto druhu dělení kompaktní systém obsahuje speciální kategorii pro určení SYNTAKTICKÉHO TYPU zájmen: *nominální*, *adjektivní*. Tato kategorie souvisí s existencí zájmen subjektivních a adjektivních, které se dle MČ2 odrážejí jak ve formální stránce zájmen, tak v jejich funkci, kterou plní ve větě.[99] Další kategorií, je REFERENČNÍ TYP zahrnující typ *osobní* a *přivlastňovací*, tato kategorie blíže specifikuje subjekt reality, o němž zájmena referují. Referenční typ úzce souvisí se skupinou tříd zájmen představenou v MČ2 pod názvem: třídy deiktických slov spjaté vztahem ke komunikačním „osobám".[100] Taktéž určení OSOBY zde nechybí, tradičně zde najdeme rozdělení na tři osoby: *první*, *druhou* a *třetí*. Podobně jako u systému pozičního (viz výše) i zde najdeme kategorie, jež blíže specifikují vlastníka, pro ROD VLASTNÍKA jsou zde definovány tři hodnoty: *mužský*, *ženský* a *střední*, pro ČÍSLO VLASTNÍKA *singulár* a *plurál*. Kategorií speciálně určenou pro zájmena představuje také KLITIKA čili určení, zda jde o zájmeno *neklitické*, či *klitické*. Souvisí s další specifickou vlastností zájmen, tentokrát v oblasti slovosledu, ohledně kterého nás MČ2 upozorňuje právě na zvláštní postavení tzv. příklonných (klitických) tvarů zájmen.[101] Klitika souvisí

[99] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 84.
[100] Tamtéž, s. 89–92.
[101] Tamtéž, s. 87.

rovněž s poslední kategorií kompaktního systému věnovanou zájmenům, a to KLIT. „S“, zde je upřesněno, zda je k zájmenu připojeno sloveso „být“ ve druhé osobě jednotného čísla ve své klitické podobě „-s“, či nikoliv. O klitickém „-s“ viz výše (3.2.1.2).

Jediná rozšiřující kategorie zájmen v **Xerox tagsetu** souvisí s rozdělením pronomin do funkčně-sémantických tříd. Jedinou definovanou hodnotou je *possesive* [správně possessive] *pronoun „se“* (zn. +PSE). Zájmeno „se“ je zde představeno jako zájmeno přivlastňovací (possessive), i když se jedná o zájmeno zvratné. Je pravda, že MČ2 rozděluje zájmena zvratná na zvratná osobní a zvratná přivlastňovací,[102] avšak „se“ patří právě do skupiny zvratných osobních.

Kódovník PMK taktéž obsahuje třídění zájmen z hlediska funkčně-sémantického. V kategorii DRUH jsou definovány hodnoty: *osobní*, *zvratné osobní*, *přivlastňovací*, *zvratné přivlastňovací*, *neurčité*, *ukazovací*, *tázací*, *vztažné* a *záporné*. I když ani kódovník neodráží klasifikaci druhů zájmen v naprostém souladu s MČ2, jako jediný odlišuje zájmena zvratná přivlastňovací a osobní. V kategorii druh je kromě výše zmíněných tříd zájmen reflektována také víceslovnost, najdeme zde hodnoty pro zájmeno *víceslovné* a *víceslovné vztažné*. MČ2 s pojmem zájmeno víceslovné neoperuje. Obdobně jako u substantiv (či adjektiv) i zde jsou určovány kategorie VALENCE, FUNKCE a STYL, které však už sahají za hranice morfologické roviny jazyka, více viz výše (3.2.1.2).

3.2.4 Číslovky

3.2.4.1 Základní kategorie

Mezi základní gramatické kategorie určované u číslovek patří podobně jako u substantiv **rod**, **číslo** a **pád**, viz kap. č. 3.2.1.1. Avšak MČ2 zároveň upozorňuje, že jsou u číslovek tyto gramatické kategorie zastoupeny různým způsobem: „Některé číslovky mají v plné míře všechny tři gramatické prostředky. […] Jiné číslovky některé gramatické kategorie vyjadřují omezeně […] Některé číslovky mají pádové tvary, skloňují se, nevyjadřují však význam rodu a čísla.“[103]

[102] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 90–92.
[103] Tamtéž, s. 125.

V **pozičním systému** jsou pro číslovky definovány všechny tři základní gramatické kategorie. Co se týče hodnot v rámci těchto kategorií obsažených, shodují se s hodnotami přítomnými v pozičním tagsetu u zájmen, viz kap. č. 3.2.3.1.

Atributivní systém rovněž odráží všechny zmíněné gramatické kategorie číslovek. Kategorie disponují hodnotami, jež byly popsány v témže systému u adjektiv, viz kap. č. 3.2.2.1.

V **systému kompaktním** jsou také zahrnuty všechny tyto kategorie. Hodnoty definované pro ROD, ČÍSLO, PÁD a zvlášť vydělenou kategorii ŽIVOTNOSTI se shodují s hodnotami pozorovanými u substantiv, adjektiv a zájmen. Viz kap. č. 3.2.1.1.

Xerox tagset podobně jako u zájmen i u číslovek zohledňuje jedinou základní kategorii, a to PÁD. Obsahuje hodnoty: *numeral: accusative*; *numeral: dative*; *numeral: genitive*; *numeral: instrumental*; *numeral: locative*; *numeral: nominative* a *numeral vocative*, tedy všech sedm pádů.

V **kódovníku PMK** jsou opět postiženy všechny základní kategorie, rod, číslo i pád. Hodnoty pro ROD jsou totožné s těmi definovanými u zájmen, viz kap. č. 3.2.3.1. U kategorie ČÍSLA jsou však definovány pouze hodnoty pro *singulár* a *plurál*. Pojetí PÁDU je tradiční, viz kap. č. 3.2.1.1. U všech tří kategorií je definována hodnota *nelze určit* (viz kap. č. 3.2.1.1).

3.2.4.2 Rozšiřující kategorie

MČ2 rozděluje číslovky do několika významových skupin, jádrem jsou číslovky základní, dále se vydělují číslovky numerické, úhrnné, souborové, druhové, velikostní, násobné, dílové, skupinové a řadové. V rámci těchto skupin MČ2 ještě rozlišuje číslovky určité, neurčité a úplnostní. Jednotlivé tagsety se značně liší v míře, s jakou zmíněné třídy číslovek rozlišují.

Poziční tagset disponuje na druhé pozici (SLOVNÍ PODDRUH) opět značným množstvím hodnot. Nalezneme jich celkem osmnáct a převažuje situace, kdy je pro jednu významovou skupinu číslovek definováno více hodnot.
Z významových druhů číslovek jsou reflektovány číslovky základní, druhové, řadové, násobné a dílové, zvlášť jsou vyděleny číslovky tázací, neurčité a mnohé specifické příklady číslovek. Pro číslovky základní jsou definovány dvě hodnoty:

číslovky základní 1–4, „půl“, ...; sto a tisíc v nesubst. *skloňování* a *číslovky základní >=5*. Následují číslovky druhové se čtyřmi hodnotami: *číslovka druhová, adjektivní skloňování („jedny“, „dvojí“, „desaterý“, ...)*; *číslovky druhové „jedny“ a „nejedny“*; *číslovka druhová >=4 substantivní postavení („čtvero“, „desatero“, ...)* a *číslovka druhová >=4 adjektivní postavení krátký tvar („čtvery“, ...)*. Číslovky řadové, násobné a dílové mají po jedné hodnotě, a to: *číslovky řadové*; *číslovky násobné („-krát“: „pětkrát“, „poprvé“ ...)* a *zlomky zakončené na „-ina“ (značkováno jako slovní druh číslovka – ‚C‘)*. Zvlášť jsou vyděleny číslovky tázací: *číslovka tázací násobná „kolikrát“*; *číslovka tázací řadová „kolikátý“* a *číslovka „kolik“*, pro číslovky druhové tázací hodnota vydělena není. Číslovky neurčité jsou rovněž reflektovány: *číslovka neurčitá („mnoho“, „málo“, „tolik“, „několik“, „kdovíkolik“, ...)*; *číslovky násobné neurčité („-krát“: „mnohokrát“, „tolikrát“, ...)* a *číslovky neurčité s adjektivním skloňováním („nejeden“, „tolikátý“, „několikátý“, ...)*. Mezi specifické hodnoty patří: *číslovky psané číslicemi (značkováno jako slovní druh: číslovka – ‚C‘)*; *číslovka psaná římskými číslicemi* a *zkratka jako číslovka*. Poslední kategorií určovanou u číslovek je VARIANTA, STYLOVÝ PŘÍZNAK APOD., více viz kap. č. 3.2.1.2.

V **atributivním tagsetu** jsou pro DRUH ČÍSLOVKY k dispozici dvě kategorie, v první nalezneme hodnoty pro číslovku *základní*, *řadovou* a *druhovou*, dále pak dvě blíže nespecifikované hodnoty s názvem *gramatika*. Druhá kategorie pojmenovaná druh číslovky obsahuje hodnoty pro číslovku *zápornou* a *neurčitou*. U číslovek jsou v atributivním systému dále určovány již představené kategorie STYLISTICKÝ PŘÍZNAK TVARU a TYP TVARU, viz kap. č. 3.2.1.2. U číslovek je jako u jediného slovního druhu v atributivním systému nabídnuta kategorie TERMINÁL V GRAMATICE.

Kompaktní tagset operuje s bližším určením druhu číslovky ve dvou kategoriích. První z nich je TYP s hodnotami pro číslovky: *základní*, *řadové*, *násobné* a *druhové*. Druhou klasifikující kategorií je TŘÍDA s hodnotami číslovka: *určitá1*, *určitá2*, *určitá34*, *ukazovací*, *neurčitá*, *tázací* a *vztažná*. Kompaktní tagset neobsahuje tak podrobné určení číslovek jako MČ2. Vydělením hodnot pro číslovky určité a neurčité umožňuje postihnout určitost a neurčitost u všech významových druhů číslovek prezentovaných v kategorii typ. Problematika významového třídění číslovek je tedy zpracována systematicky a přehledně, v souladu s obecným konceptem

tagsetu. Ukazovací číslovky, kterým se kompaktní tagset v kategorii třída věnuje, jsou ve výkladu MČ2 zařazeny mezi deiktická slova (po vzoru zájmen a proadverbií). Obdobně je tomu u číslovek tázacích a vztažných. Kompaktní systém také přináší klasifikaci číslovek z hlediska formálního v kategorii FORMA, kde rozlišuje: *číslo arab.*, *číslo řím.* a *slovo*.

V **Xerox tagsetu** jsou v rámci číslovek zvlášť vyděleny pouze číslice (*digit*, +NUM_INS).

Kódovník PMK obsahuje několik rozšiřujících kategorií. V kategorii DRUH je vedle významových druhů číslovek brána v potaz také víceslovnost některých číslovek. To souvisí s významnou specifickou vlastností číslovek, jak uvádí MČ2: „většina vyšších číselných hodnot se označuje kombinací číslovkových výrazů […] Taková pojmenování se nazývají kombinované číslovkové výrazy.“[104] Kategorie obsahuje hodnoty: *základní (určitá jednoslovná)*; *řadová (určitá jednoslovná)*; *druhová (určitá jednoslovná)*; *násobná (určitá jednoslovná)*; *neurčitá, i zájmenná*; *víceslovná základní*; *víceslovná řadová* a *víceslovná neurčitá*. Další kategorií pro číslovky specifickou je PÁD POČÍTANÉHO PŘEDMĚTU (tradičně sedm pádů a hodnota pro situaci, kdy pád *nelze určit*), tato kategorie již zasahuje do syntaxe. Podobně také zbývající dvě kategorie: VALENCE a FUNKCE. Ani zde nechybí kategorie reflektující stylistický příznak tvaru, STYL.

3.2.5 Slovesa

3.2.5.1 Základní kategorie

MČ2 připisuje slovesům pět základních gramatických kategorií: osobu, číslo, čas, způsob, slovesný rod a vid. Rozlišujeme tři **slovesné osoby**: *první*, *druhou* a *třetí*. Kategorie osoby funguje v rámci širší kategorie personálnosti, která vychází ze vztahu komunikačního aktu a komunikačních osob, a to: mluvčího (první osoba), adresáta (druhá osoba) a nepartnerské osoby, kterou může reprezentovat živá bytost či věc (třetí osoba). U sloves nalezneme dvojí mluvnické **číslo** *jednotné* (*singulár*) a *množné* (*plurál*). Tradičně rozlišujeme tři základní slovesné **časy**: *přítomný* (*prézens*), *minulý* (*préteritum*) a *budoucí* (*futurum*). U **slovesného způsobu** MČ2

[104] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 107.

uvádí tři hodnoty: *indikativ* (*oznamovací způsob*), *kondicionál* (*podmiňovací způsob*) a *imperativ* (*rozkazovací způsob*). V rámci **slovesného rodu** rozlišujeme dvě hodnoty: *aktivum* (*rod činný*) a *pasivum* (*rod trpný*). Z hlediska **vidu** dělíme slovesa na: *dokonavá*, *nedokonavá* a *obouvidová*. MČ2 u kategorie vidu poznamenává, že tato kategorie stojí na pomezí mluvnice a slovníku a někdy se označuje jako kategorie lexikálně-gramatická či gramaticko-lexikální.[105]

Slovesa jsou z hlediska základních kategorií velmi problematickým slovním druhem. Tagsety jsou vybaveny kategoriemi, které odrážejí gramatické kategorie daného slovního druhu. V případě sloves se však málokdy kryjí kategorie tagsetu se základními gramatickými kategoriemi u sloves určovanými. Často dochází k situaci, kdy je v rámci tagsetu přítomna kategorie, která neodráží striktně jednu základní gramatickou kategorii slovesa, nýbrž kombinuje více kritérií (např. v sobě spojuje čas a způsob). Případně může nastat situace ještě složitější.

V **pozičním tagsetu** jsou určovány všechny základní gramatické kategorie kromě slovesného způsobu. Situace ohledně určování slovesného způsobu je poněkud problematická, neboť v rámci tagsetu sice nedisponuje vlastní kategorií, ale je alespoň částečně reflektován na druhé pozici, tato bude vzhledem ke svému charakteru podrobněji rozebrána dále. Pro kategorii ČÍSLA jsou v tagsetu definovány hodnoty *plurál (množné číslo)* a *singulár (jednotné číslo)*. Dále je zde také hodnota *„W“ = pouze v kombinaci se jmenným rodem „Q“ (singulár pro feminina, plurál pro neutra)*, tato kombinující hodnota nemá z hlediska lingvistického opodstatnění. Určována je zde také kategorie OSOBY s hodnotami: *1. osoba*, *2. osoba* a *3. osoba*. Pro čas jsou zde definovány tradiční hodnoty: *futurum (budoucí čas)*, *prézens (přítomný čas)* a *minulý čas*. U osoby i času jsou přítomny také hodnoty kombinující hodnoty základní, *libovolná osoba (1/2/3)* pro osobu a *libovolný čas (F/R/P)* a *minulost nebo přítomnost (P/R)* pro kategorii času. Tyto hodnoty také vznikly spíše z technických důvodů, situace je zde podobná jako u tzv. kombinačních hodnot u adjektiv (viz kap. č. 3.2.2.1). Kategorie slovesného rodu pojmenovaná AKTIVUM/PASÍVUM rozlišuje dvě hodnoty: *aktivum nebo ‚nikoli pasívum‘* a *pasívum*. V kategorii VIDU nalezneme hodnoty: *perfektivum (dokonavé sloveso)*, *imperfektivum (nedokonavé sloveso)* a *obouvidé sloveso*.

[105] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 156–187.

Atributivní tagset disponuje vlastními kategoriemi pro osobu, číslo a vid. Pro kategorii OSOBY jsou zde definovány tradiční hodnoty: *první osoba*, *druhá osoba* a *třetí osoba*. V kategorii ČÍSLA u sloves rozlišuje atributivní systém *číslo jednotné* a *číslo množné*. U VIDU nalezneme všechny tři hodnoty: *perfektivum, imperfektivum* a *obouvidé*. Kategorie způsob, slovesný rod a čas částečně vyplývají z hodnot definovaných v kategorii TYP (MÓD), ale vlastními kategoriemi v rámci tagsetu nedisponují. Typ (mód) reflektuje syntetické tvary sloves, najdeme zde částečně hodnoty odrážející slovesný rod, způsob a čas, patří zde hodnoty: *infinitiv, indikativ prézentu, imperativ, příčestí činné (minulé), příčestí trpné, přechodník přítomný (současnost), přechodník minulý (dřívější děj)* a *indikativ futura*.

V **kompaktním tagsetu** jsou vyjádřeny všechny základní gramatické kategorie sloves s výjimkou vidu. V kategorii ZPŮSOB jsou kromě tradičních hodnot (*indikativ, imperativ, kondicionál*) zahrnuty také hodnoty označující některé specifické slovesné tvary: *infinitivu, příčestí* a *přechodníku*, tyto slovesné formy jsou z hlediska způsobu problematické, infinitiv a přechodník nevyjadřují slovesný způsob vůbec, příčestí jsou zase součástí analytických tvarů, mohou být tedy součástí kondicionálu nebo indikativu préterita apod. Z hlediska ČASU rozlišuje kompaktní tagset klasicky *prézens*, *futurum* a čas *minulý*. Pojetí kategorie OSOBY (*první, druhá, třetí*), ČÍSLA (*singulár, plurál*) i slovesného RODU (*aktivum, pasívum*) je tradiční.

Xerox tagset neobsahuje komplexní ztvárnění žádné z výše zmíněných základních gramatických kategorií. Způsob je částečně reflektován prostřednictvím hodnot: *verb: imperative* (+VERB_IMP, imperativ) a *verb: present indicative* (+VERB_PRI, indikativ prézentu). Z uvedených hodnot vyplývá, že kategorie způsobu je sice zohledněna, ale ne zcela.

Kódovník PMK odráží všechny základní kategorie sloves. VID obsahuje tradičně hodnoty pro *imperfektivum, perfektivum* a sloveso *obouvidové*. Kategorie čísla a osoby jsou spojeny v jedné kategorii kódovníku nazvané OSOBA, ČÍSLO, OSOBOVOST, DRUH INFINITIVU. Osoba a číslo jsou pojaty tradičně, údaje o nich jsou však spojeny, tudíž zde nalezneme hodnoty pro: *1. sg., 2. sg, 3. sg., 1. pl., 2. pl.* a *3. pl.* Vedle těchto hodnot jsou v kategorii přítomny také hodnoty pro *infinitiv aktivní*, *infinitiv pasivní*, u kterých je reflektována rovněž kategorie slovesného rodu, a hodnoty *neosobní tvar singuláru* a *neosobní tvar plurálu*, kde je však již brána

v potaz stránka významová, formálně se jedná o tvary 3. os. singuláru a plurálu. Zbývající základní kategorie sloves (rod, způsob a čas) jsou v kódovníku shrnuty do kategorie ZPŮSOB, ČAS, SLOVESNÝ ROD, přičemž způsob je přítomen i v kategorii IMPERATIV A NEURČITÉ TVARY. Kategorie způsob, čas, slovesný rod obsahuje hodnoty kombinující zmíněné gramatické kategorie: *ind. préz. akt.*, *ind. préz. pas.*, *kond. préz. pas.*, *ind. prét. akt.*, *ind. prét. pas.*, *kond. prét. pas.*, *ind. fut. akt.*, *ind. fut. pas.* Z hodnot vyplývá, že z kategorie způsobu je zde reflektován pouze indikativ a kondicionál, spolu s hodnotami času a slovesného rodu, jež u nich mohou být určovány. *Imperativ* nalezneme v kategorii následující, vedle neurčitých tvarů: *imperativ aktivní*, *participium pasivní*, *přechodník přítomný aktivní*, *přechodník minulý aktivní*. Z uvedených hodnot je patrné, že i zde je určován slovesný rod. Jedná se již o třetí kategorii v rámci kódovníku, která zohledňuje gramatickou kategorii slovesného rodu. Kódovník PMK je typickým příkladem tagsetu, který nakládá se slovesnými gramatickými kategoriemi velmi volně.

3.2.5.2 Rozšiřující kategorie

Poziční tagset obsahuje několik rozšiřujících kategorií. Jak již bylo zmíněno výše, na druhé pozici (SLOVNÍ PODDRUH) částečně reflektuje kategorii způsobu, explicitně je vyjádřena hodnotami: *kondicionál slovesa být („by“, „bych“, „bys“, „bychom“, „byste“)* a *slovesný tvar rozkazovacího způsobu*. Dále zde najdeme přechodníky: *slovesný tvar přechodníku přítomného („-e“, „-íc“, „-íce“)*; *slovesný tvar přechodníku minulého, příp. (zastarale) přechodník přítomný dokonavý*; kromě přechodníků také infinitiv nebo hodnoty pro příčestí: *slovesné tvary minulého aktivního příčestí (vč. přidaného „-s“)* a *slovesné tvary pasívního příčestí (vč. přidaného „-s“)*. Je zde také reflektována archaičnost některých tvarů (je tudíž zohledněn stylistický příznak, i když je pravdou, že se tyto tvary vyznačují specifickou formou zakončení): *archaické slovesné tvary minulého aktivního příčestí (zakončení „-ť“)* a *archaické slovesné tvary přítomného a budoucího času (zakončení „-ť“)*. Na druhé pozici je rovněž definována hodnota kombinující různé tvary sloves: *sloveso, tvar přítomného nebo budoucího času*. Zvlášť jsou vydělena *slovesa jako zkratky*.

Další rozšiřující kategorií pozičního systému je gramatický ROD, hodnoty jsou stejné jako u substantiv (viz kap. č. 3.2.1.1), není zde však obsažena hodnota pro

maskulinum inanimatum (rod mužský neživotný), na rozdíl od substantiv zde najdeme hodnoty kombinující nemající oporu v teorii: *femininum nebo neutrum (tedy nikoli maskulinum)*; f*emininum singuláru nebo neutrum plurálu (pouze u příčestí a jmenných adjektiv)*; *masculinum inanimatum nebo femininum (jen plurál u příčestí a jmenných adjektiv), masculinum animatum nebo inanimatum*. Určují se kategorie NEGACE a VARIANTA, STYLOVÝ PŘÍZNAK APOD. více viz kap. č. 3.2.1.2.

V **atributivním systému** se u sloves vyskytuje také několik rozšiřujících kategorií. Kategorii NEGACE viz kap. č. 3.2.2.2 a ROD viz kap. č. 3.2.2.1. STYLISTICKÝ PŘÍZNAK TVARU a TYP TVARU pak viz kap. č. 3.2.1.2.

Kompaktní systém rovněž uvádí několik rozšiřujících kategorií: typ, rod (jm.), negace, životnost a klit. „s“. V kategorii TYPU vyděluje hodnoty pro sloveso: *významové, pomocné, modální* a *sloveso „být“*. ROD (JM.) a ŽIVOTNOST viz u substantiv v kap. č. 3.2.1.1. Kategorie NEGACE s hodnotami *negace* a *klad* je podobná stejnojmenné kategorii pozičního systému, více viz kap. č. 3.2.1.2. Poslední kategorii KLIT. „S“ viz kap. č. 3.2.3.2 (a potažmo 3.2.1.2).

Xerox tagset obsahuje hodnoty blíže určující slovesný tvar. Kromě těch již zmíněných v souvislosti se způsobem u kategorií základních to jsou: *verb: infinitive* (+VERB_INF, infinitiv), *verb: past participle* (+VERB_PAP, příčestí činné) a *verb: transgressive* (+VERB_TRA, přechodník). Nepostihuje komplexně všechny syntetické tvary (není zde například zohledněno příčestí trpné apod.).

Kódovník PMK neobsahuje žádné podrobné určení slovního druhu slovesa. JMENNÝ ROD A ČÍSLO jsou shrnuty do kategorie jediné, nalezneme zde určení všech čtyř rodů, všechny v singuláru a v plurálu: *maskulinum životné v singuláru, maskulinum neživotné v singuláru, femininum v singuláru,* n*eutrum v singuláru, maskulinum životné v plurálu, maskulinum neživotné v plurálu, femininum v plurálu, neutrum v plurálu*, případně hodnotu pro případ, kdy rod a číslo *nelze určit*. Další kategorií je KLAD A ZÁPOR s hodnotami *forma pozitivní* a *forma negativní,* tato kategorie je podobná kategorii NEGACE u pozičního systému, viz kap. č. 3.2.1.2. Analytičnost tvarů je reflektována v kategorii VÍCESLOVNOST A REZULTATIVNOST: *forma jednoslovná, forma víceslovná nezvratná, forma zvratná nesložená* a *forma zvratná složená*; rezultativ tamtéž nabývá hodnot: *rezultativ přítomný, rezultativ minulý,*

rezultativ budoucí, *rezultativ v infinitivu* a *rezultativ kondicionálový*. Syntakticky zaměřená kategorie VALENCE je u sloves neobvykle rozsáhlá, dokonce jí byly v rámci kódu věnovány dvě pozice. Poslední je kategorií určovanou u sloves je STYL, více viz výše.

3.2.6 Příslovce

U příslovcí nejsou primárně určovány žádné gramatické kategorie. MČ2 upozorňuje, že: „Na rozdíl od ostatních základních slovních druhů jsou adverbia slova neohebná,“[106] avšak připouští, že: „některá se ovšem (stejně jako adjektiva) stupňují.“[107]

Kategorie SLOVNÍ PODDRUH se v **pozičním tagsetu** odvíjí od kategorií na dalších pozicích, zohledňuje, zda jsou u daných příslovcí určovány, či nikoli. Obsahuje hodnoty: *příslovce (bez určení stupně a negace; „pozadu“, „naplocho“, ...)* a *příslovce (s určením stupně a negace; „velký“, „zajímavý“, ...)*. STUPEŇ je pojat stejně jako u adjektiv (viz 3.2.2.1). NEGACI a kategorii VARIANTA, STYLOVÝ PŘÍZNAK APOD. pak viz u substantiv kap. č. 3.2.1.2. U příslovcí jsou v pozičním systému také definovány hodnoty po VID, to je způsobeno pravděpodobně skutečností, že jako příslovce jsou označeny ustrnulé tvary přechodníku: „chtě“, „nechtě“.

V **atributivním systému** je specifikován i významový druh příslovce, a to ve dvou stejnojmenných kategoriích (obdobně jako u zájmen): DRUH ZÁJMENNÉHO PŘÍSLOVCE, které obsahují hodnoty pro příslovce: *ukazovací*, *vymezovací*, *způsobové*, *stavové*, *tázací*, *vztažné*, *záporné* a *neurčité*. Kategorie negace a stupeň obsahují stejné hodnoty, jako tomu je u adjektiv, NEGACI viz kap. č. 3.2.2.2 (potažmo 3.2.1.2), STUPEŇ viz 3.2.2.1. Dále jsou pro příslovce definovány také kategorie STYLISTICKÝ PŘÍZNAK TVARU a TYP TVARU, viz kap. č. 3.2.1.2.

Pokud jde o **tagset kompaktní** v kategorii TYP je obsažena pouze jediná hodnota pro příslovce *obecné*, klasifikaci příslovcí z hlediska významového zde na rozdíl od atributivního tagsetu nenajdeme. Nechybí kategorie STUPNĚ, více viz 3.2.2.1.

[106] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 188.
[107] Tamtéž, s. 188.

Xerox tagset nedefinuje pro příslovce, kromě určení SLOVNÍHO DRUHU, žádnou gramatickou kategorii.

Kódovník PMK obsahuje v rámci bližší specifikace slovního druhu příslovce dvě kategorie druh a třídu. U kategorie druh převažuje hledisko formální, zatímco u kategorie třída významové a funkční. Pro DRUH jsou definovány hodnoty: *nespecifické*, *predikativum*, *zájmenné nespojkové*, *spojovací výraz jednoslovný* a *spojovací výraz víceslovný*. Pro TŘÍDU zde najdeme hodnoty: *deskriptivní*, *evaluativní*, *intenzifikační*, *restriktivní* a *deskriptivní časoprostorové*. Ani v kódovníku nechybí kategorie STUPNĚ, více viz kap. č. 3.2.2.1. Přítomná je také do syntaxe zasahující VALENCE, či do lexikální roviny zasahující STYL.

3.2.7 Předložky

MČ2 uvádí předložky jako nezákladní slovní druh, který: „je tvořen třídami funkčně i sémanticky nesamostatných morfémových útvarů, které specifikují a modifikují pád substantiva a jeho funkčních ekvivalentů a spolu s ním se podílejí na výstavbě věty a textu.“[108] Předložky tedy samy pád nevyjadřují. Jsou však s pádem spojeny prostřednictvím substantiva či jeho funkčního ekvivalentu, např. zájmena. V tagsetech je tato skutečnost zohledňována, mnohé obsahují určení **pádu**, se kterým se daná předložka pojí. Avšak určování této kategorie u předložek již zasahuje do syntaxe.

V **pozičním tagsetu** je přítomna kategorie PÁDU, nabízí určení všech sedmi pádů kromě vokativu, s nímž se žádná předložka nepojí. Další kategorií určovanou u předložek je SLOVNÍ PODDRUH, jenž uvádí hodnoty: *předložka, obyčejná*; *předložka vokalizovaná („ve“, „pode“, „ku“, ...)* a *součást předložky, která nikdy nestojí samostatně („nehledě“, „vzhledem“, ...)*. Hlavním kritériem je forma předložky, poslední hodnota také reflektuje výrazy vícečlenné. Určována je také kategorie VARIANTA, STYLOVÝ PŘÍZNAK APOD. (viz kap. č. 3.2.1.2).

Atributivní systém uvádí u předložek pouze kategorii PÁDU, obsahuje všechny tradiční hodnoty kromě vokativu.

[108] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 197.

Kompaktní tagset rovněž u předložek určuje kategorii PÁDU, na rozdíl od předchozích tagsetů nenabízí vedle vokativu ani hodnotu pro nominativ. Tím je v rozporu s MČ2, která uvádí, že se i nominativ může pojit s předložkami, tyto předložky jsou vesměs cizího původu, např. kontra, versus, via apod.[109] Kategorie TYP s jedinou hodnotou, *předložka*, pouze reduplikuje určení slovního druhu. V kategorii TYP-FORMA je brán zřetel na vícečlenné tvary předložek, najdeme zde hodnoty: *předložka jednoduchá* a *složená nebo část složené*.

Xerox tagset u předložek určuje pouze kategorii SLOVNÍ DRUH.

Kódovník PMK obsahuje určení pádu, přiřazuje mu kategorii VALENČNÍ PÁD, v níž jsou definovány hodnoty pro všech sedm pádů, navíc také hodnota věnovaná problematickým případům: *nejasné a jiné*. Kategorie DRUH kombinuje hledisko původu a vícečlennosti předložek, reflektuje předložky vícečlenné, ale také zda se jedná o předložky primární či sekundární, a to hodnotami: *vlastní jednoslovná*, *nevlastní jednoslovná* a *víceslovná*. Kategorie TŘÍDA je založená na funkčně-sémantických vlastnostech předložek, rozlišuje prepozice: *lokální*, *temporální* a *jiné*. Kategorií již silně zasahující do roviny syntaktické je FUNKČNÍ ZÁVISLOST. Obdobně jako u předchozích slovních druhů i zde je definována kategorie STYL (viz kap. č. 3.2.1.2).

3.2.8 Spojky

U spojek nejsou primárně určovány žádné gramatické kategorie. MČ2 charakterizuje spojky jako slova neohebná (s výjimkou *aby, kdyby*).[110] Existence několika málo ohebných spojek se v tagsetech projevuje tím, že jsou v souvislosti s těmito tvary definovány kategorie a hodnoty, jež ve spojitosti s jinými spojkami nenajdeme.

V **pozičním systému** je z gramatických kategorií určována osoba a číslo. Pro ČÍSLO zde najdeme hodnoty *singuláru* a *plurálu*, dále pak hodnotu kombinační: *libovolné číslo (P/S/D)*. Pro OSOBU pak tradičně hodnoty: *1. osoba*, *2. osoba* a *3. osoba*. Kategorie SLOVNÍ PODDRUH bere v úvahu povahu vztahů vyjadřovaných spojkami, vyjadřování gramatických vztahů mezi jednotkami jazyka se již vztahuje k syntaxi, z tohoto hlediska se spojky dělí na parataktické (souřadicí) a hypotaktické

[109] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 208.
[110] Tamtéž, s. 214.

(podřadicí). Nalezneme zde hodnoty: *spojka podřadicí (vč. „aby“, a „kdyby“ ve všech tvarech)*, *spojka souřadicí* a *slovo „krát“ (slovní druh: spojka)*. Slovo „krát“ je však příslovce, nikoli spojka. Poslední určovanou kategorií je VARIANTA, STYLOVÝ PŘÍZNAK APOD., viz kap. č. 3.2.1.2.

Atributivní tagset v kategorii DRUH spojky také rozlišuje spojky na *souřadicí* a *podřadicí*. Dále je zde přítomna také kategorie TYP TVARU, viz kap. č. 3.2.1.2.

Kompaktní systém podobně jako systém poziční určuje u spojek kategorii OSOBY a ČÍSLA (viz v kap. č. 3.2.5.1). Také v kategorii TYP rozlišuje spojky *souřadicí* a *podřadicí*.

Xerox tagset pro spojky nedefinuje žádnou gramatickou kategorii.

Kódovník PMK v rámci kategorie DRUH zohledňuje opět několik kritérií, jednak druh spojky v závislosti na gramatickém vztahu, jenž vyjadřuje, jednak reflektuje spojky vícedílné. Je prezentován hodnotami: *souřadicí jednoslovná*, *podřadicí jednoslovná*, *souřadicí víceslovná*, *podřadicí víceslovná* a *nelze určit*. Kategorie TŘÍDA je založena na funkci spojek, nenajdeme zde však tradiční hodnoty představené i v MČ2 (např. spojky slučovací, odporovací, stupňovací apod.), ale skupiny, které v sobě těchto tradičních tříd spojek obsahují hned několik, jsou to spojky: *kombinační*, *specifikační*, *závislostní*, *časové* a *jiné*. Kategorií typicky syntaktickou i zde zůstává VALENCE. Přesah do lexikální roviny je reprezentován kategorií STYL.

3.2.9 Částice

U částic nejsou určovány žádné gramatické kategorie. Partikule jsou: „nezákladní slovní druh, vyjadřující vztah mluvčího ke sledované skutečnosti, k adresátovi, k obsahu či členění textu.“[111] Vzhledem k charakteru tohoto slovního druhu u něj nenalezneme v tagsetech mnoho kategorií, které by u něj byly určovány.

Poziční systém uvádí u částic kategorii SLOVNÍ PODDRUH, která však pouze reduplikuje informaci o slovním druhu, tj. že je to *částice*. Poslední kategorií určovanou je VARIANTA, STYLOVÝ PŘÍZNAK APOD., viz kap. č. 3.2.1.2.

[111] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 228.

V **atributivním systému** je pro částice definována pouze kategorie typ tvaru, viz kap. č. 3.2.1.2.

Kompaktní tagset neobsahuje žádnou kategorii blíže specifikující slovní druh částice.

Xerox tagset rovněž nezahrnuje kromě určení slovního druhu žádnou kategorii vztahující se k částicím.

Oproti tomu v **kódovníku PMK** nalezneme v rámci částic kategorií hned několik. Bližší určení slovního druhu je obsaženo v kategoriích druh a třída. Kategorie DRUH reflektuje, zda se jedná po stránce tvarové o částici homonymní s jiným slovním druhem, či nikoli. Najdeme zde hodnoty pro částice: *vlastní, nehomonymní*; *homonymní s adverbiem*; *homonymní se spojkou* a *jiné*. MČ2 uvádí, že až čtvrtina všech částic je homonymní s jinými slovními druhy, resp. slovními tvary.[112] Kategorie TŘÍDA již specifikuje částice z hlediska funkčně-sémantického, jsou zde definovány hodnoty: *faktuální*, *faktuálně-evaluativní*, *faktuálně-intenzifikační*, *voluntativní*, *voluntativně-evaluativní*, *expresivně-evaluativní/intenzifikační*, *emocionálně-evaluaitvní/intenzifikační*, *faktuálně-expresivní* a *jiné kombinace*. Třídění částic do skupin na základě funkčně-sémantického hlediska je však velmi obtížné a závisí mimo jiné na kontextu, v němž se daná částice nachází. Funkčně--sémantická klasifikace částic se v kódovníku od pojetí MČ2 liší, a to i z hlediska terminologie. Syntakticky zaměřenou kategorií je VALENCE. Další kategorií zasahující do syntaxe je zde MODUS VĚTY. Poslední kategorií zde určovanou je výše zmiňovaný STYL.

3.2.10 Citoslovce

U citoslovcí neurčujeme žádné gramatické kategorie. MČ2 charakterizuje citoslovce jako nezákladní slovní druh, jehož základní morfologickou charakteristikou je neohebnost.[113] V tagsetech nenajdeme mnoho kategorií, které by se vztahovaly k citoslovcím. Jejich funkce a významy jsou různorodé. Klasifikace citoslovcí je

[112] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 228.
[113] Tamtéž, s. 239.

obtížná, to také podporuje fakt, že se jedna o neuzavřený, produktivní slovní druh,[114] který se vyznačuje velkou variabilitou.

V **pozičním systému** je v rámci citoslovcí definovaná pouze kategorie SLOVNÍ PODDRUH, která přináší pouze zmnožení informace o určení slovního druhu, *citoslovce*.

Atributivní tagset citoslovcím kromě určení slovního druhu žádnou kategorii nepřisuzuje.

Systém kompaktní rovněž neobsahuje žádnou kategorii blíže specifikující slovní druh citoslovce.

V **Xerox tagsetu** také není zahrnuta kromě určení slovního druhu žádná kategorie vztahující se k citoslovcím.

V **kódovníku PMK** nalezneme i u citoslovcí bližší určení slovního druhu. Kategorie DRUH podobně jako u částic reflektuje možnou homonymii citoslovcí s jinými slovními druhy, obsahuje hodnoty pro citoslovce: *běžné, původní*; *homonymní s tvarem substantiva*; *homonymní s tvarem adjektiva*; *homonymní s deiktikem*; *homonymní s tvarem slovesa nebo slovesného původu*; *homonymní s adverbiem* a *jiné*. Kategorie TŘÍDA reflektuje funkčně-sémantickou stránku citoslovcí, rozděluje citoslovce na: *faktuální*, *voluntativní*, *emocionální*, *kontaktové*, *onomatopoické*, *voluntativní a kontaktové*, *voluntativní a emocionální*, *voluntativní a onomatopoické*, *emocionální a kontaktové* a *jiné*. Tato klasifikace neodráží plně třídění MČ2, ale jak již bylo řečeno výše, klasifikace citoslovcí je velmi problematická, zvlášť když uvážíme, že se v rámci citoslovcí vyskytuje řada přesahů mezi skupinami a četné jevy periferní.[115] Přítomná je zde také kategorie STYL.

3.2.11 Ostatní kategorie na úrovni slovního druhu

3.2.11.1 Poziční tagset

Poziční tagset na úrovni SLOVNÍCH DRUHŮ obsahuje dvě hodnoty: *neznámý, neurčený, neurčitelný slovní druh* a *interpunkce, hranice věty*.

[114] *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia, 1986, s. 242.
[115] Tamtéž, s. 249.

Neznámý, neurčený, neurčitelný slovní druh

Tato hodnota představuje tzv. záchrannou hodnotu v rámci pozičního systému, jak již bylo naznačeno v kap. č. 1.3 v souvislosti s problematickým určením/zařazením některých slovních tvarů. Spadají sem právě ty tvary, které nelze jednoznačně přiřadit do žádné jiné skupiny. Tato hodnota slovního druhu je blíže specifikována v kategorii SLOVNÍ PODDRUH, která nabízí hodnoty pro: *nerozpoznaný slovní tvar*; *slovní tvar, který byl rozpoznán, ale značka chybí* a *zkratka, slovní druh neurčen, neznámý*.

Interpunkce, hranice věty

Poslední hodnota definovaná v rámci kategorie slovní druh (1. pozice) se zabývá označováním speciálních znaků, především interpunkce. V kategorii SLOVNÍ PODDRUH je blíže určeno, zda se jedná o *hranici věty* či *interpunkci* všeobecně.

3.2.11.2 Atributivní tagset

Atributivní tagset pro kategorii slovní druh uvádí na úrovni tradičních deseti slovních druhů ještě dvě hodnoty: zkratka a by, aby, kdyby.

Zkratka

K této hodnotě se nevážou žádné další kategorie.

By, aby, kdyby

U této hodnoty najdeme v atributivním tagsetu čtyři kategorie. První kategorie vztah k slovesnému módu reflektuje kondicionál. Dále je určována kategorie osoby a čísla, viz kap. č. 3.2.5.1. S přesahem do stylistiky a potažmo i lexikologie je zde přítomna kategorie stylistický příznak tvaru.

3.2.11.3 Kompaktní tagset

V kompaktním systému nalezneme také dvě další hodnoty na úrovni SLOVNÍHO DRUHU: *zkratky* a *zbytek*.

Zkratky

Zkratky jsou v rámci tagsetu vyděleny zvlášť, není tedy brán ohled na to, o jakou zkratku se jedná, jaké slovo či slovní spojení je zkracováno (substantivum,

adjektivum, …), jako je tomu např. v případě pozičního systému. Není přítomna žádná další specifikace této hodnoty.

Zbytek

I tato hodnota patří mezi tzv. záchranné hodnoty, více viz výše. Nenajdeme zde žádnou bližší specifikaci.

3.2.11.4 Xerox tagset

Xerox tagset obsahuje také hodnoty, které nejsou přiřaditelné k žádnému z deseti slovních druhů. Jsou to hodnoty: ***clitic*** (klitika), tři hodnoty pro interpunkci – ***comma***, čárka; ***punctuation***, interpunkce; ***sentence final punctuation***, interpunkce značící konec věty – a ***date*** (datum). K žádné z uvedených hodnot se nevážou kategorie, které by je blíže specifikovaly. Hodnota klitika obsahuje příklonné tvary slov napříč různými slovními druhy např. sloveso „být“, či spojka „-li“). Datum je v Xerox tagsetu vyděleno zvlášť, není naznačeno, že by se mohlo přiřadit k určitému slovnímu druhu, i když je tu možnost, že by mohlo být řazeno k číslovkám, jelikož příklady u této hodnoty uvedené jsou psány pouze číslicemi (měsíc není ani jednou vypsán slovy). Avšak zařazení není jasné, a proto datum uvádíme zde.

3.2.11.5 Kódovník PMK

V kódovníku PMK jsou na úrovni SLOVNÍHO DRUHU přítomny hodnoty: *frazém/idiom* a *jiné*.

Frazém/idiom

Touto hodnotou kódovník PMK výrazně zasahuje také do syntaktické roviny jazyka, která pro morfologický tagset není určující. Zohledňuje vztahy mezi slovy. Hodnota frazém/idiom je blíže specifikována kategorií DRUH s hodnotami: *verbální, substantivní, adjektivní, adverbiální, propoziční* a *jiné*. Dále se k hodnotě frazém/idiom váží různé kategorie VALENCE vztahující se k syntaxi. Opět zde najdeme i kategorii STYL.

Jiné

K této hodnotě se váže několik kategorií. Blíže je specifikována kategoriemi druh a poddruh. DRUH se soustřeďuje na charakter označovaného tvaru a obsahuje tyto

hodnoty: *cizojazyčný výraz*, *zkratka* a *proprium*. PODDRUH pak s hodnotami *jednoslovné* a *víceslovné* reflektuje případnou víceslovnost daného jevu. Dále jsou zde přítomny kategorie RODU, ČÍSLA a PÁDU, viz kap. č. 3.2.1.1. Poslední je kategorie STYL.

3.3 Výsledky komparace morfologických tagsetů s teorií prezentovanou v Mluvnici češtiny 2

V předchozí kapitole jsme provedli analýzu morfologických tagsetů a jejich komparaci s teorií vyloženou v tzv. *akademické Mluvnici češtiny 2*. Naše analýza potvrdila, že se od sebe tagsety v míře odrazu morfologických kategorií češtiny značně liší. Nejmenší rozdíly byly zaznamenány v okruhu základních gramatických kategorií určovaných u jednotlivých slovních druhů. Pokud jde o tzv. kategorie rozšiřující, pozorujeme u systémů značkování značnou diverzitu. V souvislosti s těmito rozšiřujícími kategoriemi byla u morfologických tagsetů zjištěna poměrně vysoká míra přesahu z roviny morfologické do jiných rovin jazyka (lexikální – např. STYL, syntaktické – např. VALENCE).

Analýza také odhalila variabilitu v pojetí jednotlivých kategorií i v rámci jednoho tagsetu. Stává se, že v rámci jednoho anotačního systému jsou pro jednu kategorii (např. rod) definovány u různých slovních druhů různé hodnoty, např. v atributivním tagsetu je u substantiv přítomná hodnota *rodina (příjmení)*, kdežto třeba u adjektiv ji už nenalezneme. V této kapitole přineseme shrnutí základních poznatků vyplývajících z analýzy jednotlivých tagsetů a jejich konfrontace s lingvistickou teorií.

3.3.1 Poziční tagset

Bylo zjištěno, že poziční tagset u ohebných slovních druhů odráží všechny základní gramatické kategorie, avšak nebylo tomu tak vždy, např. kategorie VIDU byla do systému přidána až se zveřejněním korpusu SYN2005 (rok prvního zveřejnění: 2005), 16. pozici obsahuji již všechny anotované korpusy ČNK uvedené po roce 2005. U sloves byla ohledně základních gramatických kategorií také zaznamenána

výjimka v podobě slovesného způsobu, který nemá v tagsetu vlastní kategorii, ale je reflektován na druhé pozici v rámci bližší specifikace slovního druhu.

Druhá pozice u pozičního systému se jeví jako velmi problematická. Napříč slovními druhy je svým zastoupením velmi nevyrovnaná, např. u substantiv obsahuje pouze dvě hodnoty, to je počet ve srovnání s jinými (zvláště ohebnými) slovními druhy (adjektiva: osm hodnot, zájmena: 20 apod.) zanedbatelný, když nadto uvážíme, kolik prostoru je pro substantiva vyčleněno ve výkladech MČ2, je tato situace zarážející. Právě na této kategorii (SLOVNÍ PODRUH, druhá pozice) lze demonstrovat variabilitu v pojetí jednotlivých kategorií u různých slovních druhů, jinými slovy: stejná kategorie určovaná u různých slovních druhů nemusí obsahovat vždy stejný počet hodnot.

U pozičního tagsetu také dochází ke kombinování různých kritérií při definování hodnot v rámci jedné pozice, tato situace byla zaznamenána opět u druhé pozice tagsetu, např. u adjektiv, kdy jsou v tagsetu k dispozici hodnoty vytvořené s ohledem na formální podobu adjektiva a navíc je zde obsažena hodnota, která reflektuje skutečnost, jež má na formální podobu adjektiva naopak vliv minimální: *slovo před pomlčkou*.

Zmíněná hodnota nás přivádí k dalšímu poznatku, který jsme ohledně pozičního tagsetu získali. Poziční systém obsahuje řadu hodnot, které nemají z hlediska lingvistické teorie opodstatnění, ať se již jedná o podobné případy jako výše, či tzv. hodnoty kombinační, které pouze kombinují známé hodnoty základní, např. hodnota pro rod: *masculinum (animatum nebo inanimatum)*. Podobného charakteru jsou také hodnoty typu *libovolné číslo, libovolný rod* apod., avšak v jejich existenci se, jak vyvozujeme v kapitole č. 1.3, promítá fakt, že zařazení některých tvarů je velmi problematické. Problematické určování některých slovních tvarů se projevuje i na vyšší úrovni, a to existencí tzv. záchranné hodnoty na úrovni slovního druhu, zde *neznámý, neurčený, neurčitelný slovní* druh.

V neposlední řadě je třeba zmínit, že poziční tagset obsahuje také kategorie přesahující morfologickou rovinu jazyka, patří zde zejména kategorie VARIANTA, STYLISTICKÝ PŘÍZNAK APOD. s přesahem do lexikální roviny.

3.3.2 Atributivní tagset

V rámci atributivního tagsetu jsou reflektovány všechny základní gramatické kategorie. U sloves je situace problematická, nemají v tagsetu k dispozici vlastní kategorie pro způsob, slovesný rod a čas, jsou reflektovány v kategorii TYP (MÓD).

Také zde se setkáváme ohledně jednotlivých kategorií s vnitřní variabilitou systému (pro totožnou kategorii jsou definovány u různých slovních druhů různé hodnoty), jak bylo poznamenáno výše, např. pro ROD je u substantiv přítomná hodnota *rodina (příjmení)*, kterou u dalších slovních druhů nenalezneme. I když je pravda, že vnitřní variabilita je méně zastoupená než v systému pozičním.

I zde se objevuje hodnota související s problematickým určováním slovních tvarů, avšak pouze jedinkrát, a to u zájmen v kategorii OSOBY: *první nebo druhá nebo třetí*. Nejen, že se zde hodnota takového charakteru objevuje pouze jednou, ale dokonce v systému chybí jakákoliv tzv. záchranná hodnota, která by byla na úrovni slovního druhu.

V atributivním systému rovněž najdeme přesah z morfologické roviny. V kategorii STYLISTICKÝ PŘÍZNAK TVARU se např. odráží rovina lexikální. U číslovek se objevuje také kategorie spjatá s formální gramatikou, TERMINÁL V GRAMATICE.

3.3.3 Kompaktní tagset

V kompaktním systému rovněž pozorujeme určení téměř všech základních gramatických kategorií u ohebných slovních druhů, u sloves chybí kategorie vidu. Kompaktní tagset se také od ostatních liší v tom, že z kategorie rodu je zvlášť vydělena životnost, kategorie rodu je tedy v systému reflektována v podobě dvou oddělených kategorií (ROD a ŽIVOTNOST).

Nepozorujeme zde téměř žádnou vnitřní variabilitu, hodnoty definované pro totožné kategorie u různých slovních druhů jsou takřka identické, jednou z mála výjimek je např. kategorie ČÍSLA a i ta je odlišná pouze u sloves, kde na rozdíl od předchozích slovních druhů nenajdeme hodnotu pro duál. Lze tedy říci, že v tomto směru je kompaktní tagset velmi jednotný.

Nenajdeme zde hodnoty kombinační. Avšak nechybí tzv. záchranná hodnota na úrovni slovního druhu, zde nazvaná *zbytek*.

Z analýzy a následné komparace s teorií vyplynulo, že se tento tagset velmi pevně drží morfologické roviny. Neobsahuje kategorie výrazně zasahující např. do lexikální roviny (stylistický příznak apod.) jako výše uvedené tagsety.

3.3.4 Xerox tagset

Xerox tagset je nejméně podrobným systémem značkování z uvedených pěti. Neobsahuje určení téměř žádné základní gramatické kategorie u ohebných slovních druhů. Uvádí pouze kategorii pádu u zájmen a číslovek, částečně je reflektován způsob u sloves.

Zásadní odlišností Xerox tagsetu od ostatních systémů morfologického značkování je jeho stručnost, neobsahuje takřka žádné kategorie uvedené u ostatních tagsetů. Také se místy liší v pojetí slovních druhů, například když zařazuje číslovky řadové mezi adjektiva.

3.3.5 Kódovník PMK

Jedná se zřejmě o nejpodrobnější tagset pro značkování češtiny. Odráží všechny základní gramatické kategorie ohebných slovních druhů. Specifická je situace u sloves, kdy jsou sice reflektovány všechny základní gramatické kategorie, avšak pouze vid má svou samostatnou kategorii. Číslo a osoba jsou spojeny do jedné kategorie, podobně pak způsob s časem a se slovesným rodem.

Jak je patrné z komentáře slovesných kategorií, i v kódovníku se vyskytují kategorie, které odrážejí více než jedno kritérium a jsou tak vytvářeny kategorie poměrně komplikované.

Problematické označování určitých slovních tvarů se zde projevuje existencí hodnot: *nelze určit, jiné*. Také jedna hodnota na úrovni slovního druhu je pojmenována *jiné*.

Kódovník PMK jako nejrozsáhlejší tagset z pěti uvedených také oplývá největším množstvím kategorií přesahujících morfologickou rovinu jazyka, za lexikální rovinu

jmenujme kategorii STYL, za rovinu syntaktickou pak např. kategorie VALENCE či FUNKCE.

3.3.6 Otázka možné vzájemné kompatibility

Z uvedených poznatků vyplývá, že se mezi sebou morfologické anotační systémy značně liší, rozsáhlostí, pojetím, a jak jsme mohli pozorovat v části práce věnující se analýze tagsetů a jejich komparaci s MČ2, tak i terminologií. Největší míru shody bychom mohli nalézt u základních gramatických kategorií, které jsou reflektovány téměř všemi tagsety. Zde spatřujeme jako nejrozporuplnější otázku sloves. Jak je patrné z výše uvedených výsledků, patří kategorie určované u sloves mezi nejproblematičtější a jsme v jejich případě také svědky největší variability ohledně jejich zpracování.

V tagsetech najdeme jen omezené množství hodnot, které jsou totožné a schopné převoditelnosti mezi systémy. Dále u tagsetů najdeme hodnoty, které mohou být společné pro omezený počet systémů a také hodnoty unikátní, nevyskytující se jinde než právě v jednom ze systémů. Z našeho výzkumu vyplývá, že nejlépe převoditelný napříč systémy by mohl být paradoxně Xerox tagset, a to z toho prostého důvodu, že je nejstručnější, a tudíž obsahuje hodnoty natolik elementární, že je u nich pravděpodobnost, že by se mohly dostat do rozporu s hodnotami v jiných systémech minimální, téměř nulová. Je ale otázkou, zda je v praxi skutečně použitelný a ve svých kategoriích dostatečně vyvážený.

4 Závěr

Hlavním cílem této bakalářské práce bylo srovnání morfologických tagsetů, kategorií a hodnot, které obsahují, s lingvistickou teorií, reprezentovanou *Mluvnicí češtiny 2*. Všech pět morfologických tagsetů (poziční, atributivní, kompaktní, Xerox tagset i kódovník PMK) bylo podrobeno analýze a konfrontováno s teorií. Důraz byl kladen na jádro morfologického tagsetu, tedy určování morfologických kategorií. Vzhledem k různorodým koncepcím tagsetů jsme se však nevyhnuli ani jiným rovinám jazyka vedle morfologické (např. rovině lexikální či syntaktické), a to zejména u kódovníku PMK, který je z uvedených tagsetů nejpodrobnější.

Předložili jsme podrobnou analýzu konstruovanou s ohledem na lingvistickou teorii. Dospěli jsme tak k charakteristikám systémů morfologického značkování značně ovlivňujícím jejich možnou vzájemnou kompatibilitu. Naše analýza potvrdila velkou vzájemnou odlišnost tagsetů. Nejmenší rozdíly byly zaznamenány v rámci základních gramatických kategorií, v tomto užším okruhu jsou jako nejproblematičtější slovní druh vnímána slovesa, která se ohledně zpracování gramatických kategorií k nim vztažených vyznačují velkou variabilitou.

Velká diverzita však byla odhalena také uvnitř jednotlivých systémů, pro jednu kategorii jsou uvnitř tagsetu definovány v souvislosti s různými slovními druhy různé hodnoty. Nejmenší mírou variability ve zpracování jednotlivých gramatických kategorií uvnitř systému byla zjištěna u kompaktního tagsetu. Můžeme říci, že se v tomto směru vyznačuje velkou jednotností a také systematičností.

Dalším vypozorovaným prvkem systémů byla existence tzv. záchranných hodnot souvisejících s problematickým určováním některých slovních tvarů. Zde se promítly problémy ohledně lingvistické anotace naznačené v první části práce.

Poněkud stranou výše zmíněných charakteristik stojí Xerox tagset, který je nejstručnější, právě díky tomu, že obsahuje takřka výhradně elementární hodnoty, obsahuje nejméně sporných bodů s lingvistickou teorií. I zde se rovněž potvrdily teze z první části práce.

U jednotlivých systémů lze vypozorovat hodnoty schopné převoditelnosti mezi všemi tagsety (velmi malé množství), mezi omezeným okruhem tagsetů a také

hodnoty unikátní, vyskytující se právě v jednom z tagsetů. Zde lze zařadit hodnoty nadstavbové přítomné především v kódovníku PMK ve spojitosti s kategoriemi zasahujícími do jiných plánů jazyka než do toho morfologického. Jistá míra převoditelnosti je tedy možná, ovšem v omezené míře a se zvláštním ohledem na specifika jednotlivých systémů vyplývající z jejich podstaty založené na unikátní koncepci.

Anotace

Jméno a příjmení: Adéla Hanová

Název fakulty: Filozofická fakulta

Název katedry: Katedra bohemistiky

Název práce: Morfologické značkování korpusů češtiny – komparace

Anglický název: Morphological Tagging of Czech Corpuses – a Contrastive Study

Vedoucí práce: PhDr. Petr Pořízka, Ph.D.

Počet znaků: 127 081

Počet příloh: 0

Počet titulů použité literatury: 25

Klíčová slova: anotace, morfologické tagsety, tagy, gramatické kategorie, lingvistická teorie, vzájemná kompatibilita

Key words: annotation, morphological tagsets, tags, grammatical categories, linguistic theory, mutual compatibility

Shrnutí

Tato práce se zabývá morfologickými tagsety dostupnými pro značkování češtiny. Tagsety jsou konstruovány na základě odlišných koncepcí a morfologické kategorie odrážejí v různé míře a s různou komplexností. Cílem této práce je porovnat pět systémů morfologického značkování češtiny s lingvistickou teorií. Zaměřujeme se na systém poziční, atributivní, kompaktní, Xerox tagset a kódovník PMK, které stručně charakterizujeme v první části práce. Na základě analýzy tagsetů a jejich konfrontace s teorií, které tvoří jádro práce, pozorujeme charakteristické rysy těchto systémů, které mají vliv na vzájemnou kompatibilitu jednotlivých tagsetů, k níž se vyjadřujeme v závěru práce.

Resumé

This thesis deals with morphological tagsets used for annotation of Czech language. Several morphological tagsets of Czech language exist nowadays. These tagsets differ one from another in their conceptions. They differ by the reflection of morphological categories in various degrees of complexity. The aim of this thesis is to compare five morphological tagsets with the linguistic theory. We focus on these morphological tagsets: positional, attributive, compact, Xerox tagset and the coder PSC, briefly characterized in the first part of thesis. Main part of this thesis consists of analysis of aforementioned tagsets and its comparison with linguistic theory. Based on the analysis and comparison we observed main characteristics of tagsets influencing their mutual compatibility commented in the final part of thesis.

Seznam použité literatury

Literatura

BEDNAŘÍKOVÁ, Božena. *Slovo a jeho konverze*. Olomouc: Univerzita Palackého v Olomouci, 2009. 253 s. ISBN 978-80-244-2220-6.

ČERMÁK, František et al. *Frekvenční slovník mluvené češtiny*. Praha: Karolinum, 2007. 510 s. ISBN 978-80-246-1425-0.

ČERMÁK, František. Korpusová lingvistika dnešní doby. In: ČERMÁK, František a Renata BLATNÁ, eds. *Korpusová lingvistika – stav a modelové přístupy*. Praha: NLN, 2006, s. 9–18. ISBN 80-7106-861-6.

ČERMÁK, František. Korpusy včera, dnes a zítra. In: ČERMÁK, František, ed. *Korpusová lingvistika Praha 2011. 2, Výzkum a výstavba korpusů*. Praha: NLN, 2011, s. 10–29. ISBN 978-80-7422-115-6.

JELÍNEK, Tomáš a Vladimír PETKEVIČ. Systém jazykového značkování korpusů současné psané češtiny. In: PETKEVIČ, Vladimír a Alexandr ROSEN, eds. *Korpusová lingvistika Praha 2011. 3, Gramatika a značkování korpusů*. Praha: NLN, 2011, s. 154–170. ISBN 978-80-7422-116-3.

JELÍNEK, Tomáš. Morfologické značkování a lemmatizace v korpusech ČNK. In: ŠTÍCHA, František a Mirjam FRIED, eds. *Gramatika a korpus 2007: sborník příspěvků ze stejnojmenné konference, 25.–27.9.2007, Liblice = Grammar & Corpora 2007: selected contributions from the conference Grammar & Corpora, Sept. 25–27, 2007, Liblice*. Praha: Academia, 2008, s. 169–180. ISBN 978-80-200--1634-8.

LEECH, Geoffrey. Anotační systémy pro značkování korpusů. In: ČERMÁK, František et al. *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 185–197. ISBN 80-7184-893-X.

LOTKO, Edvard. *Slovník lingvistických termínů pro filology*. 3. nezměn. vyd. Olomouc: Univerzita Palackého, 2003. ISBN 80-244-0720-5.

Mluvnice češtiny 2. Tvarosloví. Praha: Academia, 1986.

POŘÍZKA, Petr a Markus SCHÄFER. MorphCon – A software for Conversion of Czech Morphological Tagsets. In: LEVICKÁ, Jana, ed. a GARABÍK, Radovan, ed. *NLP, Corpus Linguistics, Corpus Based Grammar Research*. Brno: Tribun, 2009, s. 292–301. ISBN 978-80-7399-875-2.

ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999. 94 s. ISBN 80-7184-847-6.

Časopisecké články

ADAM, Robert. Znovu a šířeji o formě komunikace. *Naše řeč*. 2006, č. 4, s. 198–204. ISSN 0027-8203.

JEDLIČKA, Alois. Druhý svazek akademické Mluvnice češtiny. *Naše řeč*. 1989, č. 3, s. 140–151. ISSN 0027-8203.

KOŘENSKÝ, Jan. Jaké gramatiky češtiny dnes a zítra? *Naše řeč*. 2007, č. 4, s. 169–174. ISSN 0027-8203.

Internetové zdroje

CZPJ. *Centrum zpracování přirozeného jazyka* [online]. © 2001–2013 [cit. 2014-04--16]. Dostupné z: http://nlp.fi.muni.cz.

ČERMÁK, František. Pražský mluvený korpus. ÚČNK. *Český národní korpus* [online]. [cit. 2014-03-30]. Dostupné z: http://ucnk.ff.cuni.cz/pmk.php.

HAJIČ, Jan. Popis morfologických značek – poziční systém. In: KOPŘIVOVÁ, Marie a Jan KOCEK. *Manuál korpusového manažeru bonito* [online]. [cit. 2014-03--16]. Dostupné z: http://ucnk.ff.cuni.cz/bonito/znacky.php.

PETKEVIČ, Vladimír. Popis morfologických značek použitých v korpusu orw-mte. *Český národní korpus* [online]. [cit. 2014-03-28]. Dostupné z: http://ucnk.ff.cuni.cz/orwell.php.

RYCHLÝ, Pavel. Bonito. In: CZPJ. *Centrum zpracování přirozeného jazyka* [online]. [cit. 2014-04-24]. Dostupné z: http://nlp.fi.muni.cz/projects/bonito/bonito.html.cz.

RYCHLÝ, Pavel. NoSketch Engine. In: CZPJ. *Centrum zpracování přirozeného jazyka* [online]. Poslední aktualizace 3. 3. 2013 [cit. 2014-04-24]. Dostupné z: http://nlp.fi.muni.cz/projects/bonito/bonito.html.cz.

SEDLÁČEK, Radek. *AJKA tagset* [online]. © 2006 [cit. 2014-03-19]. Dostupné z: http://nlp.fi.muni.cz/projekty/ajka/tags.pdf.

SKOUMALOVÁ, Hana. *Program pro vytváření tagů* [online]. [cit. 2014-03-16]. Dostupné z: http://utkl.ff.cuni.cz/~skoumal/morfo/?lang=cs.

ÚČNK. Korpus ORWELL. *Český národní korpus* [online]. [cit. 2014-03-19]. Dostupné z: http://ucnk.ff.cuni.cz/orwell.php.

ÚFAL. Corpora. *Institute of Formal and Applied Linguistics* [online]. © 2014 [cit. 2014-04-24]. Dostupné z: http://ufal.mff.cuni.cz/projects/corpora.

XEROX CORPORATION. *Open Xerox* [online]. ©1999–2014 [cit. 2014-03-23]. Dostupné z: http://open.xerox.com/.

Seznam tabulek

Tabulka 1: Ukázky tagů na příkladech substantiva „dům“ a spojky „a“.....24
Tabulka 2: Popis kategorií a hodnot pozičního systému na příkladu tagu pro substantivum „dům“.....24
Tabulka 3: Popis kategorií a hodnot atributivního systému na příkladu tagu pro substantivum „dům“.....24
Tabulka 4: Popis kategorií a hodnot kompaktního systému na příkladu tagu pro substantivum „dům“.....25
Tabulka 5: Popis kategorií a hodnot Xerox tagsetu na příkladu tagu pro substantivum „dům“.....25
Tabulka 6: Popis kategorií a hodnot kódovníku PMK na příkladu tagu pro substantivum „dům“.....25