

CJBB84 - Morfologie a korpus

Úkol č. 1: Praktické problémy lemmatizace

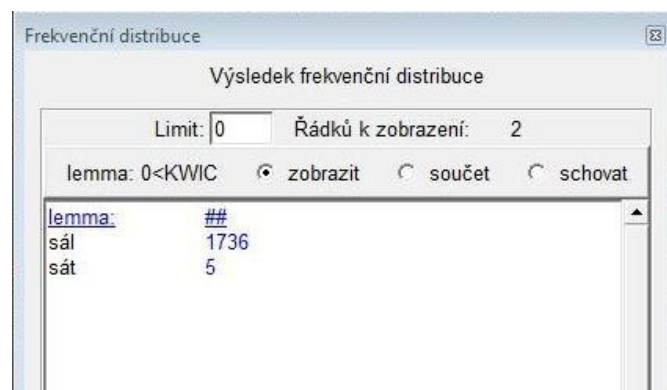
V korpusu SYN2010 hledáme případy chybné lemmatizace slovního tvaru "sál", tj. takové, ve kterých byl nesprávně označován coby substantivum a kdy se ve skutečnosti jedná o tvar slovesa "sát" v minulém čase (příp. kondicionálu).

I.

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

Query: "sál"

Počet výskytů: 1741



The screenshot shows a window titled "Frekvenční distribuce" with a subtitle "Výsledek frekvenční distribuce". It contains a search bar with "lemma: 0<KWIC" and three radio buttons: "zobrazit" (selected), "součet", and "schovat". Below the search bar is a table with two columns: "lemma:" and "##". The table lists "sál" with a frequency of 1736 and "sát" with a frequency of 5.

lemma:	##
sál	1736
sát	5

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 řadeů, z nichž však, jak vidíme po bližším prozkoumání, mnohé nejsou správně lemmatizovány:

... cířku si pomeřemou končetin **sál/sál** . Třebač nevěstičce zakřovili ma
i náhle vzpommeľa na operační **sál/sál** , z něhož Hoyt uprchl . Pouta ! b
pronikavý pach moči . Pitevní **sál/sál** je dokonale tichý a celý se čistou
erná nádra . " " Libal jste ji je , **sál/sál** a olizoval ? " " Přimo posedle . "
cientovi k ústům . Muž žiznivě **sál/sál** . pak polkl - - a vzápětí zaúpěl bc
chvíli ho odvezou na operační **sál/sál** . Ale zitra byste snad . . . " Jake
o jeho rodné město . Elegantní **sál/sál** plný proslulých osobností , nadše
stolu své fantazie konferenční **sál/sál** , hotel i restauraci , všechny ty v
itkou , kterou vezli na operační **sál/sál** a z níž mu v paměti utkvěl jen sti
unkou holčičku , co přivezli na **sál/sál** . . . " Sál byl pokoj s mnoha lůžky
Měli tam velký , velice světlý **sál/sál** . U širokých oken vedoucích na
na svém stehně . Z ní chlívnik **sál/sál** živiny . V noci pak broušil od jed
to vzduchu . Cestou ven mįjela **sál/sál** s velkou mapou Islandu . Sledov
trůnu se nenamáhal . Samotný **sál/sál** vypadal přesně tak , jak si ho Jaš
schodišti . Přestože je ponděli , **sál/sál** překypuje známými tvářemi , kte
teré lemovaly vchod . Taneční **sál/sál** jiskřil od stropu visícími lustry a c
iň stejně elegantní jako taneční **sál/sál** , byla nasáklá vážnou , tichou atr
stopky skořápku a labužnický **sál/sál** sladké voňavé mléko . Vždyť je
dalším soustem , které chytře **sál/sál** a dlouho převaloval v ústech jak
Í : podle matky v půli cesty na **sál/sál** , na chodbě porodnice ADRESA
se objevili Američané - se celý **sál/sál** vzdouval a uvnitř se vznášela tak
š lupala v reproduktoru , a celý **sál/sál** a všichni přítomní jako by se na c
oví ve vodním salonku a život **sál/sál** zase postupně nanhí vodou . Z i

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:

me šli s Frantou před sál a čekali až přijde
dostí . Odvezli mě na sál bez všech obligát
livat . Vedla mě přes sál . " Je v ranním pc
se obrátil , pohlédl na sál a potřásl hlavou .
; a dvě dámy , a přes sál nám laškovně zar
i naděje . " Jedete na sál , " konstatoval sa
i naděje . " Jedete na sál , " konstatoval sa
im zamířil rázně přes sál . Van Vleck se k
že má platit nájem za sál na příští měsíc a
/ezou - li ho např. na sál k operaci kyčelní
začal !) a rychle na sál . Milé sestřičky a
zli moji kopretinku na sál . Vypil jsem tři kc
e skokem ocitli mimo sál . Tančíme někde

<-1, -1>.

N filter: [pos="R"]

Počet výskytů: 1660

- tvaru "sál" bezprostředně předchází adjektivum:

o . " Viděl jsem přes celý sál , jak se chudáček červe
" Brzy to jméno volá celý sál , někteří jistě ani nevědí
rychle přeběhly přes celý sál , aby se s otcem přivita
ný salon a velký sloupový sál , kde byla recepce . Ná
k ho Zani přes zakouřený sál pozoruje a hosté u ostat
můžete připravit operační sál a chirurgický tým . Za j
ohlížel si obrovský nabitý sál . Od stolu k němu vzhlé
ze lidí . Samotný Hodovní sál pojme přes pět tisíc , hl
tisíc , hlavní ' konferenční sál ' deset tisíc , a to ještě :
ide převezen na operační sál , jakmile ho pro něj příp
snažili evakuovat . Velký sál RadioCity byl zaplněný
sále , i když je to taneční sál tvého hotelu , je příliš ri
t v patře , kde je zasedací sál . " " Ne , raději si s ní a
lidal vězně ? " " Operační sál představuje jakousi ster
i si prohlédnout koncertní sál . " Hlídač nevěřicně zal
le vzpomněla na operační sál , z něhož Hoyt uprchl .
ikavý pach moči . Pítevní sál je dokonale tichý a celý
li ho odvezou na operační sál . Ale zítra byste snad .
o rodné město . Elegantní sál plný proslulých osobnos
své fantazie konferenční sál , hotel i restauraci , vše

<-1, -1>

N filter: [pos="A"]

Počet výskytů: 471

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

" Vysvětlím tento sál pomoci náčrtku ,
 : ' Strýčku , tento sál je skvělý . Jak je
 dite dobře . Tento sál nemá s rytířským
 se nenazývá tento sál Nebeskou zahrad
 ů a darů , že tento sál je v podstatě tím ;
 hly postavit tenhle sál , jsou už minulosti
 ských domů . Ten sál byl původně " So
 bude divit . Žádný sál a zmechanizovan
 jistá , že když ten sál pronajímali , naps
 nádvoří , a každý sál byl na určitou dot
 ho stavu . Nějaký sál na Pražském Hra
 Záleží , jestli je to sál pro sedící publiku
 oru s tím , že je to sál víceúčelový . Tak
 n projektu je tento sál označen jako spe
 r - stejně . Tentýž sál , tytéž kulisy , židl

<-1, 1>

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

Počet výskytů: 441

vůni kadidla se sál ruské ambasády r
 ? " Až když se sál znovu rozšuměl , l
 udu , a opustila sál se svěšenou hlavc
 , poslouchá . A sál se ztrácí v soustře
 , " štěkl šerif a sál se spontánně roztl
 V okamžiku se sál zaplnil tančícími p
 ipnou , když se sál vyprázdni , jak se
 řeli restauraci a sál se začal pomalu v
 rou a napařují ; sál se bavi a tleská .
 ednu hodinu se sál naplnil k prasknutí
 její šálek čaje a sál se výborně bavil .
 nu dceru máte a sál se zaplnil výskajic

<-1, 1>

N filter: [word="se"]

Počet výskytů: 415

- tvaru sál předchází nebo po něm bezprostředně následuje číslovka:

Hotel zamluvil **sál** dvakrát , takž
 . Byl - li **první sál** celý bílá , zdo
 iě , " ten **druhý sál** , veřejnosti př:
 řistavěn **druhý sál** a chata dostal
 uškami . **První sál** je obrovská m
 1 , **Konfereční sál** 19 . 10 . 2006
 o a čas konání : **sál** 102 , 103 , ad
 budova BVV , **sál** 102,103 , 15.0
 ůňovaný **druhý sál** je nepoměrně
 ždy připravoval **sál prvního** posch
 a - živé . **První sál** ukazuje vliďně
 íe . " Na **jeden sál** je to až moc ,
 . 7 . , Prokyšův **sál** 19.30 Johan S
 ide ještě **jeden sál** , ve kterém z:
 tržby na **jeden sál** třikrát až čtyř
 im sále . **První sál** je dokumenta
 rativní budova , **sál** 102 Energetic
 . ledna . **První sál** je Jessie plný

<-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: *žiznivý 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, přičestí 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).

bedlivě přehlížel **sál** , a i když policie
 estou ven mijela **sál** s velkou mapou
 hou . Geologové **sál** vypustili jistě p
 nich , takže byl **sál** už naplněn nat
 . Světla zhasla a **sál** ztichl . Sisa stí
 ruk flétny naplnil **sál** , Úžasný Albe
 ylo . Než opustili **sál** , Jupiter se na
 nimi se rozkládal **sál** s vysokým kle
 ě právě opouštěli **sál** ve skrytu seve
 omylem , opustil **sál** . " Je to součá
 m zvolna naplnili **sál** všichni ostatní
 k JIPce . Přelétl **sál** pohledem . Vě
 řlasy . Pak našla **sál** . Pohlédla ode
 rychle obhlédla **sál** , uviděla , že z
 až jsme opustili **sál** a spatřili před
 nkova , opouštěli **sál** dětského oddě

<-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.

. Blízko je **sál** Lucerna . T
 také trůní **sál** . Z divánu se
 . Když je **sál** už zpola plný
 i opouštějí **sál** . Osvícení v
 ntrolu , zda **sál** je naplněn ,
 poda huči **sál** tleská paňác
 n ovládne **sál** . Také z ruz
 lgie opusti **sál** . Jednoho m
 i , ve které **sál** je pítevní a s
 ě Konečně **sál** buráci potles
 n zhasnou **sál** a skončí zde
 řhan získá **sál** další rozměr
 vázala , že **sál** bude pět let
 těší , když **sál** nezeje prázd
 Gorlice je **sál** uvnitř kasem
 rrvní tóny , **sál** začne zpívat
 připravuji **sál** i mladého m
 h freskách **sál** připomíná st
 tom , jestli **sál** nedostane ta
 'aké to , že **sál** přejde po pa

<-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>.

l a přelétl pohledem sál . Zastavil s
stna pronajmout svůj sál komukoli ,
otože to byl vpravdě sál v rozloze p
lších dveří . Byl tu sál , trochu zaj
děl a představoval si sál plný lidí , c
agenturu a najala si sál v kulturáku
zkusilo - pronajal si sál . Snad ho k
teta přechází ztěžka sál . Teď je te
slet . Obcházel jsem sál , abych po
o jiné najít například sál věnovaný r
z Broumova , byl by sál poloprázdn
ř hotelu je připraven sál napěchova
: něhož byl postaven sál pro jeho kc
é prohlédl Zanderův sál . Nejvíce b
muset půjčovat svůj sál a opět se o
účasti podniku je těž sál a právě v r
í technologie a je tu sál pro tiskové
pod , kde mají vzadu sál , schází se
c . Městu chybi také sál pro společ
stříhů obsadila včera sál prostějovsk
ka lidí zaplnila včera sál v Auditoriu

<-2, -1>

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

Počet výskytů: 169

Zajdeme 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.)

obhlédl unaveným zrakem **sál** a chtěl se z:
 ovalí . Postavili jsme ve vsi **sál** . " Pani Ma
 měli ? Já znám vlastně jen **sál** na Stadionu
 oněnou látku jsem do sebe **sál** teplý podzim
 otěhlosti kladu vedle sebe **sál** plný intelekt
 př. ve spojení jít do sálu/na **sál** , viz dále) ,
 o invalidu vezou do sálu/na **sál** , vnímáme t
 ihu se nalézá velký sítrový **sál** , jehož starý
 iek . " Příště bude při party **sál** Vřídla uzam
 " bavil slovenský showman **sál** . Dlouho bu
 t .) a ukazoval nám přítom **sál** kulturního d
 y v Grandu , není ve městě **sál** , kde by mo
 ění muzice patří dnes večer **sál** Klubu Na R

<-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):

. Nařídila správci , aby **sál** bezpečně uzamkl a nikomu neotvíral .
 i pronikavá disonance a **sál** ho nevypískal jenom proto , že největš
 paláci , patřícím městu , **sál** to byl spíš konferenční , pro velká zas
 ěna , a když vcházíme , **sál** už je zaplněn . Ráno , když jsme do ně
 ěmi upřely na admirála a **sál** opět ovládlo ticho . Admirál Clay si da
 vyhovující . A publikum **sál** opravdu naplnilo , ocenilo výborné výl
 . Bylo krátce po volbě a **sál** Fouskovou odměnil bouřlivým aplause
 ě hlavně jsme chtěli , aby **sál** zastupitelstva byl jen pro zastupitele ,
 ůrga a Jane Birkinové , **sál** večer oceni nadšeným potleskem a je

<1, 2>

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

Počet výskytů: 147

o život , **sál** zase postupně naplní vodou . Z jakýchk
dál víc , **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 ním žil , smál se a mrzl v silných mo
. I když **sál** u Cmirahů 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 korpusech doložených případech o substantivum:

sobota , **sál** Kina Blansko
. Brno , **sál** B. Bakaly -
. večer - **sál** Dělnického d
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 Center
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
dín hosti **sál** Lidových sac
centru , **sál** A , je zaměř
u 2005 , **sál** Lipa pak po j
vilon E , **sál** Business Cer
nažlice (**sál** MKS) kyva
oba dny **sál** Městského d

<1, 1>

N filter: [word="[ABCDEFGHGIJKLMNOPQRSTUVWXYZ].*"]

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 konjunktci.

imní zahrada a sál , který vy
ný pokoj nebo sál , tak by r
, dole výčep a sál . Jako za
tato kaple jako sál s kupolí ,
normní divadlo a sál i celá bu
u divadlo nebo sál plně ene
vní prostory a sál často ho
né prostředí a sál , " vysvě

<-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":

řikalo salon nebo sál " velký " , a ten z
ědnout si " velký " sál , v němž seděli .
i opustila " velký " sál s galerii předků r
jným účelům jako sál velký , využití je
ti a velký tanečný sál v zadní zámecké
velký audivizuální sál a ještě jeden , kt

<-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*.

" Líbal **jste** ji je , **sál** a olizoval ? " "
ama , když **jsem** **sál** a slíntal , a pře
mla , když **jsem** **sál** z jejího prsu , a
:onečně dal říct , **sál** **jsem** tak zuřivě
žítým dubům , a **sál** **jsem** tu čerstvě
:anil **jsem** hlinu a **sál** si ránu . Když
:eděl na písčíně ; **sál** **jsem** do sebe t
: častokrát **jsem** **sál** do sebe tiše slá

<-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 teoreticky 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 kondicionálních tvarů *bych, bys, byste* apod.