Markus G i g e r – Pavel Š t ě p á n (Praha)

ČESKÁ DEVERBÁLNÍ PŘÍJMENÍ A PROBLÉM JEJICH HOMONYMIE V ELEKTRONICKÝCH KORPUSECH[1)]

0. Úvod

Formální homonymie různých morfologických tvarů je zásadním problémem při automatickém morfologickém značkování (‚tagging') českých textů[2)]. Významným zdrojem homonymie jsou také vlastní jména, která mohou být formálně totožná s apelativy (srov. toponyma jako *Písek, Most* nebo antroponyma jako *Bílý, Cihla*). Zatímco v případech jako *Písek, Most* a *Cihla* dochází k homonymii mezi nominálním apelativem a rovněž nominálním propriem (lemmata *písek1* a *Písek2*, *most1* a *Most2*, *cihla1* a *Cihla2*)[3)], tak v případě jako *Bílý* se navíc mění slovní druh, což se při morfologickém značkování obtížně zjišťuje a odpovídající příjmení je často lematizováno jako adjektivum.

Stejná situace nastává u mnoha deverbálních příjmení, převážně formálně homonymních s *l*-ovými tvary sloves typu *Vybíral, Odložil*, ale i jiných. Výrazný rozdíl mezi syntaktickými úlohami substantiva a slovesa ovšem poskytuje širší prostor pro hledání řešení (disambiguace), což bude hlavním tématem tohoto příspěvku.

1. Deverbální příjmení vyvolávající homonymii

Základní přehled o slovotvorném i motivačním systému českých příjmení poskytuje J. Beneš[4)]. Jedna kapitola publikace je věnována deverbálním příjmením[5)]. Beneš rozlišuje příjmení z tvarů minulého času, příjmení z tvarů přítomného času,

---

[1)] Tento příspěvek vznikl v rámci projektu Informační společnosti Grantové agentury Akademie věd ČR „Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů" (reg. číslo 1ET101120503). V rámci tohoto projektu jsme také sestavili seznam dosud v odborné literatuře neuvedených příjmení homonymních s *l*-ovým tvarem slovesa a nabídli několik úvah o jejich motivaci (M. Giger – P. Štěpán, Některá dosud nepopsaná česká deverbální příjmení typu *Vybíral, Odložil*, Čeština doma a ve světě, v tisku).

[2)] Srov. např. J. Králík, Vyvážení zdrojů Synchronního korpusu češtiny SYN2000, Slovo a slovesnost 62, 2001, s. 38 – 53; V. Petkevič, Neprojektivní konstrukce v češtině z hlediska automatické morfologické disambiguace českých textů. In: Z. Hladká – P. Karlík (eds.), Čeština – univerzália a specifika 3, Brno 2001, s. 197 – 205; M. Šulc, Tematická reprezentativnost korpusů, Slovo a slovesnost 62, 2001, s. 53 – 61; K. Oliva, O dialektické jednotě protikladů morfologického značkování korpusů a závislostní syntaktické analýze vět. In: M. Šimková (ed.), Aktuálne otázky súčasnej syntaxe, Bratislava 2005, s. 41 – 56.

[3)] Částečně se liší různými paradigmatickými tvary: *cihle – Cihlovi, o cihle – o Cihlovi*; onymické tvary v SYN2000 částečně vůbec nejsou identifikovány.

[4)] J. Beneš, O českých příjmeních, Praha 1962.

příjmení z tvarů rozkazovacího způsobu, příjmení z tvarů neurčitých a příjmení ze slovesného základu. Nejproduktivnější mezi nimi jsou příjmení z tvarů minulého času. Mezi nimi lze zase rozlišovat příjmení končící na *-l*, homonymní s tvarem mužského rodu slovesa, příjmení končící na *-la*, homonymní s tvarem rodu ženského, a příjmení na *-lo*, homonymní s tvarem rodu středního[6)].

Sestavili jsme reprezentativní databázi těchto příjmení na základě retrográdního rejstříku k práci J. Beneše[7)] a na základě abecedního seznamu 60 000 nejrozšířenějších příjmení v České republice, který poskytuje Ministerstvo vnitra ČR na svých internetových stránkách (dále CEO)[8)]. Ohled byl brán i na příjmení, která podle Beneše nejsou deverbální, homonymii s tvarem slovesa však přesto vyvolávají. Příjmení na *-l, -la* a *-lo* bylo Doležalovou excerpováno necelých 800, z nichž přes 540 je homonymních s finitním slovesným tvarem v současné spisovné češtině. Dalších příjmení homonymních s *l*-ovým tvarem slovesa (s jinou motivací nebo u Beneše neuvedených a excerpovaných námi z CEO) je 165. Ostatní tvary jsou mnohem méně početné: Beneš uvádí *Nepovím, Nevím, Pereš, Věříš, Víme*. Z příjmení excerpovaných z CEO jsou relevantní z hlediska homonymie navíc příjmení *Dám, Tučím, Valím, Broulím, Haraším* (homonymní s 1. os. sg. slovesa); *Nemáš, Rubáš, Mokříš, Bereš, Kapeš, Kopeš* (homonymní s 2. os. sg.); *Bude, Mate, Píše, Pomeje, Doleží, Duží, Hluší, Jančí* (homonymní s 3. os. sg.). Dále lze zde uvést i příjmení *Pils; Skočilas, Koplas, Kulas, Mrklas*, která vyvolávají homonymii se staženými tvary 2. os. sg. minulého času.

Jak je patrné, vznik homonymie není vždy dán deverbálním původem příjmení; na druhé straně u některých příjmení, jež skutečně vznikla ze slovesných tvarů, k homonymii nedochází – jedná se o případy, kdy hláskoslovně nářeční anebo obecně-české tvary neodpovídají slovesnému tvaru použitému obyčejně v psaných textech[9)], např. příjmení *Nevečeřal* oproti slovesnému tvaru *nevečeřel* nebo příjmení *Vodehnal* oproti slovesnému tvaru *odehnal*, někdy tvary hláskoslovně obměněné nepravidelně jako příjmení *Nužil* od *neužil*). Příjmení se může také odlišovat pouze pravopisně od slovesného tvaru, např. příjmení *Schrabal* oproti slovesnému tvaru *shrabal* či příjmení *Shíbal* oproti slovesnému tvaru *shýbal*. Nakonec příjmení může být vytvořeno ze

---

[5)] J. Beneš, o. c. v pozn. 4, s. 268 – 296.
[6)] K další literatuře viz M. Giger – P. Štěpán, o. c. v pozn. 1.
[7)] V. Doležalová, Retrográdní rejstřík ke knize Josefa Beneše „O českých příjmeních“, Zpravodaj Místopisné komise ČSAV 15, 1974, s. 309 – 507.
[8)] Nejčetnější příjmení v ČR. http://www.mvcr.cz/statistiky/jmena/.
[9)] V případě obecněčeských tvarů může k homonymii dojít, když se sloveso v obecněčeské formě v korpusu používá, srov. příjmení *Hejbal* a sloveso *hejbat*. Úloha řešení homonymie však zůstává stejná; obecněčeský a spisovný tvar jsou lematizovány jako dva lexémy.

slovesa, které je nářeční anebo zastaralé, popř. vůbec vyšlo z úzu, srov. např. příjmení *Šámal* od nářečního slovesa *šámat, šámat se*[10], příjmení *Šamonil* od nářečního slovesa *šamonit*[11]. Všechny tyto případy nepředstavují z hlediska řešení homonymie zvláštní problémy.

Vytvořili jsme databázi z korpusu SYN2000, která sloužila jako inspirace pro nacházení možných řešení (238 konkordancí obsahujících buď jednotlivé tvary – příkazy typu *Bořil, Sekal* – anebo všechny tvary určitého lemmatu – příkazy typu [lemma="bořit"], [lemma="sekat"]). Někde byly použity i příkazy typu [word="Čekal" & tag="V.*"], resp. [word="Čekal" & tag="N.*"], které dovolily přímo zjistit, kde tvary příjmení jsou morfologicky značkovány jako tvary slovesa a naopak.

2. Stav morfologického značkování tvarů typu *Vybíral, Odložil* v SYN2000 v roce 2005

Tvary deverbálních příjmení byly často interpretovány mylně jako slovesné a naopak. Lze rozlišit několik skupin:

A. Jak sloveso, tak příjmení je poměrně vzácné; neexistuje obyčejně lemma pro příjmení, tvary příjmení jsou lematizovány jako tvary slovesa. V krajním případě tvary příjmení představují jediné zdánlivé doklady na dané sloveso:

– *Dosoudil* – sloveso v korpusu není, 3 doklady na příjmení jsou lematizovány pod ním;

– *Drkal* – v korpusu slovesný tvar *drkal* není, 2 doklady na příjmení *Drkal*, lematizovány pod slovesem;

– *Hvízdala* – tvar má 7 dokladů, 1 na příjmení, lematizován pod slovesem;

– *Nešetřil* – tvar má 17 dokladů, 6 na příjmení, lematizovány pod slovesem;

– *Nezkusil* – tvar má 9 dokladů, 4 na příjmení, lematizovány pod slovesem

– *Papala* – tvar má 4 dokladů, všechny na příjmení, lematizovány pod slovesem.

– *Šetřil* – tvar *Šetřil* má 13 dokladů, 3 na příjmení, lematizovány pod slovesem;

– *Škrábal* – tvar *Škrábal* má 3 doklady, 1 na příjmení, lematizován pod slovesem;

B. Jde o případy, kdy příjmení má poměrně vysokou frekvenci, slovesný tvar poměrně nízkou. Obyčejně existuje zvláštní lemma pro příjmení, dochází však k oboustranným přehmatům, nadsazen bývá spíše počet slovesných tvarů. Srov. např.:

– *Kasal* – tvar *Kasal* má 795 dokladů, z nichž je 62 lematizováno pod slovesem, všechny jsou však ve skutečnosti doklady na příjmení;

– *Musil* – tvar *Musil* má 853 dokladů, všechny jsou lematizovány jako tvary příjmení; v rámci prvních 500 dokladů je ovšem *Musil* 66krát tvarem slovesa;

– *Nakládal* – tvar má 72 dokladů, vše příjmení, avšak lematizováno pod slovesem;

---

[10] Příruční slovník jazyka českého 5, Praha 1948 – 1951, s. 1017.
[11] Příruční slovník jazyka českého 5, o. c. v pozn. 11, s. 1017.

– *Navrátil* – tvar má 967 dokladů, 73 z nich je lematizováno pod slovesem; všechny takto lematizované tvary jsou ovšem doklady na příjmení;

– *Nekvasil* – tvar má 187 dokladů, vše příjmení, z toho je 13 lematizováno pod slovesem;

– *Rozehnal* – tvar *Rozehnal* má 76 dokladů, z toho 71 lematizováno jako jméno, 2 jsou přitom slovesné tvary. 5 dokladů je lematizováno pod slovesem, z nich jsou však 4 doklady na příjmení;

– *Šmejkal* – tvar *Šmejkal* má 348 dokladů, z toho 19 lematizováno pod slovesem, z nichž slovesný tvar je ve skutečnosti pouze jediný;

– *Šplíchal* – příjmení má 97 dokladů, z toho je 7 lematizováno pod slovesem;

– *Zdráhal* – tvar *Zdráhal* má 220 dokladů v korpusu, z toho je 196 lematizováno jako příjmení. Z nich jsou 3 ve skutečnosti doklady na sloveso; naopak z 24 tvarů lematizovaných jako tvary slovesa je ve skutečnosti 19 dokladů na příjmení;

C. Jak příjmení, tak i slovesný tvar mají poměrně vysokou frekvenci. Srov. např.

– *Běhal* – tvar *Běhal* má 66 dokladů, všechny jsou lematizovány pod slovesem, ale v 39 případech se jedná o příjmení;

– *Odložil* – tvar má 197 dokladů, 141 z nich je lematizováno jako příjmení, z toho 41 je tvary slovesa; z 56 tvarů lematizovaných pod slovesem je 5 dokladů na příjmení;

– *Vysloužil* – tvar má 71 dokladů, 56 je lematizováno jako jméno, 20 z nich jsou doklady na sloveso; z 15 dokladů lematizovaných pod slovesem jsou 4 tvary příjmení.

D. Slovesný tvar má poměrně vysokou funkci, příjmení poměrně nízkou, počet tvarů interpretovaných jako slovesných bývá nadsazen (někdy je tomu však naopak):

– *Nosil* – tvar má 86 dokladů, 4 na příjmení, všechny lematizovány pod slovesem;

– *Opravil* – tvar má 86 dokladů, všechny jsou lematizovány pod slovesem, ale v 71 případech jde přitom o příjmení.

– *Radil* – tvar má 51 dokladů, 8 na příjmení, všechny lematizovány jako příjmení; takto je lematizováno i 14 tvarů slovesa. 29 slovesných tvarů je správně lematizováno;

– *Soudil* – tvar má 27 dokladů, z toho 2 na příjmení, lematizovány pod slovesem;

– *Vrátil* – tvar má 659 dokladů, všechny jsou lematizovány pod slovesem, dokladů na příjmení jsme identifikovali 7;

E. Slovesný tvar má velmi vysokou frekvenci, takže nebylo možné ručně dohledávat výskyt příjmení. Srov. např.:

– *Čekal* – tvar má 579 dokladů, všechny jsou lematizovány pod slovesem, mezi prvními 500 doklady je 6 na příjmení (nositelů příjmení je podle CEO 319);

– *Dodal* – tvar *Dodal* má 2654 dokladů v korpusu, všechny lematizovány pod slovesem (příjmení má podle CEO 11 nositelů);

– *Mohl* – tvar má 1887 dokladů, všechny jsou lematizovány pod slovesem (příjmení má podle CEO 15 nositelů).

– *Musel* – tvar má 1475 dokladů, všechny jsou lematizovány pod slovesem (příjmení má podle CEO 90 nositelů);

F. V některých případech dochází k další homonymii mezi slovesným tvarem a apelativem nebo mezi dvěma příjmeními. Srov. např.:

– *Bouchal* – tvar *Bouchal* má 13 dokladů, 1 na sloveso, 12 na příjmení; všechny jsou lematizovány pod apelativem *bouchal*;

– *Přibyla* – tvar *Přibyla* má 133 dokladů, od příjmení *Přibyla* je jich 13; lematizovány jsou až na 1 pod příjmením *Přibyl*, 1 je lematizován pod slovesem; velkou část dokladů tvoří doklady na genitiv-akuzativ příjmení *Přibyl*.

– *Váhala* – tvar má 16 dokladů, 3 na příjmení *Váhala*, 2 jsou lematizovány pod slovesem *váhat*, 1 pod příjmením *Váhal*;

3. Produktivita deverbálních příjmení na *-l, -la, -lo* z hlediska slovesných tříd

Jak ukázal již Beneš (1962), deverbální příjmení na *-l, -la, -lo* se tvoří převážně od sloves 4. a 5. slovesné třídy[12)]. V rámci seznamu potenciálně homonymních sloves mají obě skupiny téměř stejný rozsah, kolem 280 příjmení z celkového počtu 700 homonymních příjmení. V rámci skupiny příjmení derivovaných od sloves 4. třídy můžeme vymezit podskupinu příjmení derivovaných od sloves typu B na *-el (Skácel, Sháněl)*. Někdy se vyskytují paralelní příjmení jako *Musil/Musel, Vláčil/Vláčel*, jež jsou dány výskytem konkurujících podob sloves. Apelativně regresivní typ na *-il* se ustaluje v příjmení, kdežto typ na *-el* dominuje u slovesa.

Ze zbývajících cca. 140 příjmení na *-l, -la, -lo* je cca. 30 derivováno od sloves 2. slovesné třídy (skoro výlučně typu A[13)]): např. *Koukl, Lízl, Klesla*. Jenom několika příjmeními je zastoupena 1. slovesná třída (bez typu B) – např. *Mohl, Pekl, Nesl* – a stejně tak 3. slovesná třída (skoro výlučně typ A), např. *Kryl, Hodoval, Pila*. Několika příjmeními je zastoupeno i nepravidelné sloveso *být* a jeho deriváty: např. *Byl, Přibyl, Nedobyl*. Nápadné je zejména, že v apelativní sféře velmi produktivní typ B 3. slovesné třídy (slovesa na *-ovat* s prézentním kmenem na *-uj-e-*) je zastoupen pouze příjmeními *Hodoval* a *Koštoval*.

U ostatních morfologických typů deverbálních anebo homonymních příjmení nelze pro nízký počet dokladů uvažovat o zastoupení slovesných tříd.

4. Tvary, v nichž k homonymii dochází

V rámci tvarosloví zkoumaného typu příjmení jsou homonymní vůči slovesným tvarům nominativní a genitivní podoba, v případě příjmení na *-l* a *-lo* i podoba akuzativní, formálně shodná s genitivní:

– nominativní podoba příjmení na *-l* – mužský rod sg. *l*-ového tvaru slovesa: *Odložil šel nakoupit – Odložil jsem kabát*;

[12)] Včetně několika příjmení od sloves typu B 1. slovesné třídy, která v apelativní slovní zásobě často kolísají s 5. slovesnou třídou (srov. Mluvnice češtiny 2, Praha 1986, s. 436n.), např. *Dlabal, Dřímal, Mazal*.

[13)] K typu C patří sloveso *začít*, od něhož je zastoupeno příjmení *Začal*.

– genitivní podoba příjmení na *-l* – ženský rod sg. / střední rod pl. *l*-ového tvaru slovesa: *Odložila si ani nevšiml – Odložila jsem kabát / Odložila* [děvčata] *kabát*;

– akuzativní podoba příjmení na *-l* – ženský rod sg. / střední rod pl. *l*-ového tvaru slovesa: *Odložila jsem neviděl – Odložila jsem kabát / Odložila* [děvčata] *kabát*;

– nominativní podoba příjmení na *-la* – ženský rod sg. / střední rod pl. *l*-ového tvaru slovesa: *Váhala tam nebyl – Váhala jsem dlouho / Váhala* [děvčata] *dlouho*;

– genitivní podoba příjmení na *-la* – ženský rod pl. *l*-ového tvaru slovesa: *Váhaly jsem si ani nevšiml - Váhaly* [přítelkyně] *dlouho*;

– nominativní podoba příjmení na *-lo* – střední rod sg. *l*-ového tvaru slovesa: *Obešlo šel nakoupit – Obešlo se to bez následků*;

– genitivní podoba příjmení na *-l* – ženský rod sg. / střední rod pl. *l*-ového tvaru slovesa: *Obešla jsem si ani nevšiml – Obešla jsem všechny / Obešla* [děvčata] *všechny*;

– akuzativní podoba příjmení na *-l* – ženský rod sg. / střední rod pl. *l*-ového tvaru slovesa: *Obešla jsem neviděl – Obešla mě hrůza*.

Dále může vzniknout homonymie mezi genitivem/akuzativem příjmení na *-l* a nominativem příjmení na *-la*, např. *Bouchala jsem neviděl. / Bouchala nepřišel*[14)].

5. Strategie pro řešení homonymie

Pro automatické rozpoznání příjmení homonymního se slovesným tvarem lze navrhnout elementy z několika různých strategií, které lze také spojit:

5. 1. První krok k řešení homonymie zde popsaných příjmení, ale i identifikace pojmenovaných entit vůbec, je založen na **ortografickém** principu: využívá se faktu, že jména se píší s počáteční majuskulí, čímž se liší od apelativní slovní zásoby. Pokud tedy potenciálně homonymní tvar je psán s počáteční minuskulí, pak lze vyloučit, že by se jednalo o jméno. Poněvadž ovšem majuskule stojí i na začátku věty, je třeba udělat několik dalších kroků: je nutné v případě tvaru psaného s majuskulí na začátku prohlédnout okolí: pokud nestojí bezprostředně před ním jedna ze skupin znaků tečka – mezera, otazník – mezera nebo vykřičník – mezera, pak se jedná o jméno. Pokud potenciálně homonymní tvar s majuskulí na začátku stojí bezprostředně po jedné z uvedených tří kombinací znaků, pak je nutné učinit další kroky[15)].

[14)] Z úzu je doložen výjimečně měkký tvar vokativu *pane Vykypěli*, homonymní s plurálovou formou *l*-ového tvaru slovesa (za informaci děkujeme Bohumilu Vykypělovi, Brno).

[15)] Je několik periferních případů, kde identifikace příjmení mimo začátek věty pouze na základě psaní s majuskulí není jednoznačná, např. když celá věta je psána v majuskulích: doc#178133 *ŘEDITEL OKRESNÍHO ŘEDITELSTVÍ POLICIE ČESKÉ REPUBLIKY RYCHNOV NAD KNĚŽNOU LUBOŠ DOSTÁL ŘEKL PRÁVU: [...]*. Dalším takovým případem je situace, kdy konec předcházející věty není označen tečkou (otazníkem, vykřičníkem), resp. kde do věty je vložena víceslovná pojmenovaná entita, obsahující apelativní slovesný tvar na začátku. To se týká zejména titulů, citátů nebo básní, kde se

Problém představuje používání tečky, které nesignalizuje konec věty, nýbrž zkratku uvnitř věty. Zde po tečce může stát minuskule a apelativní slovesný tvar se liší od jména. Lze proto vycházet z toho, že používání majuskule po tečce signalizující zkratku identifikuje potenciálně homonymní tvar jednoznačně jako onymický (pokud tečka nesignalizuje zároveň konec věty). Je tudíž nutné pracovat se seznamem zkratek, po nichž tečka obyčejně nesignalizuje konec věty, a následující homonymní tvar s majuskulí na začátku bude interpretován jako příjmení. Takové zkratky jsou zejména *p*. (pan), *kol*. (kolega), akademické tituly jako *mgr*. (magistr), *dr*. (doktor), *dra*.[16)] (doktora), *doc*. (docent), *prof*. (profesor), *PhDr*. (doktor filozofie), *MUDr*. (doktor medicíny), *JUDr*. (doktor práv), *ThDr*. (doktor teologie), *RNDr*. (doktor přírodních věd), běžné zkratky jako *např*. (například), *popř*. (popřípadě), *tj*. (to jest), *zejm*. (zejména) a další. Pro dané účely by se neměly do seznamu zahrnovat zkratky *atd*. (a tak dále), *aj*. (a jiní/é) a *apod*. (a podobně), protože stojí obvykle na konci věty či dílčí věty. Specifický subtyp tohoto problému představují iniciálky: velmi často homonymní tvar vystupuje ve spojeních typu *J. Musil, F. Navrátil, J. M. Navrátil*, kde je prakticky vyloučeno, že by iniciálka mohla tvořit konec předcházející věty. Je proto nutné počítat se zvláštní skupinou mezera – majuskule – tečka – mezera před poten-ciálně homonymním tvarem, po níž se tento tvar bude identifikovat jako příjmení.

Zvláštní problém představuje dvojtečka, protože po ní se používají jak majuskule, tak minuskule, a v případě přímé řeči se někdy používají uvozovky, jindy nikoliv[17)]. Tento fakt se odráží i v korpusu SYN2000: vystupuje zde přímá řeč s uvozovkami, přímá řeč bez uvozovek, dvojtečka bez přímé řeči, po níž následuje minuskule, a dvojtečka bez přímé řeči s následující majuskulí. Vzhledem ke kvantitativním poměrům bude nejlepší považovat skupinu dvojtečka – mezera – uvozovka za začátek věty (zacházet s ní stejně jako se skupinami tečka – mezera, otazník – mezera, vykřičník – mezera) a pouhou dvojtečku za začátek věty nepovažovat (zacházet se skupinou dvojtečka – mezera stejně jako se skupinou čárka – mezera či středník – mezera).

5. 2. Část řešení může být **lexikální**, tj. může se opírat o seznam relevantních příjmení. Vzhledem k tomu, že Svozilová et al. uvádějí ve svém slovníku 16 000 jednotek, z

označuje konec řádky lomítkem (konec věty se někdy označuje, někdy ne), srov. např. doc#196718 *Zazní i zpěv písní Skákal pes a Běží liška k Táboru*; doc.opus=refl9453 *„První čin můj samostatný / v Taverně měl konec špatný. / Musil jsem si sám být hostem, / banket zakončil jsem postem.“* (tvary *Skákal* a *Musil* budou interpretovány jako příjmení).

[16)] Citujeme (zde i dále) přesně podle korpusu, tedy i s pravopisnými a jinými chybami.

[17)] Srov. podrobně Pravidla českého pravopisu, Praha 1993, s. 50n.

nichž „největší rozsah mají slovesa“ (navíc existují i slovesa nevalenční)[18], lze pomocí seznamu udělat negativní výběr: seznam dovoluje určovat slovesné tvary, které nejsou homonymní s žádným doloženým příjmením. Je však třeba uvědomit si, že v CEO jsou pouze příjmení s minimálně 5 nositeli a že homonymní mohou být i cizí příjmení: srov. např. příjmení *Nadal*[19] nebo *Vadil* (ať je etymologicky jakéhokoliv původu)[20]. Dále se soubor českých příjmení homonymních se slovesným tvarem zejména v beletristice někdy rozšiřuje o příjmení, která reálně neexistují (např. u Viewegha příjmení *Naskočil* nedoložené ani u Beneše ani v CEO nebo ve filmu „Waterloo po česku“ příjmení *Rozvinul* a *Zavinul*). Využití seznamu tedy závisí na tom, zda se mají automaticky rozpoznat i dosud nedoložená příjmení.

5. 3. Rozhodující část řešení by měla být **syntaktická**, tj. měla by využívat různé syntaktické funkce slovesa a podstatného jména. Tato část je proto specifická pro vlastní jména homonymní vůči slovesným tvarům a nedá se aplikovat na vlastní jména homonymní vůči nominálním apelativům. Základním principem je hledání jiného finitního slovesného tvaru[21] v rámci (dílčí) věty. Hranice věty se určují interpunkčními znaménky před homonymním tvarem a po něm. Interpunkčními znaménky, jež stojí na konci věty, po níž následuje homonymní tvar, jsou tečka, otazník, vykřičník nebo dvojtečka – mezera – uvozovka (viz 5.1). Interpunkčními znaménky ukončujícími (dílčí) větu s homonymním tvarem na začátku jsou tečka, otazník, vykřičník, dvojtečka, středník, čárka. Jestliže se mezi znaménkem ukončujícím předcházející větu a znaménkem ukončujícím větu obsahující homonymní tvar nenajde další finitní slovesný tvar, pak je velmi pravděpodobné, že homonymní tvar je slovesný: doc#84 *Normana vzbudil kvílející poplachový signál a blikající červené světlo. Vyskočil z lehátka, hodil na sebe izolované boty a vyběhl ze dveří, kde se srazil s Beth.* Zde se tvar *Vyskočil* jako jediný finitní slovesný tvar mezi tečkou ukončující předcházející větu a čárkou ukončující dílčí větu, v níž stojí, identifikuje jako sloveso. Srov. obdobně doc#54 *Cítil dech děvčete. Přikryl ji dvěma dekami.* (tvar *Přikryl* bude identifikován jako slovesný). Oproti tomu bude homonymní tvar, který stojí na začátku (dílčí) věty s jiným finitním slovesným tvarem, určen jako příjmení: doc.opus=havel ... *nejoptimističtějších a nejnápaditějších lidí, které divadlo mělo.*

---

[18] N. Svozilová – H. Prouzová – A. Jirsová, Slovník slovesných, substantivních a adjektivních vazeb a spojení, Praha 2005.

[19] Z 25 dokladů pro tvar *Nadal* v SYN 2000 je jeden od slovesa *nadat*, ostatní jsou příjmení různých španělských sportovců. Všechny doklady jsou lematizovány pod slovesem *nadat*.

[20] Srov. dánského(?) fotbalistu Erharta Vadila pod http://www.ch-net.dk/managerzone/.

[21] Finitní tvary obsahují kategorii osoby a jsou tak schopné tvořit samostatně predikát věty (Mluvnice češtiny 2, o. c. v pozn. 12, s. 408). Infinitní tvary jsou infinitiv, příčestí a přechodníky.

Vyskočil ho poznal, když byl Vomáčka ještě barovým klavíristou v Olympii (tvar *Vyskočil* bude identifikován jako příjmení, protože se nachází v dílčí větě s jednoznačně slovesným finitním tvarem *poznal*). Srov. obdobně doc#135072 *případ nezákonného propojení podsvětí a státních úředníků. Přikryl totiž podal trestní oznámení na dva kriminalisty, kteří se ...* (tvar *Přikryl* bude identifikován jako příjmení na základě přítomnosti tvaru *podal*).

Vzniká zde však opět několik problémů: nejzávažnějším je přítomnost finitního slovesného tvaru spojeného s homonymním tvarem pomocí spojky *a*[22)] bez interpunkční-ho znaménka: doc.opus=des ... *smích - bylo na něm něco nezemského. Neslyšel ho už někdy? Vyskočil a otevřel dveře do haly, ale nepřicházelo to zdola* (tvar *Vyskočil* by se podle dosud nastíněného principu identifikoval jako příjmení na základě přítomnosti tvaru *otevřel*). Srov. obdobně doc#85 ... *časový údaj na monitoru. Ukazoval 1230 a pořád odpočítával. Přikryl Harryho dekou a šel k ovládacímu pultu.* (tvar *Přikryl* by se identifikoval jako příjmení kvůli tvaru *šel*). Zároveň není možné identifikovat homonymní tvar jako slovesný pouze na základě přítomnosti spojky *a*, protože ta může spojovat i nomina; je tedy žádoucí rozlišovat případy *Vyskočil a utekl pryč* (homonymní tvar je slovesný) a *Vyskočil a Novák utekli pryč* (homonymní tvar je jmenný). Bude proto třeba v případě, že se kromě potenciálně homonymního tvaru v (dílčí) větě nachází další finitní slovesný tvar, zahrnout kromě spojky *a* i další slovo: jedná-li se o sloveso, je vysoká pravděpodobnost, že slovesem je i homonymní tvar na začátku (dílčí) věty, jedná-li se o jméno, je pravděpodobné, že jménem bude i tvar stojící na začátku. Při takovém řešení zůstává „zbytek“ typu *Vyskočil Novák a Svoboda utekl pryč*, kde by se tvar *Vyskočil* nesprávně identifikoval jako příjmení (pokud se nepoužívá čárka, tedy *Vyskočil Novák, a Svoboda utekl pryč*), rovněž typ *Vyskočil a Novák utekl pryč* (pokud by se v takových případech nepoužívalo osobní zájmeno, popř. opět čárka tedy *On vyskočil(,) a Novák utekl pryč*). Tyto větné typy však budou kvantitativně marginální.

Pozornost zasluhuje výskyt tvaru 1. a 2. os. sg. pomocného slovesa *být* (auxiliár v préteritu), *jsem, jsi, jste* (v případě vykání) ve spojení s homonymním tvarem na *-l, -la, -lo* (u ostatních typů tento problém není): tyto tvary se nesmějí zohlednit při hledání dalšího finitního tvaru uvnitř (dílčí) věty s homonymním tvarem, jinak by se v případech jako doc.opus=mirak *Nemohl jsem to vydržet. Vyskočil jsem z ustlané postele, vztekle jsem praštil kelímkem, jenž ...* tvar *Vyskočil* interpretoval jako příjmení. Srov. obdobně doc#278 *Cítíš zápach?“ „Možná, nevím. K ránu jistě bude.“ „Přikryl jsem ji voskovým plátnem, dobrým americkým voskovaným plátnem ...* (tvar *Přikryl* by se nesprávně interpretoval jako příjmení). Tak bude sice systematicky nesprávně

[22)] Úloha jiných souřadicích spojek se zdá být podle našich konkordancí minimální.

interpretována věta s tvary sponového slovesa *být* po homonymním tvaru, je však jasné, že typ *Vyčítal jsem mu to* je oproti typu *Vyčítal jsem já* kvantitativně natolik v převaze, že vznikající chyba bude minimální. Jednoznačně sponové nespisovné tvary *(já) su* a *(ty) (j)seš* se vyloučit nemusejí (zejména druhý se těší v korpusu značné frekvenci). V případě tvarů pomocného slovesa analytického kondicionálu *bych, bys, by* je situace jednodušší; jejich výskyt je vázán na kondicionál, čili pokud se vyskytují v (dílčí) větě s homonymním tvarem na začátku a nevyskytuje se v ní další *l*-ový tvar (s minuskulí), homonymní tvar lze interpretovat jako slovesný.

Vyloučit z finitních tvarů určujících homonymní tvar jako příjmení se musí i tvar 3. os. sponového slovesa *být*, tedy *je*, pro jeho homonymii s akuzativem osobního zájmena 3. os. pl. Jedná se o problém, který daleko přesahuje naši problematiku, projevuje se však i zde: pokud by se tvar *je* stojící s homonymním tvarem v (dílčí) větě připouštěl jako určující tento tvar jako příjmení, pak by se v případě jako doc.opus=vrazlek *„Máte taky ráda rybky?“ „Koupil je pro mě,“ vyhnula se odpovědi, „bylo to od něj moc milé ...“* tvar *Koupil* interpretoval jako příjmení. Zároveň jeho nezohlednění vede k nesprávné interpretaci případů jako doc.opus=proste98 ... *vystavených grafik, kde dokáže využít rezerváže, akvatinty a lepty. Odehnal sám je již přes třicet pedagogem na Střední škole uměleckých řemesel v Brně* (tvar *Odehnal* by se interpretoval jako slovesný). Kvantitativní poměry budou velmi různé: v případě početného výskytu slovesa a marginálního výskytu příjmení (jako např. *Koupil*) se dosáhne dobrého výsledku, v případě příjmení s velkým počtem dokladů a zároveň nízké frekvence homonymního slovesného tvaru tomu tak nebude (v korpusu je třikrát na začátku věty spojení *Kasal je* a jedná se vždy o sponové sloveso *být*, ačkoliv tranzitivní sloveso *kasat něco* podle Slovníku spisovného jazyka českého existuje[23)]). Mnoho deverbálních příjmení je navíc odvozeno od intranzitivních sloves, kde není třeba počítat s akuzativním objektem. Lze ještě uvažovat o využívání slovosledového kritéria: jestli tvar *je* nestojí přímo po homonymním deverbálním tvaru, pak se zohledňuje jako tvar sponového slovesa. Tím by se ošetřily případy typu *Odehnal sám je již přes třicet pedagogem...*, ovšem nikoliv případy typu *Kasal je přesvědčen...*

Dalším problémem jsou tvary příjmení stojící samostatně mezi dvěma větnými znaménky. Lze rozlišovat různé případy: 1. Příjmení stojí před dvojtečkou, např. v titulech bez verba dicendi: doc#182533 *Čadil: Salcburk bude lehký soupeř.* 2. Příjmení stojí před čárkou, po níž sleduje vztažné zájmeno: doc#134494 ... *je po něm vyhlášeno celostátní pátrání. Přikryl, jehož firma má například licenci na vývoz vojenské techniky, byla agentem nejen speciálního útvaru* ... 3. Příjmení stojí po tečce, jíž předchází číslice, a před čárkou či tečkou (v shrnutích sportovních utkání):

[23)] Slovník spisovného jazyka českého I, Praha 1960, s. 847.

doc.opus=lnd94056 *Třinec – Znojmo 4:1 (0:1), branky: 51. Kříž, 70. Činčala, 74. Gajger, 88. Körner – 37. Osladil. Rozhodčí: Janata. Žluté karty: Kubeš, Kříž, Jančík – Samek, Stříž, Kovář. Divácí: 750.*, doc.opus=mf921109 *Branky: 30. Fryč, 75. Poborský – 28. Luhový. Rozhodčí: Velický. ŽK: Přibyl, Fryč, Vácha – Solár, Rehák, Forgáč. Diváků: 4105*. Podle dosud přijatých postupů by se tvary *Čadil, Přikryl, Osladil* a *Přibyl* identifikovaly jako slovesné. Nejméně problémů představuje situace 2: homonymní tvar stojící sám v (dílčí) větě před čárkou, po níž následuje tvar vztažného zájmena (*který, kterého, jenž, jehož* atd.), – buď bezprostředně anebo dělen jen předložkou – se bude interpretovat jako příjmení. Pro situaci 3 lze navrhnout, aby se homonymní tvary stojící mezi skupinou číslice-tečka-mezera a čárkou či tečkou interpretovaly jako příjmení. Tím by se zohlednila situace tvaru *Osladil* v uvedeném příkladě, ovšem nikoliv situace tvaru *Přibyl* (přesnost mechanismu by se dala vylepšit sestavením seznamu zkratek, které se v takových shrnutích sportovních utkání vyskytují). Řešení této problematiky se doporučuje kvůli vysoké frekvenci tohoto druhu textu v korpusu. Nejméně snadná je situace 1, protože ve stejné pozici mohou stát verba dicendi: doc#19 *konzervace uchylující se za života do umělého rámce, se vždycky vymluvil na nějaké tajemné zaneprázdněn. Říkal: jen pitomá sladká brambora sama dodává liánu, která ji zadusí*. Vzhledem k tomu, že počet verb dicendi mezi slovesy, od nichž jsou odvozená doložená deverbální příjmení, je velmi nízký, se nabízí řešení, že se v případě situace 1 homonymní tvar interpretuje jako příjmení, pokud se nachází na seznamu navrženém v 5.2., s výjimkou tvarů odvozených od sloves dicendi, která by se určila ve zvláštním seznamu *(Říkal, Radil, Myslil...)*.

Je tu ještě jeden problém: stojí-li homonymní tvar mezi tečkou a čárkou, čárka nemusí označovat vždy konec dílčí věty, ale může též oddělovat několikanásobné větné členy nebo může signalizovat apozici, vsuvku, vloženou větu apod., např. *Vyskočil z Prahy, Novák z Brna a Kučera z Olomouce se bohužel setkání nemohli zúčastnit.* Nebo: *Vyskočil v domnění, že už má málo času, utíkal, jak nejrychleji mohl.* V takových případech by zvýšit pravděpodobnost úspěšné disambiguace mohla nanejvýš ještě textová strategie (viz 5.4.).

5. 4. Jenom náznakem uvádíme další strategie, totiž strategii textovou a strategii kvantitativní. Strategie textová by spočívala v hledání jednoznačně nominálních tvarů příjmení v (nadvětném) kontextu homonymního tvaru, tedy tvar *Vybíral* by se na základě výskytu tvarů jako *Vybíralovi, Vybíralem* v kontextu identifikoval jako příjmení. Strategie kvantitativní by využívala pravděpodobnost výskytu tvaru příjmení oproti tvaru slovesnému. Využívání těchto strategií by mělo smysl pouze tehdy, pokud by se dala s ostatními zde nastíněnými strategiemi kombinovat.

6. Závěr

Představili jsme problémy spojené s homonymii českých deverbálních příjmení při automatickém anotování v korpusu. Navrhli jsme řešení pro vylepšení, která se zakládají na ortografických a interpunkčních principech, na seznamu odpovídajících

příjmení a pak zejména na syntaktických vlastnostech jmen a příjmení. Úspěšnost zde navržených kroků při správné disambiguaci homonymních deverbálních tvarů bude (v závislosti na technickém řešení) dost vysoká. Prošli jsme řadu konkordancí (*Nežádal, Dodal, Běhal, Prchal, Trkal, Kasal* aj.) a můžeme odhadnout úspěšnost na 95% – 99%.

Summary

Czech Deverbal Surnames and the Problem of Their Homonymy in Electronic Corpora

The article discusses the problems connected with the homonymy of Czech deverbal surnames which must be solved during automatic annotation in electronic corpora. Solutions for improvement of these procedures, based on orthographic and punctuation criteria, on the list of corresponding surnames, and especially on syntactic features of surnames, are suggested in the paper. The degree of success of the steps proposed for disambiguation of the surnames will be very high; it can be estimated around 95 – 99 %.