Jiří Berka, učo 360766

CJBB84 - Morfologie a korpus

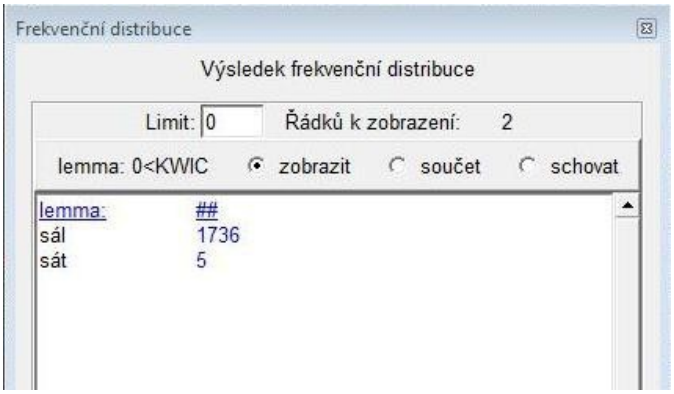
Úkol č. 1: **Praktické problémy lemmatizace**

V korpusu SYN2010 hledáme případy chybné lemmatizace slovního tvaru "sál", tj. takové, ve kterých byl nesprávně označkován coby substantivum a kdy se ve skutečnosti jedná o tvar slovesa "sát" v minulém čase (příp. kondicionálu).

I.

Nejprve vyhledáme všechny doklady slovního tvaru "sál" v korpusu.

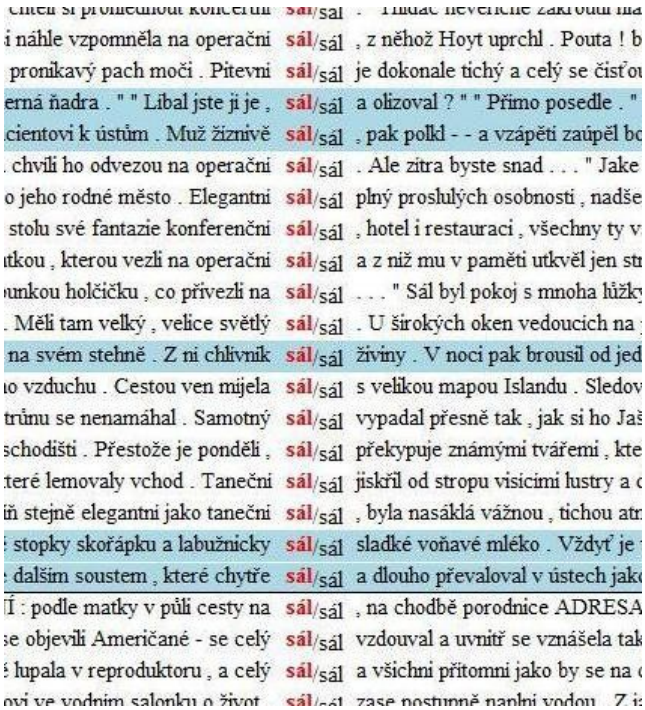
```
Query: "sál"
Počet výskytů: 1741
```



Negativním filtrem odstraníme výskyty, kdy se jedná o správně provedenou lemmatizaci k základnímu tvaru "sát" (celkem pět případů, viz obrázek frekvenční distribuce výše).

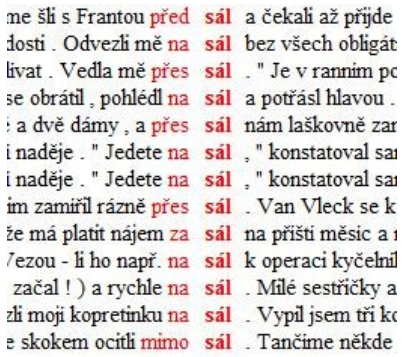
```
N filter: [lemma="sát"]

Počet výskytů: 1736
```

Zbývá 1736 řadků, z nichž však, jak vidíme po bližším prozkoumání, mnohé nejsou správně lemmatizovány:



Tyto chyby se nyní pokusíme izolovat odstraněním případů, kdy lemmatizace proběhla správně, tj. kdy je slovní tvar "sál" prokazatelně substantivem. To lze tvrdit, pokud:

- tvaru "sál" bezprostředně předchází předložka:

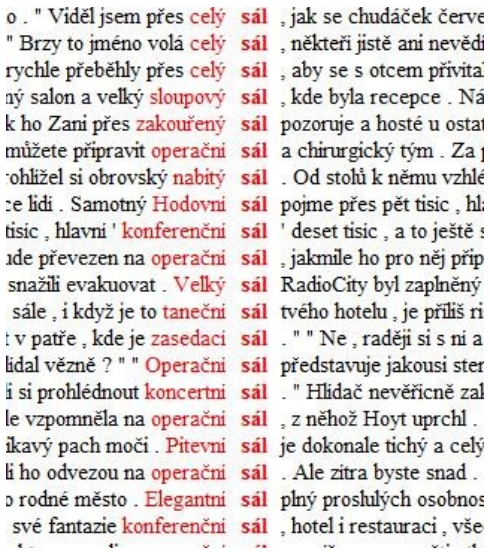


<-1, -1>.

N filter: [pos="R"]

Počet výskytů: 1660

- tvaru "sál" bezprostředně předchází adjektivum:



<-1, -1>

N filter: [pos="A"]

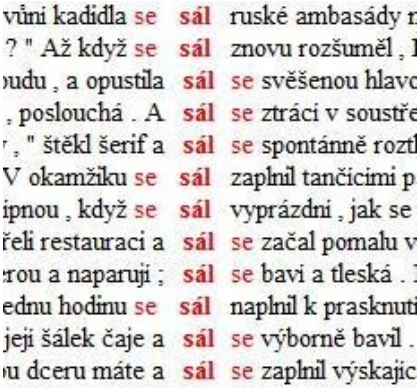
Počet výskytů: 471

- tvaru "sál" předcházejí některé další slovní druhy, např. některé typy zájmen (ukazovací a rovněž zvratné zájmeno "se"):

<-1, 1>

N filter: [tag="P[DLSWZ].*"]

Počet výskytů: 441



<-1, 1>

N filter: [word="se"]

Počet výskytů: 415

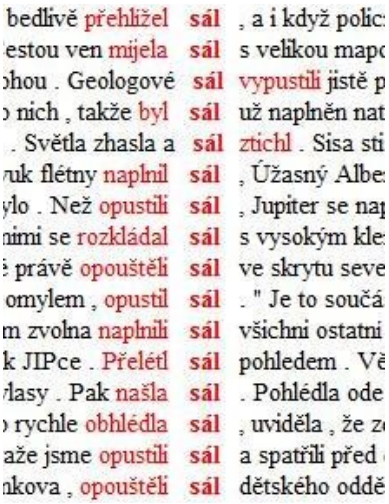
- tvaru sál předchází nebo po něm bezprostředně následuje číslovka:


```
<-1, 1>

N filter: [pos="C"]

Počet výskytů: 397
```

Poznámka: Hypoteticky by v některých případech adjektivum, zájmeno nebo číslovka mohly stát v roli podmětu a tehdy by slovní tvar "sál" byl slovesem (uvažme: *žíznivý sál vodu, tento sál vodu, druhý sál vodu*). Podobně nelze s jistotou tvrdit, že číslovka následující za tvarem jasně indikuje substantivum (hypoteticky: *sál 102 hroznů* apod.). Ruční kontrolou nicméně dospějeme k závěru, že korpus takovéto případy neobsahuje, a proto můžeme výše uvedená pravidla pro naše účely využít.

II.

Výše uvedenými kroky jsme zredukovali počet řádků výstupu na 397. Dále obraťme pozornost ke slovesům v kontextu. Vzhledem k tomu, že česká věta obsahuje pouze jeden tvar verba finita, můžeme odstranit výskyty, kdy tvaru "sál" předchází či po něm bezprostředně následuje infinitiv nebo určitý tvar slovesný (imperativ, příčestí minulé). Nechceme se však zbavit tvarů jako např. *sál jsem (vodu)*, do seznamu eliminovaných tvarů proto zatím nezahrneme slovesné tvary přítomného času (a tudíž zároveň budoucího, vzhledem k tomu, že pro ně korpusový manažer Bonito má stejnou značku).

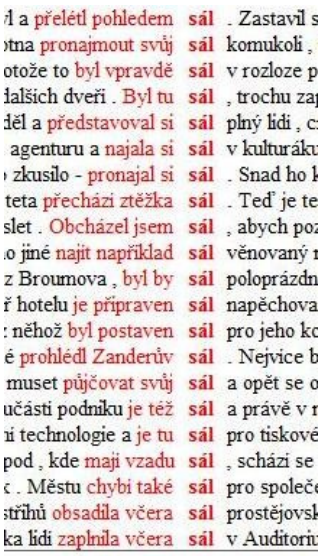



```
<-1, 1>

N filter: [tag="V[fip].*"]

Počet výskytů: 248
```

Nyní můžeme odstranit i výskyty se slovesným tvarem prezénsu nebo futura, abychom se však nezbavili tvarů jako *sál jsem* apod., odstraníme pouze tvary ve 3. osobě aktiva.

```
<-1, 1>

N filter: [tag="VB-.---3.-.A.*"]

Počet výskytů: 190
```

Dále se pokusíme zabrousit do hlubšího kontextu a vyřadit případy, kdy se infinitiv nebo slovesný tvar určitý nacházejí v delší vzdálenosti od centrálního tvaru než 1. Abychom se při tom pohybovali v kontextu jedné věty a také abychom nevyřadili koordinace typu *líbal a sál,* musíme zajistit, že se mezi centrálním tvarem a slovesným tvarem nebude nacházet konjunkce či interpunkční znaménko. Začněme v nejtěsnějším kontextu na levé straně, tj. v intervalu <-2, -1>.




```
<-2, -1>

N filter: [tag="V.*"] [tag="[ACDINPRTVX].*"]

Počet výskytů: 169
```

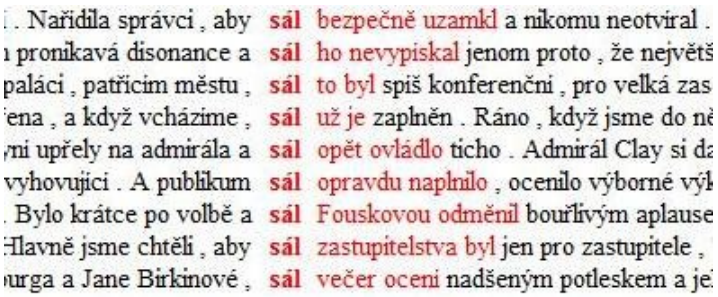
Zajděme nyní v kontextu ještě dál doleva a aplikujeme ten samý filtr v intervalu -3, -1. (Do našeho dotazu musíme přidat navíc jednu pozici, neshodnou s konjunkcí či interpunkčním znaménkem.)

<-3, -1>

N filter: [tag="V.*"] [tag="[ACDINPRTVX].*"] [tag="[ACDINPRTVX].*"]

Počet výskytů: 156

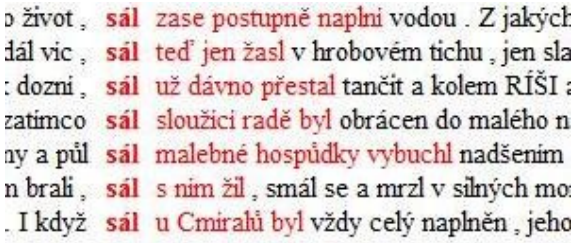
Udělejme nyní to samé s kontextem napravo (mějme při tom na paměti, že konjunkce či interpunkční znaménko v tomto případě nesmí nikoliv následovat, ale *předcházet* vyhledávaným slovesným tvarům):



<1, 2>

N filter: [tag="[ACDINPRTVX].*"] [tag="V.*"]

Počet výskytů: 147



ɔ život , sál zase postupně naplní vodou . Z jakých
dál vic , sál teď jen žasl v hrobovém tichu , jen sla
: dozni , sál už dávno přestal tančit a kolem RÍŠI :
zatímco sál sloužící radě byl obrácen do malého n
ny a půl sál malebné hospůdky vybuchl nadšením
n brali , sál s nim žil , smál se a mrzl v silných mo
. I když sál u Cmiralů byl vždy celý naplněn , jeho

<1, 3>

N filter: [tag="[ACDINPRTVX].*"] [tag="[ACDINPRTVX].*"] [tag="V.*"]

Počet výskytů: 140

III.

Více než polovina řádků stále obsahuje případy správně provedené lemmatizace. Můžeme je však postupně dále eliminovat jistými "mikropravidly". Například, následuje-li bezprostředně po tvaru "sál" slovo s velkým počátečním písmenem nebo samotné velké písmeno, jedná se ve všech korpusem doložených případech o substantivum:

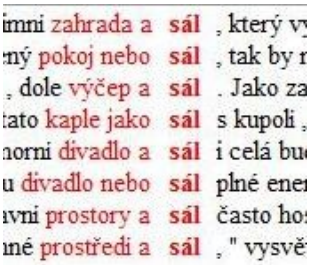
sobota , sál Kina Blansk
. Brno , sál B. Bakaly -
. večer - sál Dělnického
entrum , sál B , 14.00 - 1
vilon A , sál Rotunda , 19
entrum , sál B , 9.00 - 12
lon A3 , sál Rotunda , 10
lon A3 , sál Morava , 13.
vilon E , sál Press Cente
entrum , sál C , 9.00 - 13
entrum , sál C , 9.30 - 11
lon A3 , sál Morava , 10.
entrum , sál B , 9.00 - 14
ovského sál Národního d
uzeum , sál Jiřího Popela
po ránu sál ABC klubu r
din hosti sál Lidových sa
centru , sál A , je zaměř
u 2005 , sál Lípa pak po
vilon E , sál Business Ce
nažlice (sál MKS) kyva
oba dny sál Městského d

<1, 1>

```
N filter: [word="[ABCDEFHGIJKLMNOPQRSTUVWXYZ].*"]

Počet výskytů: 118
```

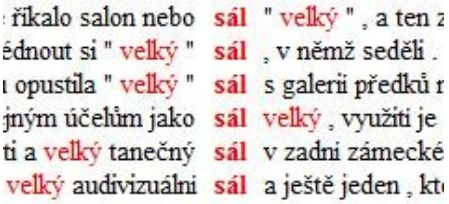
Pozn.: Případy, kdy půjde o sloveso, nelze pochopitelně ani v těchto případech zcela vyloučit, ale očividně jsou statisticky méně zastoupené.

Dále můžeme s jistotou odstranit všechny doklady, kde se vyskytují slovní spojení typu *zahrada a sál,* tj. takové, kde interval <-2, -1> obsahuje za sebou substantivum v nominativu a konjunkci.




```
<-2, -1>

N filter: [tag="NN..1.*"] [pos="J"]

Počet výskytů: 110
```

Můžeme také vyhledat případy, kdy se v kontextu širším než je pozice bezprostředně vlevo nachází adjektivum "velký":




```
<-2, 2>
```

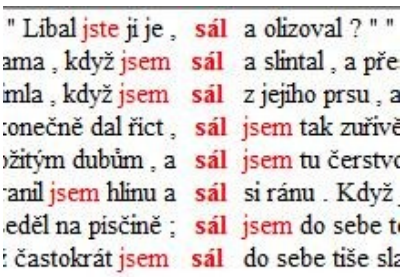
```
N filter: [word="velký"]

Počet výskytů: 104
```

Pozn.: Opět se nejedná o univerzální pravidlo. Uvažme například hypotetické *sál velký hrozen.*

IV.

Takto bychom mohli pokračovat dál a odstraňovat dalšími více či méně niternými (a často ne univerzálními) pravidly další případy správné lemmatizace. Seznam řádků nyní však již obsahuje více jak z poloviny případy lemmatizace chybné. Můžeme proto naopak vydělit tyto výskyty použitím pozitivních filtrů.

Vyčleňme například doklady, kdy je tvar "sál" slovesem v minulém času nebo kondicionálu 1. či 2. osoby, tj. jeho kontext obsahuje tvary *jsem, jsi, jste, bych, bys, byste* nebo *by.*




```
<-4, 2>

P filter: [word="jsem|jsi|jste|bych|bys|byste|by"]

Počet výskytů: 8
```

Zamyšlení:

Povšimněme si, že jsme v posledním případě zvolili kontext intuitivně na interval od -4 do 2. S tím se pojí otázka, **jak daleko může stát tvar pomocného slovesa "být" od základového tvaru?** Uvážíme-li kontext vlevo, zjistíme, že hledat zde jakákoliv omezení je značně problematické. Mezi základovým slovesem a pomocným tvarem

slovesným může totiž teroeticky stát větší množství více či méně rozvitých větných členů a jejich vzdálenost proto může být relativně dlouhá. V pravém kontextu je situace podstatně jednodušší. Zde se již může dostat mezi základové sloveso a pomocný tvar maximálně jeden výraz, jehož výběr je navíc omezený úzkou doménou slov. Můžeme si představit tvary jako např. *sál už jsem, sál prý jsem, sál kdysi jsem,* ale stěží je rozvineme tak, aby tvar pomocného slovesa "být" stál v pravém kontextu dále než na 2. pozici za základovým tvarem.

Stejná pravidla platí také pro vzdálenost základového slovesa a kondicionálových tvarů *bych, bys, byste* apod.