Problémy desambiguace proprií a apelativ se sufixem -ář

Na datech korpusu syn v8

Jaro 2020

Štěpán Mach (UČO 333368)

Východiska pro sběr dat

- Všechna slova s velkým písmenem a se sufixem -ář
  - Dotaz na lemma: *"[AÁBCČDĎEÉFGHIÍJKLMNŇOÓPQRŘSŠTŤUÚVWXYÝZŽ].*ář"*
  - Celkem 1434 položek, zohledněno nejčastějších 50 (50. má 513 výskytů)
    - Blíže zkoumána propria s alespoň 1000 výskyty
- Homonyma s apelativy
  - Vybráno 27 jmen, která se pravděpodobně vyskytují i jako apelativa
  - Příjmení: Kolář, Kovář, Šindelář, …
  - Jiná: Hraničář, Olejář

Východiska pro sběr dat

- Apelativa s proprii sdílejí velké písmeno na začátku věty
  - Tyto případy netřeba blíže zkoumat
  - Dotaz na slovní tvar s verzálkou musí odfiltrovat první slova ve větách
    - Např.: *[word=",|;|-" | tag!="Z.*"][word="Kolář.*" & lemma="Kolář"]*
    - Dotaz na dvojici tokenů, z nichž první není interpunkce, nebo je taková, která nezakončuje/neuvozuje větu (čárka, středník, pomlčka).
    - Výhoda: Pracuje s bezprostředním levým kontextem, odhaluje možné víceslovné entity
    - Nevýhoda: Není 100% účinný, předcházející text často bývá nevětného charakteru (např. nadpis)
    - Levý kontext ignorován v případě nulové shody => rozšíření vyhledávání

Východiska pro sběr dat

- U každého jména použity minimálně tři typy dotazů na druhý token:
  - *[word="Kolář.*" & lemma="Kolář"]*
    [word="Kolář." & lemma="kolář"]*
    [word="kolář." & lemma="kolář"]* – levý kontext nechán, není nutný
    - Navíc odfiltrují nezpracovatelné zápisy verzálkami (*MICHAL KOLÁŘ*)
- Kontrola lemmat dvojic
  - Nejčastější užití, jednoduchý kontext, sémantika
  - Např.: *Jan Kolář*, *útočník Kolář*, *umělecký kovář*…
- Kontrola výsledků v širším kontextu
  - Méně časté páry slov, slova mimo MWE
  - Větší frekvence (10^5) => kontrola namátkou

Kolář (109 064 výskytů)

- Slovní tvar i lemma *Kolář* (106 351 výskytů)
  - V levém kontextu takřka výhradně křestní jména (*Petr* 11 000×), tituly/funkce (*pan, prezident, doktor, …*), uvozovací slovesa (*říci, uvést, dodat, …*)
  - Desambiguaci lze považovat za korektní
- Slovní tvar *Kolář*, lemma *kolář* (1 020 výskytů)
  - Podobný levý kontext: křestní jméno (1., 6.–9.), oslovení (2.)
    - Nejčastěji za slovem *Jožou* (40×, tag *X* => není ve slovníku)
  - V názvech firem (*Rybářství Kolář*)
  - Desambiguováno špatně

Kolář – pokračování

- Slovní tvar i lemma *kolář* (1 665 výskytů)
  - Nejčastěji za čárkou, verbem *být*, spojkou *a* – neprůkazné
  - Na dalších pozicích (4.–6.) *vyučený, mistr, rodina* či *zručný* (10.)
  - Chybně napsané příjmení (10× *Jiří kolář*)
- Srovnání frekvencí slovních tvarů (uprostřed věty)
  - *Kolář* (> 100 000 výskytů) vs. *kolář* (< 2 000 výskytů)
  - Slovo se min. 50× častěji používá jako vlastní jméno než jako řemeslo
  - Desambiguace slovního tvaru *Kolář* vždy jako propria se jeví výhodnou

Kovář (77 559 výskytů)

- Slovní tvar i lemma *Kovář* (72 589 výskytů)
  - Ambivalentní užití (kovář s příjmením Kovář, přezdívka):
    - *Richard Kovář má na celou zimu práci v teple, dělá tyče, skoby a čepy pro klenbu.*
    - *Zalman Kovář nebyl v kování mečů a kopí příliš zručný.*
    - *Několik jeho přátel, zejména Kovář a Harfeník, nad jeho odchodem truchlilo…*
  - Člen MWE
    - *… hlavní osoba z povídky Gottfrieda Kellera Kovář svého štěstí…*
    - *… piloval svou novou povídku Kovář z Wooton Major.*

Kovář – pokračování

- Slovní tvar *Kovář*, lemma *kovář* (490 výskytů)
  - Správně: neodhalitelné první slovo věty
    - *Přiznali se i bez mučení Kovář vyrobil mučidla ze středověku.*
  - Chybně: vlastní jméno
    - Lidé: *Plzni doma pomohlo útočné eso Jan Kovář...*
    - Firmy: *… řeznictví Kovář plus z Podolí u Kunovic.*
- Slovní tvar i lemma *kovář* (33 247 výskytů)
  - Nejčastější: *umělecký kovář* (7 531 výskytů)
  - Zajímavé MWE: *hřib kovář* (4.), pohádka *O statečném kováři* (11.)
- Srovnání frekvencí slovních tvarů
  - *Kovář* (cca 72 000×) vs. *kovář* (cca 33 000×)
  - Upřednostňovat proprium se nevyplatí, 1/3 výskytů je obecné jméno

Bednář (54 683 výskytů)

- Slovní tvar i lemma *Bednář* (52 624 výskytů)
  - Název s předložkou
    - *… jeden útulný bar (říkalo se tam „U Bednářů“)…*
    - Nemusí jít sémanticky o proprium, je ho však nutné tak brát
- Slovní tvar *Bednář*, lemma *bednář* (589 výskytů)
  - Nejčastější spojení: *dvojice* (45×), *duo* (29×)
  - Ve všech případech pozorováno jako proprium
- Slovní tvar i lemma *bednář* (1 917 výskytů)
- Srovnání frekvencí slovních tvarů
  - *Bednář* (cca 53 000×) vs. *bednář* (cca 2 000×) => proprium 26× častější

Šindelář (28 662 výskytů)

- Slovní tvar i lemma *Šindelář* (27 974 výskytů)
- Slovní tvar *Šindelář*, lemma *šindelář* (235 výskytů)
  - Vždy proprium (např. po slovech *dvojice/duo, Petr, „Dědek“*)
  - Apelativum snad jen ve větě: *Kniha purgkrechtní Šindeláři a všichni, kteří z borů užitku mají, dávají od 100 kop 12 kop.*
    - Jediný výskyt slova *purgkrechtní* (zdroj *Obecní radní Stoklasné Lhoty vydraživší za 37 Kč vycpaného jezevce pro potřeby školního kabinetu*)
- Slovní tvar i lemma *šindelář* (293 výskytů)
  - Nejčastěji (155×) ve výčtu řemesel, pak jako špatně napsané proprium (4. a 6.)
- Srovnání frekvencí slovních tvarů
  - Šindelář (asi 28 000×) vs. šindelář (< 300×) => proprium skoro 100× častější

Mlynář (22 765 výskytů)

- Slovní tvar i lemma *Mlynář* (22 022 výskytů)
  - Předložkové názvy (*hospoda U Mlynáře*), názvy uměleckých děl (*Mlynář čaroděj, taškář a dohazovač Alexandra Onisimoviče Ablesimova*)
- Slovní tvar *Mlynář*, lemma *mlynář* (273 výskytů)
  - Také názvy uměleckých děl (*Mlynář a jeho dítě*, 42×)
- Slovní tvar i lemma *mlynář* (9 460 výskytů)
- Srovnání frekvencí slovních tvarů
  - Mlynář (asi 22 300×) vs. mlynář (asi 9 500) => cca třetina výskytů apelativum

Sklenář (18 663 výskytů)

- Slovní tvar i lemma *Sklenář* (18 326 výskytů)
- Slovní tvar *Sklenář*, lemma *sklenář* (223 výskytů)
  - 5× řemeslo, po větě bez interpunkce nebo mimo větu
    (*Umělecký truhlář, řezbář 15 945 Kč Sklenář 15 526 Kč Zámečník 15 451 Kč*)
- Slovní tvar i lemma *sklenář* (1 387 výskytů)
- Srovnání frekvencí slovních tvarů
  - *Sklenář* (cca 18 500×) vs. *sklenář* (cca 1 400×) => proprium 13× častější

Sklář (3 308 výskytů)

- Slovní tvar i lemma *Sklář* (2 791 výskytů)
  - Seriálový *Jakub Sklář* (4.), 187 z 577 výskytů MWE => ve 32 % chybně zapsáno (Jakub, povoláním sklář, je příjmením Cirkl), spojení *obec Skláře* (54×)
    - Minimálně 5% chyba na vstupu / v desambiguaci
- Slovní tvar *Sklář*, lemma *sklář* (458 výskytů)
  - Mimo začátek věty 27× jako kulturní dům, hotel, středisko, restaurace, penzion, 9× příjmení, 1× značka piva
  - V ostatních případech 62× sportovní tým (*Skláři díky gólům Vachouška a Vůcha slavili na stadionu Liberce.*), většinou pak řemeslo
- Slovní tvar i lemma *sklář* (22 871 výskytů)
- Srovnání frekvencí slovních tvarů
  - *Sklář* (2 587) vs. *sklář* (23 292) => apelativum 9× častější
    - Bez 5% chyby, s přičtením *Sklář.*/sklář* dle skutečnosti

Neckář (16 577 výskytů)

- Slovní tvar i lemma *Neckář* (15 986 výskytů)
  - 9877× *Václav/Vašek/V.*, 2. *Stanislav* (1 085×), 3. *Jan* (725×)
- Lemma *neckář* není ve slovníku
  - Pro apelativum používá korpus tvar neckař
  - Slovní tvar *Neckář*, lemma *neckař* v korpusu 1 074×
- Slovní tvar *neckář(e|i|ovi|em|ů|ům|ích)?* (17× celkem)
  - Vždy lemma *neckař*, 9× řemeslo, 8× chybně příjmení)
- Apelativa *neckář/neckař* ve slovnících
  - Chybí v SSČ, SSJČ, IJP, jen *neckář* je v PSJČ

Rybář (14 357 výskytů)

- Slovní tvar i lemma *Rybář* (12 515 výskytů)
  - Oikonymum *Rybáře* (např. *čtvrt/čtvrť* (323×), *část* (215×), *osada* (115×))
  - Drama *Rybář*, film *Král Rybář* (75×, *Král rybář* 235× => 24 %)
- Slovní tvar *Rybář*, lemma *rybář* (2 220 výskytů)
  - Nejčastěji *do/z Rybář* (1. a 2., 249×)
    - Pravidlo pro shodu v předložkovém pádě
  - Dále *ryba* (3., 52), *výlov* (6., 27), *rybník*, *kapr*, *voda*, …
- Slovní tvar i lemma *rybář* (101 537 výskytů)
- Srovnání frekvencí slovních tvarů
  - Apelativum cca 8× častější

Korytář (11 124 výskytů)

- Slovní tvar i lemma *Korytář* (10 941 výskytů)
  - *Jan* (1., 4501×), *Radomír* (2., 887×)
- Slovní tvar *Korytář*, lemma *korytář* (57 výskytů)
  - Vždy příjmení, někdy na začátku věty, ne však pravidlem
- Slovní tvar i lemma *korytář* (111 výskytů, 121 celkem)
  - Posun významu **řemeslo > politik** (*Jsou to korytáři, zkorumpovaní darmošlapové, kteří slouží sami sobě.*)
- Srovnání frekvencí slovních tvarů
  - Proprium 90× častější

Rychtář (10 710 výskytů)

- Slovní tvar i lemma *Rychtář* (10 333 výskytů)
  - *Josef* (1., 2 915×), *pivo/pivovar* (2., 3., 1 410×)
- Slovní tvar *Rychtář*, lemma *rychtář* (441 výskytů)
  - *pivovar/pivo* (1., 2., 116×), *amfiteátr* (3., 28×), *Josef* (7., 6×)
  - MWE *motorest U Rychtáře*
- Slovní tvar i lemma *rychtář* (5 575 výskytů)
  - Kolokace s pivem jen 1×, jinak zpravidla funkce
- Srovnání frekvencí slovních tvarů
  - Proprium >2× častější
  - Pravidla pro psaní značek (např. aut) – jsou data s pivem Rychtář spolehlivá?

Cihlář (10 108 výskytů)

- Slovní tvar i lemma *Cihlář* (9 763 výskytů)
  - Příjmení, název rybníka, jméno postavy
- Slovní tvar *Cihlář*, lemma *cihlář* (103 výskytů)
  - Pouze 2× apelativum
- Nepravidelné desambiguace
  - Jméno postavy z díla Konec masopustu (*I před Tajemníkem a Předsedou se Cihlář snaží Krále hájit.*) – 72× lemma *Cihlář*, 27× *cihlář*
  - Název rybníka (kolokace -10 *rybník*) – 245× *Cihlář*, 10x *cihlář*
- Slovní tvar i lemma *cihlář* (414 výskytů)
- Srovnání frekvencí slovních tvarů
  - Proprium 23× častější

Bečvář (9 974 výskytů)

- Slovní tvar i lemma *Bečvář* (9 727 výskytů)
- Slovní tvar *Bečvář*, lemma *bečvář* (93 výskytů)
  - Jen příjmení, též chybný genitiv obce Bečváry (*Pojďme se do Bečvář ještě naposledy vrátit.*)
- Slovní tvar i lemma *bečvář* (97 výskytů)
  - Příjmení s minuskulí 2×
- Srovnání frekvencí slovních tvarů
  - Proprium >100× častější

Šafář (8 508 výskytů)

- Slovní tvar i lemma *Šafář* (8 281 výskytů)
- Slovní tvar *Šafář*, lemma *šafář* (55 výskytů)
  - Vždy příjmení
- Slovní tvar i lemma *šafář* (800 výskytů)
- Srovnání frekvencí slovních tvarů
  - Proprium cca 10× častější

Hraničář (8 471 výskytů)

- Slovní tvar i lemma *Hraničář* (6 479 výskytů)
  - Kino v Ústí nad Labem, divadelní soubor, ZD v Loděnicích, týdeník, kopec
  - Fantasy literatura (Tolkien) – diskutabilní, jde vlastně o povolání (*Představili se jako Mablung a Damrod, gondorští vojáci a Hraničáři z Ithilienu.*)
- Slovní tvar *Hraničář*, lemma *hraničář* (501 výskytů)
  - Nejčastěji kino (156×), kopec, sportovní tým (*TJ MTG Hraničář Cheb*)
- Slovní tvar i lemma *hraničář* (740 výskytů)
- Srovnání frekvencí slovních tvarů
  - Tvar propria 9× častější

Sedlář (6 387 výskytů)

- Slovní tvar i lemma *Sedlář* (6 256 výskytů)
  - První slovo v nápisu (*… s opršelým nápisem Sedlář a řemenář.*)
- Slovní tvar *Sedlář*, lemma *sedlář* (53 výskytů)
  - Vždy příjmení
- Slovní tvar i lemma *sedlář* (1 125 výskytů)
- Srovnání frekvencí slovních tvarů
  - Proprium >5× častější

Punčochář (4 617 výskytů)

- Slovní tvar i lemma *Punčochář* (4 217 výskytů)
- Slovní tvar *Punčochář*, lemma *punčochář* (43 výskytů)
  - Vždy příjmení
- Slovní tvar i lemma *punčochář* (164 výskytů)
  - 1× chyba v příjmení, jinak řemeslo
- Srovnání frekvencí slovních tvarů
  - Proprium 26× častější

Bakalář (4 494 výskytů)

- Slovní tvar i lemma *Bakalář* (4 300 výskytů)
  - Seriál *Bakaláři*, lokalita *Na Bakalářích*, příjmení, značka piva, školní IS, …
- Slovní tvar *Bakalář*, lemma *bakalář* (126 výskytů)
  - Také pivo, IS, lokalita, příjmení, ale i titul (… *obdržela diplom Bakalář ošetřovatelství.*)
- Slovní tvar i lemma *bakalář* (6 627 výskytů)
  - Skoro výhradně titul, ojediněle pivo (*I nealkoholický bakalář však není zcela bez alkoholu.*)
- Srovnání frekvencí slovních tvarů
  - Apelativum jen o trochu (1,5×) častější

Truhlář (4 191 výskytů)

- Slovní tvar i lemma *Truhlář* (3 951 výskytů)
  - Diskutabilně obor (*… stáže pro učební obory Truhlář, Čalouník a Aranžér.*), titul *Truhlář roku* (57×)
- Slovní tvar *Truhlář*, lemma *truhlář* (194 výskytů)
  - *Truhlář roku* (9×)
- Slovní tvar i lemma *truhlář* (18 832 výskytů)
- Srovnání frekvencí slovních tvarů
  - Apelativum >4× častější

Nebesář (2 929 výskytů)

- Slovní tvar i lemma *Nebesář* (2 866 výskytů)
  - *Milan* (1 702×), vzácně postava z knihy *O Nebesáři*
- Slovní tvar *Nebesář* lemma *nebesář* (4×, celkem 10×)
  - Vždy příjmení
- Slovní tvar i lemma *nebesář* (celkem 5×)
  - Druh rybníka (3×), který je závislý na srážkové vodě, pohádka *O nebesáři*
- Srovnání frekvencí slovních tvarů
  - Apelativum se takřka nevyskytuje (poměr 1 : 575)

Vačkář (2 582 výskytů)

- Slovní tvar i lemma *Vačkář* (2 437 výskytů)
- Slovní tvar *Vačkář*, lemma *vačkář* (nevyskytuje se)
- Slovní tvar i lemma *vačkář* (4 výskyty)
  - Chybně napsané příjmení 2×
  - Řemeslo 2× (*… tu pracoval například i vačkář – výrobce váčků…*), není ve slovníku
- Srovnání frekvencí slovních tvarů
  - Apelativum jediné v celém korpusu

Krytinář (2 147 výskytů)

- Slovní tvar i lemma *Krytinář* (2 001 výskytů)
  - *Jiří/Jirka* (1524×)
- Slovní tvar *Krytinář*, lemma *krytinář* (nevyskytuje se)
- Slovní tvar i lemma *krytinář* (1 výskyt)
  - Zřejmě řemeslo (*Kamil Jeřábek, 24 let, krytinář, Zlín*)
  - Apelativum není ve slovníku (tag *X*)
- Srovnání frekvencí slovních tvarů
  - Apelativum jediné v celém korpusu

Mydlář (1 927 výskytů)

- Slovní tvar i lemma *Mydlář* (1 882 výskytů)
  - Takřka výhradně *kat* (1 018×) *Jan* (269×) Mydlář
- Slovní tvar *Mydlář*, lemma *mydlář* (41 výskytů)
  - Také *kat* (35×)
- Slovní tvar i lemma *mydlář* (408 výskytů)
  - 10× *Kat mydlář*
- Srovnání frekvencí slovních tvarů
  - Proprium skoro 5× častější, min. 57 % případů souvislost s katem Mydlářem

Popelář (1 757 výskytů)

- Slovní tvar i lemma *Popelář* (1 477 výskytů)
  - Příjmení, představení *Popelář přichází* či *Popeláři*
- Slovní tvar *Popelář*, lemma *popelář* (151 výskytů)
  - Většinou apelativa, někdy i propria na začátku věty
    - *Popelář* (proprium) *se stal finančním ředitelem.*
    - *Popelář* (apelativum) *má dům se 13 pokoji!*
- Slovní tvar i lemma *popelář* (8 259 výskytů, celkem >12 500)
- Srovnání frekvencí slovních tvarů
  - Apelativum >5× častější

Korbelář (1 427 výskytů)

- Slovní tvar i lemma *Korbelář* (1 366 výskytů)
- Slovní tvar *Korbelář*, lemma *korbelář* (nevyskytuje se)
- Slovní tvar i lemma *korbelář* (celkem 7 výskytů)
  - Lemma jen jedno, chybně napsané příjmení, slovo není ve slovníku (tag *X*)
  - Ostatní slovní tvary 6× (*… dnes se podíváme za mistry bednáři, bečváři a korbeláři.*)
- Srovnání frekvencí slovních tvarů
  - Apelativum není ve slovníku, poměr 1 : 227

Rybnikář vs. Rybníkář

Rybnikář

- Lemma pouze proprium (847×)
- Apelativa (64×) lemmatizována „dlouze“
  [word="rybnikář(e|i|ovi|em|ů|ích)?"]

Rybníkář

- Lemma jen apelativum (4 065×)
- Všechna propria (1 150×) lemmatizována jako apelativa
  [word=",|;|-" | tag!="Z."]*
  [word="Rybníkář(e|i|ovi|em|ů|ích)?"]

Internetová jazyková příručka připouští obě varianty.

Olejář (727 výskytů)

- Slovní tvar i lemma *Olejář* (681 výskytů)
  - Zpravidla překlad pro Edmonton Oilers, též přezdívka pro HC Chemopetrol Litvínov a HC Orlová (v r. 2010 HC Plus Oil Orlová)
- Slovní tvar *Olejář*, lemma *olejář* (28 výskytů)
  - Hokejisté, těžaři (na začátku věty), ale 2× i příjmení (*Okresní soud v Lounech se mezitím zabýval žalobou paní Olejářové na náhradu škody.*)
    - V ČR 6× příjmení Olejář, 3× Olejářová (v r. 2016) – Mostecko, Chomutovsko
- Slovní tvar i lemma *olejář* (473 výskytů)
  - Těžaři i hokejisté
- Srovnání frekvencí slovních tvarů
  - Proprium častější, kromě příjmení však nestandardní

Shrnutí

- Zdroje problémů s desambiguací apelativ a proprií
  - Úzus, autorova znalost jazyka a světa, záměr
    - *[Oo]lejář*, *Jakub [Ss]klář*, *Král [Rr]ybář*, jména a přezdívky lit. postav (*Předseda*, *Cihlář*)
  - Značkování
    - Začátek věty, nevětné fragmenty (seznamy, tabulky, nadpisy)
  - Názvy, MWE
    - Názvy uměleckých děl, spolků, firem apod.
  - Jazyk
    - Pravopis předložkových názvů (*U Rychtáře*, *Na Bakalářích*), značky vs. jednotliviny (*pivo Rychtář* vs. *zajít si na dva rychtáře*), titulků, nápisů a cedulí
    - Homonymie (*Rybář* vs. *Rybáře*, *Bečvář* vs. *Bečváry*)
  - Pokrytí slovníku
    - Dublety *rybníkář/rybnikář*, chybějící *korbelář*, *vačkář*, *neckář*

Shrnutí – pokračování

- Možná částečná řešení
  - Doplnění slovníku (*rybnikář/rybníkář*, *neckář*, …)
  - Reflektování valence předložek (*do/z Rybář*)
  - Výběrově lemmatizovat všechny tvary s velkým písmenem jako propria
    - U statisticky významných (např. *Bečvář*, *Šindelář* 100 : 1)
    - Sjednocení lemmatizace
- Další, složitější řešení
  - Rozpoznávání víceslovných entit
  - Identifikace hranic vět

Zdroje

- Český národní korpus
  - Webové rozhraní *KonText*, korpus *syn v8*
- DEBDict
  - Slovníky SSČ, SSJČ, PSJČ – heslo *neckář/neckař*
- Internetová jazyková příručka
  - Hesla *neckář/neckař*, *rybníkář/rybnikář*
- Webová aplikace KdeJsme.cz
  - Četnost příjmení *Olejář/Olejářová*

Děkuji Vám za pozornost.