Osolsobě, Klára

Morfologické značkování složených slovesných tvarů v korpusu

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 1999, vol. 48, iss. A47, pp. [33]-50

ISBN 80-210-2098-9
ISSN 0231-7567

SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 47, 1999

KLÁRA OSOLSOBĚ

MORFOLOGICKÉ ZNAČKOVÁNÍ SLOŽENÝCH SLOVESNÝCH TVARŮ V KORPUSU

V našem článku se budeme zabývat dvěma okruhy problémů. V první části se dotkneme otázek souvisejících s obtížemi, jež přináší formalizace pravidel postavení jednotlivých komponent českých složených slovesných tvarů. Pokusíme se formulovat pravidla pro automatickou analýzu složených slovesných tvarů v české větě. Ve druhé části ukážeme na základě analýzy materiálu subkorpusu ČNK.

ČNK — Český národní korpus se buduje od roku 1993 za podpory GAČR. Od roku 1996 je práce na něm koncentrována na samostatném pracovišti ÚČNK na FF UK v Praze. V současné době zahrnuje cca 100 000 000 slovních tvarů. DESAM je anotovaným subkorpusem ČNK a zahrnuje cca 1 milion označkovaných slovních tvarů. DESAM sleduje, jak se jednotlivé slovosledné typy uplatňují v textech.

Na závěr srovnáme frekvenční zastoupení jednotlivých tvarů a jejich variant a ukážeme, jaký mají tato fakta význam pro automatické morfologické značkování složených slovesných tvarů.

1. Automatická lemmatizace a desambiguace víceslovných slovesných tvarů

Při automatizaci gramatického značkování korpusu je třeba řešit problém automatické lemmatizace víceslovných pojmenování. Program LEMMA používaný s úspěchem na FF MU a FI MU pro automatické značkování korpusu (vytvoření korpusu DESAM viz Osolsobě, Pala, Rychlý, 1998), vykazuje jistá omezení, která se týkají automatické lemmatizace složených slovesných tvarů. LEMMA je schopna automaticky přiřadit základní tvar a možné gramatické významy v případě, že slovu definovanému jako řetězec znaků oddělených mezerami odpovídá lemma. Na základě slovníku pevných slovních spojení typu *ad hoc, pars pro toto,* který je součástí automatického analyzátoru LEMMA, je možné automaticky lemmatizovat i slovní spojení, která jsou vázána pevnými slovoslednými pravidly. LEMMA ale dosud neumožňuje automatické rozpoznání a lemmatizaci složených slovesných tvarů. Tento fakt značně omezuje využitelnost korpusu označkovaného LEMMOU pro statistická šetření v oblasti

výzkumu českého slovesa, ale i pro mnohá další zkoumání. Cílem tohoto příspěvku je zamyslet se nad možnostmi řešení problémů spjatých s nespojitostí složek uvnitř složených slovesných tvarů.

1.1 Značkování sloves

Automatické značkování prováděné s využitím morfologického analyzátoru LEMMA nabízí pro slovesa následující typy značek. První typ je značka pro jednoduché určité tvary (indikativ prézentu aktiva nedokonavých sloves, indikativ aktiva dokonavých sloves, imperativ). Druhý typ označuje pasívní a nebo l-ové participium, třetí typ infinitiv. Popis jednotlivých typů srov. tab. 1

k5	eA/B	p1/2/3	nS/P	tP/B/M	mI/R/K	aI/P
k5	eA/B	pI	nS/P	tM/_	mP	aI/P
k5	eA/B				mF	aI/P

V prvním sloupci je atribut slovní druh označený k, který nabývá hodnoty 5 (sloveso). Ve druhém sloupci je atribut negace (e) nabývající hodnoty A (-negace) a B (+negace). Třetí sloupec obsahuje atribut osoba (p) nabývající hodnoty 1,2,3 (první, druhá, třetí osoba) nebo hodnoty I (neosobní forma). Ve čtvrtém sloupci je uveden atribut číslo (n) s hodnotami S (singulár) nebo P (plurál). V pátém sloupci je atribut čas (t), který může mít hodnoty P (přítomný), B (budoucí), M (minulý), _ (žádný). Šestý sloupec obsahuje atribut modus nebo typ neurčitého tvaru (m) s hodnotami I (indikativ), R (imperativ), K (kondicionál), P (participium), F (infinitiv). V sedmém sloupci je atribut vid (a) nabývající hodnoty I (nedokonavý), P (dokonavý).

Slovesa jsou tudíž značkována pouze jako jednotlivé slovní tvary bez ohledu na funkci tvarů. Z toho plyne, že např. slovo ***dělám*** bude označeno následujícím způsobem: ***dělám<dělat,*** k5eAp1nStPmIaI>. Úplně stejnou značku bude mít ovšem sloveso ***jsem <být,*** k5eAp1nStPmIaI>, a to ve velmi rozmanitých větách, v nichž plní velmi rozmanité funkce (srov. např.: ***Jsem zde. Dělal jsem svou práci dobře. Byl jsem zklamán, ale teď jsem nadšen...).*** Je na první pohled patrné, že značkování sloves založené na naznačeném principu je přinejmenším zjednodušující.

1.3. Složené slovesné tvary

Složené slovesné tvary představují vysoké procento všech slovesných tvarů (cca 50%). Z toho vyplývá, že je nanejvýš potřebné zabývat se kvalitním korpusovým značkováním těchto tvarů. V tomto článku se budeme snažit formálně popsat pravidla pro rozpoznání složeného slovesného tvaru v české větě.

Nejdříve je potřeba ujasnit si, které tvary zahrneme do našeho popisu. Slovosledná pravidla organizující postavení jednotlivých složek víceslovného slovesného tvaru se opírají o pravidla přízvuku na straně jedné a pravidla aktuálního členění na straně druhé a jsou obecná v tom smyslu, že se týkají všech typů složených slovesných tvarů. Na jedné straně stojí složené tvary času, způsobu a slovesného rodu, na straně druhé se z jejich rámce vydělují tvary s reflexívním *se*, a konečně složené tvary s modálními a fázovými slovesy a celou řadou dal-

ších sloves, která se z hlediska slovosledu chovají stejným způsobem. Vzhledem k omezením, které přináší rozsah tohoto článku, se posledně jmenovanými nebudeme úžeji zabývat. Konstatujeme pouze, že tvoří podmnožinu námi sledovaných tvarů a z hlediska slovosledných variant jsou pouze variantami nad variantami, které uvedeme. Přesto bude v budoucnu třeba se jimi zabývat samostatně.

1.3.1 Inventář komponentů složených slovesných tvarů

Složený slovesný tvar v tom smyslu, jak jsme o něm hovořili výše, je složen ze dvou a více prvků (slov, řetězců). Tyto prvky nemají stejnou hodnotu ani z hlediska funkce, kterou plní, ani z hlediska formy, ani z hlediska možností svého vztahu k prvkům, které nejsou součástí složeného tvaru. Podejme tedy nejprve inventář zmíněných prvků.

Námi navržený seznam používá notace zavedené v korpusovém manažeru GCQP:

Osobní tvary [tag=„k5.*p1.*“ & tag=„k5.*p2.*“ & tag=„k5.*p3.*“]
Infinitiv [tag=„k5.*mF.*“]
reflexivní *se* [lemma=„sebe“]
l-ové participium [tag=„k5.*tMmP.*“]
tvar pomocného slovesa *být* pro tvoření tvarů minulého času [lemma=„být“ & tag=„k5.*p1.*tPmI.*“ & tag=„k5.*p2.*tPmI.*“]
tvar pomocného slovesa *být* ve futuru [lemma=„být“ & tag=„k5.*tFmI.*“]
tvary „by“ [lemma=„*by“]
Termínem tvar *„by“* označujeme osobní formy *bych, bys,....bychom,..,aby,.. kdybyste*
l-ové participium pomocného slovesa *být* [word=„byl.*’ & lemma=„být“]
tvar pomocného slovesa *být* pro tvoření tvarů prézentu pasíva [lemma=„být“ & tag=„k5.*tPmI.*“]
pasívní participium [tag=„k5.*t_mP.*“]
l-ové participium slovesa *bývat* [word=„býval.* & lemma=„bývat“]

Tyto prvky se dále mohou klasifikovat jako lexikální složky tvaru a gramatické složky tvaru. V rámci navržených pravidel automatické analýzy a generování složených slovesných tvarů má toto rozdělení vliv především na slovosledou volnost či vázanost složky, a to jak ve vztahu k prvkům nepatřícím do složeného tvaru, tak vůči prvkům, které k němu patří, a na schopnosti složek „odtrhnout se“ a vytvořit prostor pro prvek, který nespoluutváří složený slovesný tvar (nespojité složené slovesné tvary).

1.3.2 Pravidla pro syntézu a analýzu složených slovesných tvarů

V následujících řádcích budeme stručně formulovat slovosledná pravidla složených slovesných tvarů.

1.3.2.1 Indikativ prézentu aktiva — tvary s reflexívním *„se“*

Termín reflexívní *„se“* používáme pro úsporu místa za *„se“* nebo *„si“*, které jsou sou-

částí reflexívního slovesného tvaru. Termín reflexívní slovesný tvar používáme bez ohledu na funkci reflexívního zájmena *„se"* nebo *„si"* ve složeném tvaru.

Tvar je tvořen osobním tvarem indikativu prézentu aktiva a reflexívním *„se"*. Existují dvě slovosledné varianty. V první z nich stojí *„se"* na prvním místě složeného tvaru. Tato slovosledná varianta umožňuje vznik nespojitého tvaru, takže určitý slovesný tvar indikativu prezentu aktiva stojí buď bezprostředně po *„se"*, nebo po něm následuje nespojitě. Druhá varianta je opačná. Na prvním místě stojí určitý tvar indikativu prézentu aktiva a bezprostředně po něm bez možnosti nespojitého tvaru stojí *„se"*.

*Příklady: 1. ..., že <se při těchto činnostech **aktivují**> oblasti mozkové kůry.*
*2. <**Zvyšuje se**> obrat, klesá podíl...*

1.3.2.2 Analytické futurum nereflexivních tvarů

Tento složený slovesný tvar vytvářejí dvě složky, jejichž postavení vykazuje větší samostatnost, než tomu bylo u předcházejících případů. Přesto se opět setkáváme se dvěma variantami. V první z nich stojí na prvním místě určitý tvar futura slovesa *„být"*. Po něm bezprostředně nebo nespojitě následuje tvar infinitivu významového slovesa. Ve druhé variantě je pořadí složek opačné, mohou, ale nemusí být nespojité.

*Příklady: 1. <**Budeme** jej nadále **budovat**> tak,...*
*2. <**Přihlížet mu bude**> i trenér olympioniků Ivan...*

1.3.2.3 Analytické futurum reflexívních tvarů

Tvar má tři složky. Existují čtyři slovosledné varianty. Kritériem je jednak postavení zvratného *„se"*. To stojí buď uvnitř tvaru bezprostředně po první složce (varianta 1 a 2), nebo na prvním místě složeného tvaru (varianta 3 a 4). Za skupinou první složka + reflexívní *„se"* (sloveso *„být"* ve formě futura — varianta 1 nebo infinitiv významového slovesa — varianta 2) následuje bezprostředně nebo nespojitě 2. složka (infinitiv významového slovesa -varianta 1 nebo sloveso *„být"* ve formě futura — varianta 2). Ve variantě 3 a 4 následují další složky složeného tvaru spojitě nebo nespojitě v pořadí sloveso *„být"* ve tvaru futura, po němž opět spojitě nebo nespojitě následuje infinitiv významového slovesa (varianta 3) nebo v opačném pořadí (varianta 4).

*Příklady: 1. <**Budete se** na skoky **dívat**> alespoň v televizi ...*
*2. <**Jednat se bude**> také o koordinačních orgánech...*
*3. Volič <**se nebude bát**> dát svůj hlas...*
*4. ...s touto příručkou <**se** vám to **stávat nebude**>...*

1.3.2.4 Indikativ aktiva minulého času nereflexívních tvarů

Vedle jednoduchých tvarů 3. (popř. 2.) osoby existují složené tvary 1. a 2. osoby. Jsou tvořeny příslušným osobním tvarem indikativu prézentu aktiva slovesa *„být"* a l-ovým participiem významového slovesa v následujících slovosledných variantách. Pokud stojí osobní tvar slovesa *„být"* na prvním místě, pak l-ové participium stojící na druhém místě buď následuje bezprostředně nebo nespojitě (varianta 1). Pokud stojí na prvním místě l-ové participium významo-

vého slovesa, následuje určitý tvar pomocného slovesa *„být"* vždy bezprostředně po něm (varianta 2).

*Příklady: 1. Před sezónou <**jsem** takovou situaci **nečekal**>..*

*2. <**Hrál jsem**> tak, abych se....*

1.3.2.5 Indikativ aktiva minulého času reflexívních tvarů

V rámci tohoto typu existují na rozdíl od předcházejícího pouze složené tvary. Ty se liší počtem složek. Tvary 1. a 2. osoby se skládají ze tří prvků, forma 3. (popř. 2.) osoby je složena pouze ze dvou prvků. Rozlišujeme celkem čtyři slovosledné varianty, první dvě pro tvary 1. a 2. osoby, druhé dvě pro formu 3. (popř. 2.) osoby. Pokud stojí na prvním místě určitý tvar 1. nebo 2. osoby prézentu slovesa *„být"*, pak zvratné *„se"* po tomto tvaru bezprostředně následuje. Třetí složka — l-ové participium významového slovesa může následovat spojitě nebo nespojitě (varianta 1). Pokud stojí na prvním místě l-ové participium významového slovesa, pak po něm vždy spojitě následuje 1. nebo 2. osoba prézentu slovesa *„být"* a za ní opět výhradně spojitě zvratné zájmeno *„se"* (varianta 2), nebo pouze zvratné *„se"* (varianta 4). Pokud stojí na prvním místě zvratné *„se"*, před nímž nepředchází osobní tvar pomocného slovesa, následuje po něm l-ové participium spojitě nebo nespojitě (varianta 3).

*Příklady: 1. Bláhově <**jsem se** například **domníval**>...*

*2. <**Zamiloval jsem si**> hokej...*

*3. ..., když <**se** mí partneři neustále **dívali**> na hodinky...*

*4. ...<**dostala se**> do psychické pohody...*

1.3.2.6 Kondicionál přítomný nereflexívních tvarů — aktivum

Tvar kondicionálu přítomného nereflexívních tvarů sloves v činném rodě sestává ze dvou složek. V první variantě stojí na prvním místě osobní tvary *„by"*. Spojitě nebo nespojitě následuje l-ové participium významového slovesa. Druhá varianta má na prvním místě tvar l-ového participia významového slovesa, po němž vždy spojitě následuje osobní tvar *„by"*.

*Příklady: 1. ..., že <**by** mu vlastnické problémy **bránily**> ve...*

*2. ...<**bouřil bych**> proti všemu...*

1.3.2.7 Kondicionál přítomný reflexívních tvarů — aktivum

Tvar je složen ze tří prvků, z nichž lze složit dvě slovosledné varianty. Varianta 1 má na prvním místě tvar *„by"*, na druhém zvratné zájmeno *„se"*, které následuje vždy bezprostředně po něm, a třetí komponent — l-ové participium významového slovesa — stojí na třetím místě spojitě nebo nespojitě. Varianta 2 má na první místě l-ové participium významového slovesa, po němž bezprostředně následuje tvar *„by"* a zvratné zájmeno *„se"*. V rámci této varianty se nespojité tvary nevyskytují.

*Příklady: 1. ...<**by se** ještě více **aktivoval**> nelegální trh*

*2. ...<**jednalo by se**> o zákon o pojištění...*

1.3.2.7 Kondicionál minulý nereflexívních tvarů — aktivum

Teoreticky vzato se může tento tvar v textu vyskytnout ve třech slovosledných variantách, ačkoliv pouze první dvě jsou zastoupeny ve zkoumaném materiálu, poslední byla nalezena pouze v materiálu ČNK. První varianta má na prvním místě tvar *„by"*, spojitě nebo nespojitě následuje l-ové participium slovesa *„být"* a l-ové participium významového slovesa. Druhá varianta staví na první místo l-ové participium slovesa *„být"*, po němž vždy spojitě následuje tvar *„by"*. L-ové participium významového slovesa stojí na třetím místě spojitě nebo nespojitě. Varianta 3 má na prvním místě l-ové participium významového slovesa, za nímž vždy spojitě následuje tvar *„by"* a nespojitě l-ové participium slovesa *„být"*.

Příklady: *1. <**Kdyby obilí bylo zůstalo**> v republice...*
*2. ...<**byli by ho jistě vyhnali**>....*
*3. <**Považovali bychom byli**> tehdy za úspěch*

1.3.2.9 Kondicionál minulý reflexívních tvarů — aktivum

Tento tvar má dvě základní a třetí potenciální variantu, která se nevyskytla ani v materiálu subkorpusu DESAM, ani v ČNK, a je tudíž označena „!". Složený tvar sestává ze čtyř komponent. Stojí-li na prvním místě tvar *„by"*, následuje zvratné zájmeno *„se"* vždy spojitě po něm, tvar l-ového participia slovesa *„být"* a l-ového participia významového slovesa stojí na třetím a čtvrtém místě buď spojitě, nebo nespojitě (varianta 1). Stojí-li na prvním místě tvar l-ového participia slovesa *„být"*, pak bezprostředně a vždy spojitě následují tvary *„by"* a zvratné *„se"*. Na čtvrtém místě spojitě nebo nespojitě stojí l-ové participium významového slovesa. Varianta 3 staví na první místo l-ové participium významového slovesa, za nímž vždy spojitě následuje tvar *„by"* a zvratné *„se"*. Na posledním místě následuje spojitě nebo nespojitě tvar l-ového participia slovesa *„být"*.

Příklady: *1. ..., že <**by se byla** bojová povaho Markomanů **změnila**> v úsměv...*
*2. <**byla bych si** možná **poslechla**> koledy sama...*
*3. !<**ptal by se byl**>...*

1.3.2.10 Indikativ prézentu pasíva

Tento tvar se skládá ze dvou složek a vyskytuje se ve dvou slovosledných variantách. V první variantě stojí určitý tvar pomocného slovesa *„být"* na prvním místě a po něm spojitě nebo nespojitě následuje pasívní participium významového slovesa. Ve druhé variantě je tomu naopak.

Příklady: *1. ...v jakých buňkách <**je tvorba těchto faktorů aktivována**>...*
*2. ...<**Nastaveny jsou**> standardní hodnoty...*

1.3.2.11 Indikativ pasíva analytického futura

Tvar sestává ze dvou složek, slovesa *„být"* ve futuru a pasívního participia, ve dvou slovosledných variantách. Varianta 1 má na prvním místě osobní tvar

slovesa „*být*", po němž spojitě nebo nespojitě následuje pasívní participium významového slovesa. Varianta 2 má pořadí opačné.

*Příklady: 1. ...čs. tanky <**budou** začátkem ledna **dopraveny**> do Německa...*
*2. ...nikdo za zásah <**postižen nebude**> ...*

1.3.2.12 Indikativ pasíva minulého času

Tento tvar má celkem pět slovosledných variant, a to 3 varianty pro 1. a 2. osobu a 2 varianty pro 3. (popř. 2.) osobu. První tři varianty jsou složeny ze tří komponentů. V první variantě stojí na prvním místě pomocné sloveso „*být*" v osobním tvaru, na druhém místě spojitě nebo nespojitě následuje l-ové participium pomocného slovesa „*být*" a na třetím místě spojitě nebo nespojitě pasívní participium významového slovesa. Druhá varianta má na prvním místě l-ové participium slovesa „*být*", po něm vždy spojitě následuje osobní tvar pomocného sloesa „*být*" a na třetím místě spojitě nebo nespojitě pasívní participium významového slovesa. Třetí varianta má na prvním místě pasívní participium významového slovesa, po něm vždy spojitě následuje osobní tvar slovesa „*být*" a na třetím místě spojitě nebo nespojitě l-ové participium pomocného slovesa „*být*". Varianta 4 a 5 se skládá pouze ze dvou komponent, a to l-ového participia slovesa „*být*" a pasívního participia významového slovesa. L-ové participium slovesa „*být*" je ve variantě 4 na prvním místě a na druhém místě spojitě nebo nespojitě následuje pasívní participium významového slovesa. Varianta 5 má pořadí obrácené.

*Příklady: 1. ...potom <**jsem byl** na poslední chvíli **nominován**> do Anglie...*
*2. ...<**byl jsem** hrubě slovně **napaden**>...*
*3. !<**chválen jsem byl**>*
*4. kdy <**byla** tato choroba poprvé **diagnostikována**> ...*
*5. <**Privatizovány byly**> některé železárny...*

1.3.2.13 Kondicionál přítomný — pasívum

Tvar tvoří tři komponenty — tvar „*by*", l-ové participium slovesa „*být*" a pasívní participium — ve třech slovosledných variantách. První varianta má na prvním místě tvar „*by*", po něm spojitě nebo nespojitě následuje l-ové participium slovesa „*být*" a na třetím místě spojitě nebo nespojitě následuje pasívní participium významového slovesa. Ve druhé variantě je na prvním místě l-ové participium slovesa „*být*", na druhém místě vždy spojitě tvar „*by*" a na třetím místě spojitě nebo nespojitě tvar pasívního participia významového slovesa. Třetí varianta staví na první místo pasívní participium významového slovesa, po něm vždy spojitě následuje tvar „*by*" a na třetím místě spojitě nebo nespojitě l-ové participium.

*Příklady: 1. ...<**aby bylo** obchodní tajemství organizace **chráněno**>...*
*2. ...<**byla by** na mne **upřena**> příliš velká pozornost.*
*3. ...<**zrušeno by bylo**> pět nynějších úřadů.*

1.3.2.14 Kondicionál minulý — pasívum

Tvar se skládá ze čtyř složek v osmi slovosledných variantách. Varianty 1 a 2 mají na prvním místě tvar *„by"* a na posledním místě tvar pasívního participia. Tvar l-ového participia pomocného slovesa *„být"* stojí na druhém místě a tvar l-ového participia slovesa *„bývat"* na třetím místě (varianta 1). Varianta 2 má pořadí opačné. Varianta 3 má na prvním místě l-ové participium slovesa *„být"*, za nímž vždy spojitě následuje tvar *„by"*. Na třetím a čtvrtém místě následuje spojitě nebo nespojitě tvar l-ového participia slovesa *„bývat"* a pasívního participia významového slovesa. Varianta 4 je obměnou varianty 3 v tom smyslu, že má pouze zaměněny pozice l-ových participií pomocných sloves *„být"* a *„bývat"*. Tyto tři tvary fungují jako nespojité složky složeného slovesného tvaru. Varianta 5 a 6 má na prvním místě pasívní participium významového slovesa, po němž bezprostředně vždy spojitě následuje tvar *„by"*. Nespojitě na třetím a čtvrtém místě následuje l-ové participium slovesa *„být"* a *„bývat"* (varianta 5). Varianta 6 má pořadí opačné. Varianty 7 a 8 mají na prvním místě tvar *„by"*, za nímž spojitě nebo nespojitě následuje tvar pasívního participia významového slovesa. Ve variantě 7 stojí na třetím místě spojitě nebo nespojitě l-ové participium slovesa *„být"* a na čtvrtém místě spojitě nebo nespojitě l-ové participium slovesa *„bývat"*. Ve variantě 8 je pořadí třetího a čtvrtého členu opačné.

Následující příklady jsou z ČNK, příklady s ! nebyly v korpusu nalezeny, existují ovšem potencionálně.

*Příklady: 1. ...<**kdyby byly bývaly** včas **vyřčeny**>...*
*2. ...Slovensko <**by bývalo bylo rozděleno**>...*
*3. ...<**bylo by bývalo** Lesbii **probodnuto**> srdce...*
*4. !<**býval by byl chválen**>*
*5. !<**chválen by byl býval**>*
*6. !<**chválen by býval byl**>*
*7. aniž <**by mně sděleno bylo bývalo**>...*
*8. !<**by chválen býval byl**>*

1.3.2.15 Imperativ reflexívních tvarů

Tento tvar má dvě slovosledné varianty. Pokud stojí na prvním místě sloveso v imperativu, pak na druhém vždy spojitě následuje tvar zvratného *„se"* (varianta 1). Pokud stojí *„se"* na prvním místě, tvar slovesa následuje spojitě nebo nespojitě (varianta 2).

*Příklady: 1. <**Dejte si**> panáka...*
*2. ...pak <**se nedivte**>, že to...*

1.3.2.16 Imperativ pasíva

Tvar má dvě slovosledné varianty. Buď stojí na prvním místě tvar pomocného slovesa *„být"* v imperativu a za ním spojitě nebo nespojitě následuje tvar pasívního participia (varianta 1), nebo je pořadí složek opačné (varianta 2).

*Příklady: 1. ...Ale <**buďme připraveni**> na všechno...*
*2. ...!<**připraven buď**>*

2.2 Vyhledávání složených slovesných tvarů v korpusových textech

Automatické gramatické značkování korpusu naráží při značkování složených slovesných tvarů na značné obtíže. Přiřazení gramatických kategorií slovním tvarům je možné pouze s ohledem na celek, jejž spoluutvářejí, nebo lépe řečeno: gramatické významy příslušných kategorií lze rozumně přiřadit pouze celku, a ne jeho částem. Tak například prézentní tvary pomocného slovesa *„být"* mohou signalizovat jak prézens, tak minulý čas, podobně i l-ové participium (kondicionál přítomný). Formálně neosobní tvary l-ového participia signalizují 3. osobu. Automatická lemmatizace a na ní založené automatické značkování korpusu se tudíž nemůže obejít bez automatického určování složek složených tvarů, na němž pak lze vystavět značkování složeného tvaru jako celku. Jednotlivé složky a jejich řazení lze poměrně dobře formálně popsat, jak jsme se o to pokusili v předešlých odstavcích. Největší problém při automatické analýze představují nespojité složky, respektive prostor mezi nimi. Z empirie lze vysoudit, že prostor textu mezi složkami může být více méně libovolně rozsáhlý. Pro realistický přístup k řešení naznačené problematiky je ovšem třeba podívat se na skutečné zastoupení jednotlivých slovosledných variant z hlediska frekvence. Taková analýza pomůže odhalit, které typy jsou frekventované a které okrajové, a ukáže, které otázky je třeba řešit primárně a které lze ponechat „ruční" analýze. Ve druhé části našeho článku se budeme zabývat konkrétní materiálovou analýzou korpusu DESAM.

1.4 Korpus DESAM

Korpus DESAM je subkorpusem ČNK, obsahuje cca 1 milion morfologicky označkovaných tvarů. Slovesné tvary jsou označeny pouze jako jednotlivá slova. Přesto lze prostřednictvím korpusového manažeru (GCQP) vyhledávat případy složených slovesných tvarů, analyzovat jejich frekvenci a typy chyb, k nimž při analýze za pomoci GCQP dochází. Dotazy, které můžeme nad korpusem prostřednictvím GCQP klást, mají totiž formu, která se velmi blíží formálnímu pravidlu popisujícímu hledaný tvar. Uvedené výsledky mohou značně přispět k vytváření korpusových nástrojů pro automatickou desambiguaci (zjednoznačnění) automaticky označených morfologických forem.

Následující řádky obsahují převedení slovně formulovaných pravidel z první kapitoly do formy dotazů pro korpusový manažer GCQP. Pro čtení dotazu je třeba uvést, že se lze ptát na jednotlivé atributy, jimiž mohou být tvary slov (word), slova podle základního tvaru (lemma) a slova podle gramatické charakteristiky (tag). Dotazy jsou uzavřeny ve hranatých závorkách, mají formu typ= a v uvozovkách je uveden útvar, na nějž se tážeme.

Tak např. dotaz [lemma=„.*by"] [word!=„." & lemma!=„sebe"] {0,3} [word= "byl.*] [word!=„."] {0,3} [tag=„k5.*tMmP"] čteme následujícím způsobem: Vyhledej všechny kombinace, které začínají tvarem *„by"*, po nichž následuje nejméně 0 a nejvíce tři slova s výjimkou interpunkčního znaménka („ . ") nebo tvarů reflexívního zájmena *„se"* , dále slovo začínající *byl* (rozuměj tvary *byl...*, *byly*), po němž následuje minimálně 0 a maximálně 3 slova s výjimkou interpunkčního znaménka a končící libovolným tvarem slovesa v l-ovém participiu. Stručně řeče-

no chceme vyhledat všechny případy kondicionálu minulého nereflexívních tvarů ve slovosledné variantě 1 (např.:*Pokud <by vládní koalice byla vyhrála> volby...*).

2.1. Přehled zastoupení jednotlivých tvarů podle slovosledných variant

Pro zjednodušení zkoumaného materiálu jsme omezili rozsah prostoru mezi nespojitými složkami pro frekvenční analýzu na tři slova. Toto omezení by podle předběžného výzkumu nemělo přinést výrazné chyby ve frekvenční analýze. Při zkoumání takto vymezeného materiálu se celkem průkazně ukazuje, že počet spojitých tvarů obecně převažuje nad počtem nespojitých tvarů a že počet tvarů nespojitých klesá s počtem slov uvnitř prostoru tvořeného jednotlivými složkami nespojitých tvarů.

U jednotlivých tvarů a slovosledných variant uvádíme vždy dotaz pro GCQP, číslo udávající počet nalezených odpovědí, číslo, které udává počet hledaných složených tvarů a % chyb.

Tabulka uvádí vždy zastoupení slovosledných variant v poměru jejich výskytů a v %.

2.1.1 Reflexívní tvary v indikativu prézentu aktiva

[lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*tPmI.*“]
6503.....6456.....35 chyb = 0,5%
[tag=„k5.*tPmI.*“ & lemma!=„být“] [lemma=„sebe“]
1696.....1696.....0 chyb = 0 %

varianta 1	6458	79%
varianta 2	1696	21%

Poznámka: Chybné odpovědi ve variantě 1 způsobuje fakt, že nalezené zvratné *„se“* patří ke slovesu, které není součástí hledaného slovesného tvaru (infinitivu, určitého tvaru slovesa, adjektivizovaného přechodníku přítomného), 1 chyba je způsobena chybným značkováním v korpusu DESAM.

2.1.2 .1 Analytické futurum nereflexívních tvarů aktiva

[lemma!=„sebe“] {0,3} [lemma=„být“ & tag=„k5.*tFmI.*“] [lemma!=„sebe“] [word!=„.“] {0,2} [tag=„k5.*mF.*“]
1471.....1471.....0 chyb = 0%
[lemma!=„sebe“] {0,3}[tag=„k5.*mF.*“] [lemma!=„sebe“] [word!=„.“] {0,2} [lemma=„být“ & tag=„k5.*tFmI.*]
34.....34..... 0 chyb = 0%

varianta 1	1467	98,1%
varianta 2	34	1,9%

2.1.2.2 Analytické futurum reflexívních tvarů aktiva

[lemma=„být“ & tag=„k5.*tFmI.*“] [lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*mF.*“]
32.....32.....0 chyb = 0%

[tag=„k5.*mF.*“] [lemma=„sebe“] [word!=„.“] {0,3} [lemma=„být“ & tag=„k5.*tFmI.*“]
11.....11.....0 chyb = 0%
[lemma=„sebe“] [word!=„.“] {0,3} [lemma=„být“ & tag=„k5.*tFmI.*“] [word!=„.“] {0,3} [tag=„k5.*mF.*“]
258.....258.....0 chyb = 0%
[lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*mF.*“] [word!=„.“] {0,3} [lemma=„být“ & tag=„k5.*tFmI.*“]
3.....3.....0 chyb = 0%

varianta 1	32	11%
varianta 2	11	4%
varianta 3	258	84%
varianta 4	3	1%

Poznámka: Pro získání srovnatelných poměrů sečteme výsledky variant 1+3 a 2+4.

2.1.3.1 Indikativ minulého času nereflexívních tvarů aktiva

[lemma=„být“ & tag=„k5.*tPmI.*“] [lemma!=„sebe“] [word!=„.“] {0,2} [tag=„k5.*tMmP.*“]
1459.....1459.....0 chyb = 0%
[tag=„k5.*tMmP.*“] [lemma=„být“ & tag=„k5.*tPmI.*“] [lemma!=„sebe“]
645.....634.....11 chyb = 2%

varianta 1	1459	70%
varianta 2	634	30%

Poznámka: Uvedená data nelze dosti dobře statisticky srovnávat s ostatními výsledky. Je tomu tak proto, že forma 3. osoby nereflexívních tvarů minulého času v češtině je vyjádřena nesloženým tvarem — pouhým l-ovým participiem. Abychom mohli alespoň částečně usuzovat na skutečné rozložení slovosledných variant, museli bychom provést klasifikaci zbývajících jednoduchých tvarů. Druhé variantě by odpovídalo postavení tvaru l-ového participia signalizujícího 3. osobu minulého času indikativu aktiva na prvním místě ve větě.
Co se týče chyb, ve variantě 2 jsou chybně zahrnuty tvary minulého času pasíva (*<byl jsem> hrubě slovně napaden*).

2.1.3.2 Indikativ minulého času reflexívních tvarů aktiva

[lemma=„být“ & tag=„k5.*tPmI.*“] [lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*tMmP.*“]
371.....371.....0 chyb = 0%
[tag=„k5.*tMmP.*“] [lemma=„být“ & tag=„k5.*tPmI.*“] [lemma=„sebe“]
177.....177.....0 chyb = 0%
[word!=„byl.*“] [lemma!=„.*by“] [lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*tMmP.*“]
3663.....3663.....0 chyb = 0%
[tag=„k5.*tMmP.*“] [lemma=„sebe“]
640.....640.....0 chyb = 0%

varianta 1	371	8%
varianta 2	177	4%
varianta 3	3663	75%
varianta 4	640	13%

Poznámka: Uvedené percentuální zastoupení jednotlivých variant lze srovnávat s ostatními variantami, sečteme-li variantu 1+3 (81%) a variantu2+4 (19%). Dosažený výsledek je, jak je na první pohled zřejmé, blízký výsledkům poměru ostatních variant.

2.1.4.1 Kondicionál přítomný nereflexívních tvarů aktiva
[lemma=„.*by“] [lemma!=„sebe“] [word!=„.“] {0,2} [tag=„k5.*tMmP.*“]
3803.....3684.....119 chyb = 3,13%
[tag=„k5.*tMmP.*“] [lemma=„by“] [lemma!=„sebe“]
485.....476.....9 chyb = 1,8%

varianta 1	3684	88%
varianta 2	476	12%

Poznámka: Chyby v obou variantách jsou způsobeny tím, že odpovědi na dotaz v GCQP zahrnují jednak tvary kondicionálu minulého aktiva, jednak tvary kondicionálu přítomného pasíva.

2.1.4.2 Kondicionál přítomný reflexívních tvarů aktiva
[lemma=„.*by“] [lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*tMmP.*“]
894.....889.....5 chyb = 0,67%
[tag=„k5.*tMmP.*“] [lemma=„by“] [lemma=„sebe“]
103.....103.....0 chyb = 0%

varianta 1	889	90%
varianta 2	103	10%

Poznámka: Chyby ve variantě 1 jsou způsobeny tím, že odpověď na dotaz CQP zahrne rovněž tvary kondicionálu minulého.

2.1.5.1 Kondicionál minulý nereflexívních tvarů aktiva
[lemma=„.*by“] [lemma!=„sebe“] [word!=„.“] {0,2} [word=„byl*“] [word!=„.“] {0,3} [tag=„k5.*tMmP.*“]
8.....8.....0 chyb =0 %
[word=„byl.*“] [lemma=„.*by“] [lemma!=„sebe“] [word!=„.“] {0,2} [tag=„k5.*tMmP.*“]
1.....1.....0 chyb = 0%
[tag=„k5.*tMmP.*“] [lemma=„by.*“] [lemma!=„sebe“] [word!=„.“] {0,2} [word=„byl.*“]
0.....0.....0 chyb = 0%

varianta 1	8	89%
varianta 2	1	11%
varianta 3	0	0%

Poznámka: Varianta 3 je z hlediska příznakovosti variantou varianty 2.

2.1.5.2 Kondicionál minulý reflexívních tvarů aktiva

[lemma=„.*by“] [lemma=„sebe“] [word!=„.“] {0,3} [word=„byl.*“] [word!=„.“] {0,3} [tag=„k5.*tMmP.*“]
5.....5.....0 chyb = 0%
[word=„byl.*“] [lemma=„by“] [lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*tMmP.*“]
0.....0.....0 chyb = 0%
[tag=„k5.*tMmP.*“] [lemma=„by“] [lemma=„sebe“] [word!=„.“] {0,3} [word=„byl.*“]
0.....0.....0 chyb = 0%

varianta 1	5	100%
varianta 2	0	0%
varianta 3	0	0%

Poznámka: O variantě 3 platí totéž, co bylo řečeno v předcházející poznámce.

2.1.6.1 Indikativ prézentu pasíva

[lemma=„být“ & tag=„k5.*tPmI.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
1714.....1714.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [word!=„.“] {0,3} [lemma=„být“ & tag=„k5.*tPmI.*“]
45.....38.....7 chyb = 15,5%

Varianta 1	1714	98%
Varianta 2	38	2%

Poznámka: Chyby ve variantě 2 jsou způsobeny tím, že významové a pomocné sloveso nepatří do téhož tvaru.

2.1.6.2 Indikativ futura pasíva

[lemma=„být“ & tag=„k5.*tFmI.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
346.....346.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [word!=„.“] {0,3} [lemma=„být“ & tag=„k5.*tFmI.*“]
3.....2.....1 chyb = 33%

Varianta 1	346	99,4%
Varianta 2	2	0,6%

Poznámka: Chyba ve variantě 2 je způsobena chybným označkováním tvaru v korpusu DESAM.

2.1.6.3 Indikativ minulého času pasíva

[lemma=„být“ & tag=„k5.*tPmI.*“] [word!=„.“] {0,3} [word=„byl.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
32.....31.....1 chyba =0,3 %
[word=„byl.*“] [lemma=„být“ & tag=„k5.*tPmI.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
11.....11.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [lemma=„být“ & tag=„k5.*tPmI.*“] [word!=„.“] {0,3} [word=„byl.*“]

0.....0.....0 chyb = 0%
[word=„byl.*“] [lemma!=„být“ & tag!=„k5.*tPmI.*“] [word!=„.“] {0,2} [tag=„k5.*t_mP.*“]
1384.....1384.....0 chyb = 0%
[tag=„k5.&t_mP.*“] [lemma!=„být“ & tag!=„k5.*tPmI.*“] [word!=„.“] {0,2} [word=„byl.*“]
51.....51.....0 chyb = 0%

varianta 1	29	1,1%
varianta 2	11	0,9%
varianta 3	0	0%
varianta 4	1384	94%
varianta 5	51	4%

Poznámka: Analogicky k případu minulého času reflexívních tvarů je pro získání srovnatelných výsledků třeba sečíst varianty 1+4 (95,1%) 2+5(+3) (4,9%). Chyba ve variantě 1 je způsobena nesprávnou analýzou případu *<je třikrát více, než bylo plánováno>*.

2.1.7.1 Kondicionál přítomný pasíva

[lemma=„.*by“] [word!=„.“] {0,3} [word=„byl.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
110.....110.....0 chyb = 0%
[word=„byl.*“] [lemma=„.*by“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
8.....8.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [lemma=„.*by“] [word!=„.“] {0,3} [word=„byl.*“]
0.....0.....0 chyb = 0%

varianta 1	110	93%
varianta 2	8	7%
varianta 3	0	0%

Poznámka: O variantě 3 platí, to, co bylo řečeno výše.

2.1.7.2 Kondicionál minulý pasíva

[lemma=„.*by“] [word!=„.“] {0,3}[word=„byl.*“] [word!=„.“] {0,3}[word=„býval.*“] [word!=„.“] {0,3}[tag=„k5.*t_mP.*“]
0.....0.....0 chyb = 0%
[lemma=„.*by“] [word!=„.“] {0,3} [word=„býval“] [word!=„.“] {0,3} [word=„byl“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
0.....0.....0 chyb =0%
[word=„byl.*“] [lemma=„.*by“] [word!=„.“] {0,3} [word=„býval.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
0.....0.....0 chyb = 0%
[word=„býval.*“] [lemma=.*by“] [word!=„.“] {0,3} [word=„byl.*“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“]
0.....0.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [lemma=„.*by“] [word!=„.“] {0,3} [word=„byl.*“] [word!=„.“] {0,3} [word=„býval.*“]

0.....0.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [lemma=„.*by“] [word!=„.“] {0,3} [word=„býval.*“] [word!=„.“] {0,3} [word=„byl.*“]
0.....0.....0 chyb = 0%
[lemma=„.*by“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“] [word!=„.“] {0,3} [word=„byl.*“] [word!=„.“] {0,3} [word=„býval.*“]
0.....0.....0 chyb = 0%
[lemma=„.*by“] [word!=„.“] {0,3} [tag=„k5.*t_mP.*“] [word!=„.“] {0,3} [word=„býval.*“] [word!=„.“] {0,3} [word=„byl.*“]
0.....0.....0 chyb = 0%

varianta 1	0	0%
varianta 2	0	0%
varianta 3	0	0%
varianta 4	0	0%
varianta 5	0	0%
varianta 6	0	0%
varianta 7	0	0%
varianta 8	0	0%

Poznámka: Za bezpříznakovou lze považovat variantu 1. Ostatní jsou příznakové.

2.1.8.1 Imperativ reflexívních tvarů
[tag=„k5.*mR.*“] [lemma=„sebe“]
205..... 201 .....4 chyby = 1,9%
[lemma=„sebe“] [word!=„.“] {0,3} [tag=„k5.*mR.*“]
39.....35.. ....4 chyby = 10,2%

varianta 1	201	85,2%
varianta 2	35	14,8%

Poznámka: Chyby ve variantě 1 i 2 byly způsobeny chybami ve značkování korpusu DESAM

2.1.8.2 Imperativ pasívních tvarů
[lemma=„být“ & tag=„k5.*mR.*“] [tag=„k5.*t_mP.*“]
0.....0.....0 chyb = 0%
[tag=„k5.*t_mP.*“] [lemma=„být“ & tag=„k5.*mR.*“]
0.....0.....0 chyb = 0%

varianta 1	0	0%
varianta 2	0	0%

2.2 Přehled zastoupení jednotlivých tvarů podle variant v %
tab. 2

IPA r1	6456	24,8%
IPA r2	1696	6,5%
IFA 1	1467	5,6%
IFA 2	34	0,13%
IFA r1	32	0,12%

IFA r2	11	0,042%
IFA r3	258	0,99%
IFA r4	2	0,0076%
IMA 1	1459	5,6%
IMA 2	634	2,42%
IMA r1	371	1,42%
IMA r2	177	0,68%
IMA r3	3663	14%
IMA r4	640	2,46%
KPA 1	3684	14,1%
KPA 2	476	1,83%
KPA r1	889	3,41%
KPA r2	103	0,39%
KMA 1	8	0,031%
KMA 2	1	0,0038%
KMA 3	0	0%
KMA r1	5	0,019%
KMA r2	0	0%
KMA r3	0	0%
IPP 1	1714	6,59%
IPP 2	38	0,146%
IFP 1	346	1,33%
IFP 2	3	0,0076%
IMP 1	29	0,111%
IMP 2	11	0,042%
IMP 3	0	0%
IMP 4	1384	5,32%
IMP 5	51	0,196%
KPP 1	110	0,422%
KPP 2	8	0,031%
KPP 3	0	0%
KMP 1	0	0%
KMP 2	0	0%
KMP 3	0	0%
KMP 4	0	0%
KMP 5	0	0%
KMP 6	0	0%
KMP 7	0	0%
KMP 8	0	0%
R1	201	0,77%
R2	35	0,13%
RP 1	0	0%
RP 2	0	0%
CELKEM	26005	100%

3. Závěr

Na základě frekvenční analýzy se ukazuje, že ačkoliv pro každý složený tvar existují nejméně dvě slovosledné varianty a vyskytují se i případy 3 až 8 variant, vždy jedna z variant je výrazně frekventovanější než druhá (druhé). Existují tedy jakési neutrální a příznakové varianty. Zajímavé je to, že neutrální (frekventované) varianty aktivních forem se většinou neshodují s variantami, o nichž uvažujeme jako o svého druhu základních tvarech (učebnicových tvarech, tvarech uváděných v gramatikách nebo slovnících). Například ve slovníku se reflexiva tantum (podmnožina zvratných tvarů) uvádějí ve variantě, v níž zvratné *„se"* stojí za slovesem, přičemž ve většině slovosledných variant složených reflexívních tvarů stojí *„se"* před ním. Gramatiky fixují slovosledné varianty, které jsou v textech řidčeji zastoupeny. Tento výsledek frekvenční analýzy ukazuje, že je ovšem třeba primárně počítat s variantami běžnějšími. To by mohlo mít dopad především na výuku češtiny jako cizího jazyka a rovněž na popis češtiny zaměřený na automatizaci její analýzy.

Na základě zjištěných údajů a statistik lze formulovat některé obecnější závěry:

1. Poměr zastoupení příznakových a nepříznakových variant je téměř u všech tvarů srovnatelný. Existují-li např. dvě varianty, jak je tomu ve většině případů, pak bezpříznaková varianta zahrnuje 79% až 100% všech tvarů. Druhá varianta se pohybuje v rozmezí 21% až 0%. Celkový průměr je 90,2% ku 9,8%. Výrazné odchylky vykazují jednak tvary, kde nemáme srovnatelná čísla vzhledem k tomu, že část tvaru není vyjadřována složeným slovesným tvarem (indikativ prézentu aktiva, minulý čas), jednak tvary, které jsou velmi řídké (kondicionál minulý). Tvary, které mají více než dvě slovosledné varianty, lze na dvě varianty převádět (srov. poznámky u jednotlivých tvarů).
2. Z hlediska absolutní frekvence je počet chybně vyhledaných tvarů zanedbatelný. Relativně vysoké procento chyb (33%) u tvarů indikativu futura pasíva je z absolutního hlediska zanedbatelné, protože se týká pouze jednoho chybně určeného tvaru. Podobně je tomu i v případě indikativu prézentu pasíva, kde ve druhé variantě je 15,5% chybně vyhledaných tvarů, které ovšem představují pouze 7 nesprávně vyhledaných případů. V ostatních případech se chyba pohybuje mezi 3,13% až 0%, v průměru 1,47%. Analýza chybně vyhledaných tvarů přispěje v mnoha případech k opravám chyb v subkorpusu DESAM.
3. Na základě frekvenční analýzy můžeme říci, že počet chyb vzniklých při vyhledávání složených slovesných tvarů za pomoci manažeru GCQP je nízký. Lze tudíž předpokládat, že formální analýza založená na automatické analýze jednotlivých složek analyzátorem LEMMA a na pravidlech popisujících slovosledné varianty složených slovesných tvarů bude vykazovat dostatečně vysokou míru úspěšnosti.

LITERATURA

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

HAJIČ, J., HLADKÁ B.: Probabilistic and rule based tagging of an inflective language – a comparison, Technical Report No.1, ÚFAL MFF UK, November 1996.

HAJIČ, J., HLADKÁ, B.: Tagging Inflective Languages: Prediction of morphological categories for a rich, structural tagset, *Technical Report TR-1997–04*, ÚFAL MFF UK, Praha.

JELÍNEK, J., BEČKA, J. V., TĚŠITELOVÁ, M.: Frekvence slov, slovních druhů a tvarů v českém jazyce, SPN, Praha 1961.

LEECH, G.: Corpus Annotation Schemes, in: *Literary and Linguistic Computing*, Vol. 8, No. 4, 1993, 275–281.

TĚŠITELOVÁ, M., a kol.: Kvantitativní charakteristiky současné češtiny, řada *Studie a práce lingvistické*, Academia, Praha 1985.

TĚŠITELOVÁ, M. a kol.: O češtině v číslech, *Malá jazyková knižnice*, Academia, Praha, 1987.

OSOLSOBĚ, K.: Algoritmický popis české formální morfologie a strojový slovník češtiny, disertační práce, FF MU Brno 1996.

OSOLSOBĚ K., PALA. K, RYCHLÝ P.: Frekvence vzorů českých sloves (na materiálu Českého národního korpusu), SaS, 59, 1998, s. 265–277.

PALA, K.: Korpusová lingvistika — informační technologie v lingvistice, Zpravodaj ÚVT MU, Brno 1996.

PALA, K., RYCHLÝ, P., SMRŽ, P.: DESAM – Annotated Corpus for Czech, *Proceedings of SOFSEM '97*, Springer Verlag, New York, Hamburg 1997.

PETR, J. a kol.: Mluvnice češtiny II, Academia, Praha 1986.

ŠEVEČEK, P.: Morfologický analyzátor (lemmatizátor) LEMMA, program v jazyce C, Brno 1995–96.

MORPHOLOGICAL TAGGING OF THE COMPOUND VERB FORMS IN CORPUS

The article elaborates the word-order rules of the compound forms of the Czech verbs analysing the absolute and the relative frequency of the different variants. This analysis is based on electronic source: i.e. on the Czech National Corpus (ČNK), using its grammatically annotated subcorpus DESAM containing 1 026 730 word forms, with a support of the GCQP corpus manager. The present state of the subcorpus DESAM is not satisfactory. Tagging of verb forms in it is too simplified. Each form has only one tag, so that the composed forms are not tagged as different units.

It would be necessary to start with the formulation of the word-order rules of the compound forms (chapter 1). In the chapter 2 the word-order rules are transformed into the form of GCQP-querries and the analysis of obtained material follows.

Comparing the frequency of different word-chains we get a list of the marked and the unmarked variants of each compound verb form. The relation between the marked and the unmarked form is in average 9,8% : 90,2%. The absolute number of the false „answers" (wrongly designated composed forms) is relatively low (1,47% on the average). It means that the automatic analysis based on the chosen rules comes out with high efficiency.

Klára Osolsobě
Ústav českého jazyka
Filozofické fakulty MU
Arna Nováka 1
660 88 Brno
klara@ernest.phil.muni.cz