

Derivancze — Derivational Analyser of Czech, Derinet

PLIN033

osolsobe@phil.muni.cz

Nástroje pro zpracování slovo tvorby

- *Deriv*
- *Derivancze*
- *Derinet*

Deriv (deb.fi.muni.cz/deriv) – bohužel neudržováno

- webové rozhraní
- schopnost pracovat s morfologickým slovníkem analyzátoru *(m)ajka*
- propojení s tištěnými slovníky
- propojení s korpusy
- Veronika Kalivodová: **Tvorba uživatelského manuálu pro DEBDict a Deriv (Bc. DP, FF MU, 2017: https://is.muni.cz/th/mggon/?zoomy_is=1)**
- OSOLSOBĚ, Klára, Karel PALA, Pavel ŠMERK a Dana HLAVÁČKOVÁ. Relations between Formal and Derivational Morphology in Czech. In **Czech in Formal Grammar**. Mnichov: Lincom, 2009. s. 79-87, 9 s. ISBN 978-3-89586-282-3.

Derivancze:

<https://nlp.fi.muni.cz/projects/derivancze/>

- Pala K., Šmerk P. (2015) Derivancze — *Derivational Analyser of Czech*. In: Král P., Matoušek V. (eds) Text, Speech, and Dialogue. TSD 2015. Lecture Notes in Computer Science, vol 9302. Springer, Cham.
https://doi.org/10.1007/978-3-319-24033-6_58

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

OK

For more details see [TSD 2015 paper](#).

Data are not freely downloadable. If you need them, especially for research purposes, please contact ma@nlp.fi.muni.cz.

Hledání odvozených slov **generovaných** **automaticky** na základě brněnského slovníku (Ošelechů 1996)

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

For more details see [TSD 2015 paper](#).

Data are not freely downloadable. If you need them, especially for research purposes, please contact ma@nlp.fi.muni.cz.

kralovat

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

kralovat (15 15):

⇒ kralovaný k2rpaš 4 2

⇒ kralován k2pas 3 0

⇒ kralování k1verb 12 11

⇒ kralující k2prod 9 9

Description

⇒ word can be considered derived from the analyzed one

⇐ word can be considered base for the analyzed one

⇔ words can be considered equivalent

Numbers are 1 + whole part of \log_2 of frequency of the given lemma in CzTenTen12 and SYN corpora.

synv8

Korpus: [syn v8](#) | Dotaz: [kralovan.*](#) (62 výskyty) ▶ Promíchat: ✓ ▶ Pozitivní filtr: [kralovaní](#) (4 výskyty)

Výskytů: 4 | i.p.m.: 0 (vztaženo k celému korpusu) | ARF: 3,38 | Výsledek je promíchán

1 / 1

Výběr řádků: základní ▾

- | | | | | |
|--------------------------|------------------------------------|---|---|--|
| <input type="checkbox"/> | Deniky Moravia | . Mezi hlavními kandidáty , kteří by mohli jejich šestileté | kralovaní / kralovaný / AAMP1-----1A----- | přerušit , jsou finský mistr SSV Helsinky , jeho ligový |
| <input type="checkbox"/> | Maxim | stal se králem Lýdie . To samé (ale bez | kralovaní / kralovaný / AAMP5-----1A----- |) dělají kandaulisté i v roce 2013 , a mnohem |
| <input type="checkbox"/> | Prostějovský týden | tábory , školy v přírodě . Helena Hermanová , jejíž | kralovaní / kralovaný / AAMP1-----1A----- | pozvedlo školní jídelnu na jednu z nejoblíbenějších místností v areálu |
| <input type="checkbox"/> | Hospodářské noviny | všeobecně považována za slabší a svědčí o tom i čtyřleté | kralovaní / kralovaný / AAMP1-----1A----- | týmů ze západu ve Stanleyově poháru . Do role favorita |

Korpus: [syn v8](#) | Dotaz: [kralován](#) ▶ Promíchat: ✓

Výskytů: 0 | i.p.m.: 0 (vztaženo k celému korpusu) | ARF: - | Výsledek je promíchán

Výběr řádků: základní ▾

czTenTen17

lemma kralovaný 19 less than 0.01		Left context	KWIC	Right context
<input type="checkbox"/>	Details			
1	<input type="checkbox"/> lidovky.cz	stože měnil Fin pneumatiky o pět kol později než lídr závodu, Hamilton své	kralovaný kralovaný/k2eAgMnPc1d1	dotáhl až do cíle a zvítězil o 1,032 sekundy před Räikkönenem. </s><s>
2	<input type="checkbox"/> signaly.cz	hezké, ale nese to s sebou i stinné stránky. </s><s> Například zanedbává	kralovaný kralovaný/k2eAgMnPc1d1	a zachráněné princezny se vždy urazí, protože žádnou z nich si nevezme
3	<input type="checkbox"/> radiovaticana.c...	li, a tam, kam mohou vstoupit, přinášejí mír a radost. </s><s> To je způsob	kralovaný kralovaný/k2eAgMnPc1d1	Boha; to je jeho spasitelský plán; ? </s><s> tajemství? </s><s> v biblické
4	<input type="checkbox"/> wikisource.org	9. pros. 1309); opanoval' pomalu celé království a brzy nescházela mu ku	kralovaný kralovaný/k2eAgMnPc1d1	leda koruna. </s><s> Pro nabytí také té obrátil se roku 1319 ku papeži Ja
5	<input type="checkbox"/> eucharistie.cz	Pána by byl nevložil jho poddanství násilím na šiji svých národů? </s><s>	Kralovaný kralovaný/k2eAgMnPc5d1	Kristovo je však sladké - na život a na smrt jsou mu proto oddáni jeho po
6	<input type="checkbox"/> evangnet.cz	A pak se nám možná i ty chvíle úzkosti stanou znamením blízkosti božího	kralovaný kralovaný/k2eAgMnPc1d1	, znamením naděje, která si právě uprostřed temnot razí cestu. </s><s> A
7	<input type="checkbox"/> estranky.cz	toupili do Itálie roku 488 a úplně ho dobyli roku 493. </s><s> Theodorikovo	kralovaný kralovaný/k2eAgMnPc1d1	netrvalo dlouho, protože roku 526 zemřel. </s><s> Využitím boje o nástup

czTenTen17

lemma **kralován** 12 less than 0.01

Left context KWIC Right context

	Left context	KWIC	Right context
1	wikipedia.org pouze uznávají brit. královnu za svou hlavu – nejsou královstvím, ale je jim	kralováno kralován/k2eAgNnSc1d1). </s><s> Vhodnější je termín panování, doména popř. sféra – vizte také he
2	wikipedia.org nozdřejmě Spojené království. </s><s> Jinak souhlasím stím, že ostaním je	kralováno kralován/k2eAgNnSc4d1	... --Kolomaznik 23. 10. 2009, 19:25 (UTC) </s><s> Britská královna Alžběta
3	estranky.cz lo Westminster Abbey. </s><s> Byl umístěn pod trůnem, takže když byl král	kralován kralován/k2eAgMnSc1d1	za krále Anglie, byl také kralován jako Skotský král. </s><s> Takže potom s
4	estranky.cz /l umístěn pod trůnem, takže když byl král kralován za krále Anglie, byl také	kralován kralován/k2eAgMnSc1d1	jako Skotský král. </s><s> Takže potom se tihle studenti v 50. letech rozhod
5	nasemostecko.cz ála již v roce 2004. Tedy ještě předtím, než závodu Sydney - Hobart začalo	kralováni kralován/k2eAgMnPc1d1	větší jachty Wild Oats. </s><s> Loď YuuZoo, na které Martin Trčka pojede, j
6	blesk.cz a další tuneláři </s><s> vildasin: Raději ani nevědět, kolik si dědek za éru "	kralováni kralován/k2eAgMnPc1d1	" nahrabal a co všechno vlastní. </s><s> milosuvadokat: a ještě mu potrhli
7	skoda-forum.cz ovali také na čínské rallye , kterou si hned při prvním startu podmanili a byli	kralováni kralován/k2eAgMnPc1d1	na šampiony mistrovství Asia-Pacific Rally. </s><s> Na své si přišli také Ma

Pozadí morfologických slovníků

- Pražský (*MorfFlex*) i brněnský slovník byly vytvořeny na základě pravidel, které „rozgenerovaly“ potenciální (paradigmaticky tvořené) tvary.
- Od každého slovesa se tak tvoří:
 - a) úplný soubor tvarů pasivního přičestí/ tzv. krátkých tvarů (viz zde *kralován*), a to bez ohledu na to, zda jde o sloveso, které tyto tvary tvoří, či nikoliv;
 - b) úplný soubor tvarů dlouhých/adjektivních (viz zde *kralovaný*);
 - c) úplný soubor slovesných substantiv na *ní/tí* (viz zde *kralování*);
 - d) úplný soubor procesuálních adjektiv na *oucí/ící* omezený pouze videm (nedokonavým) základového slovesa (viz zde *kralující*).

Automatická morfologická analýza

- Je založena na slovnících, a tak pokud je ve slovníku ke tvaru nalezena dvojice lemma+tag, pak je použita, viz interpretace udělované překlepům.
- Dobře/špatně?
- Nikoliv, automatické nástroje lze např. využít k tomu, aby se ze slovníku odstranily přegenerované výsledky (např. na základě vyloučení/potlačení tvarů, které mají v korpusech nulovou/malou frekvenci). Tento postup je ovšem třeba aplikovat velmi opatrně (viz nepředpokládané, ale doložené tvary typu *kralováno* atd.)

Další funkce nástroje *Derivancze*: značky obsahující derivační vztahy (automaticky generované a ručně editované)

Numbers are 1 + whole part of \log_2 of frequency of the given lemma in CzTenTen12 and SYN corpora.

label	relation/example	# pairs in data both >1× in CzTenTen	both >1× in SYN	
k1dem	deminutives (<i>dům</i> — <i>domek</i>)	6342	5251	3343
k1f	nouns: masculine → feminine form (<i>doktor</i> — <i>doktorka</i>)	3196	2369	1881
k1jmf	surnames: masculine → feminine form (<i>Novotný</i> — <i>Novotná</i>)	2230	2049	2114
k1jmr	surnames: masculine → "family" form (<i>Novotný</i> — <i>Novotní</i>)	2212	1786	19
k1obyv	nouns: place/area → demonym/gentilic (<i>Kanada</i> — <i>Kanaďan</i>)	262	241	209
k2pos	noun → possessive adjective (<i>otec</i> — <i>otcův</i>)	30953	11879	6861
k2rel	noun → relational adjective (<i>mráz</i> — <i>mrazový</i>)	23531	18997	14614
k1prop	adjective → property/quality (<i>hluchý</i> — <i>hluchota</i>)	9901	7518	5988
k6a	adjective → adverb (<i>dobrý</i> — <i>dobře</i>)	45108	18947	12685
k1verb	verb → process/state/action (<i>bít</i> — <i>bití</i>)	35723	21335	15849
k2pas	verb → passive participle (<i>bít</i> — <i>bit</i>)	34847	11273	192
k2proc	verb → present active adjectival participle (<i>bít</i> — <i>bijící</i>)	15765	7040	5539
k2rakt	verb → past active adjectival participle (<i>zabít</i> — <i>zabivší</i>)	18106	1150	600
k2rpas	verb → past passive adjectival participle (<i>bít</i> — <i>bitý</i>)	35015	17842	12341
k2uce1	verb → adjective [~ purpose: beat → used for beating] (<i>bít</i> — <i>bicí</i>)	1672	1582	1390
k5freq	verb → frequentative (<i>bádat</i> — <i>bádávat</i>)	2228	1259	787
k1ag	verb → agent (<i>bádat</i> — <i>badatel</i>)	703	588	447
var	orthographical/morphological/derivational variant (no D-relation, <i>komunizmus</i> — <i>komunismus</i>)	568	409	100
Σ		268363	131516	84960

strom

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

`strom` (19 19):

⇒ `stromek` k1dem 16 16

⇒ `stroměček` k1dem 16 15

⇒ `stromový` k2rel 14 12

⇒ `strůmek` k1dem 7 5

dům

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

dům (22 21) :

⇒ domek k1dem 18 18

⇒ domový k2rel 9 0

Ajka:

<https://nlp.fi.muni.cz/projekty/wwwajka/WwwAjkaSkripty/morph.cgi?jazyk=0>

Morfologický analyzátor ajka – interaktivní režim

Zadejte krátký text nebo vložte soubor [zde](#).

domeček

Proved'

[\[Nápověda\]](#)

Jakou akci provést

Akcentovat



Segmentovat



Analyzovat



Vstupní kódování

ISO-8859-2

Windows-1250

Výsledek morfologické analýzy – interaktivní režim

(* - Vypiš všechny odvozené tvary

Analyzovaný tvar: domeček

Základní tvar	Segmentace	Číslo vzoru	Kategorie
domeček (*)	=domeč=ek==	1587-vršeček	k1gInSc1
			k1gInSc4

? *strom* → *stromek* → *stromeček*

× *dům* → *domek*

a separé *domek* → *domeček*

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

domek (18 18) :

⇒ domeček k1dem 16 14

← dům k1dem 22 21

Problém je v automatickém zpracování dat, zejména v případě hláskových alternací a dalších nepravidelností, které jsou spoluformanty derivace

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

`nosit` (19 17):

⇒ `nositel` k1ag 17 16
⇒ `nosící` k2proc 11 10
⇒ `nosívát` k5freq 12 11
⇒ `nošen` k2pas 10 0
⇒ `nošení` k1verb 16 14
⇒ `nošený` k2rpal 14 10

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

`přát` (20 19):

⇒ `přán` k2pas 1 0
⇒ `přání` k1verb 19 17
⇒ `přávat` k5freq 8 7
⇒ `přaný` k2rpal 5 1
⇒ `přející` k2proc 13 11

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

`velet` (16 15):

⇒ `velen` k2pas 11 0
⇒ `velení` k1verb 16 15
⇒ `velený` k2rpal 9 5
⇒ `velící` k2proc 13 11
⇒ `velívát` k5freq 3 1

Podobně:

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

sázet (16 16):

- ⇒ sázecí k2ucel 11 9
 - ⇒ sázející k2proc 12 12
 - ⇒ sázen k2pas 7 0
 - ⇒ sázení k1verb 15 14
 - ⇒ sázený k2rpas 11 9
 - ⇒ sázivat k5freq 6 3
-

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

stát (23 22):

- ⇒ státek k1dem 8 5
- ⇒ stan k2pas 7 5
- ⇒ stání k1verb 16 16
- ⇒ stáný k2rpas 8 0
- ⇒ stávat k5freq 19 18
- ⇒ staný k2rpas 6 1
- ⇒ stojící k2proc 17 16

Různé nesrovnalosti

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

sát (16 12):

⇐ sát k2pas 16 12
⇒ sát k2pas 16 12
⇒ sátí k1verb 6 3
⇒ sátý k2rpas 8 5
⇒ sán k2pas 9 0
⇒ sán k2rakt 9 0
⇒ sání k1verb 14 11
⇒ sávat k5freq 7 1
⇒ sající k2proc 11 9
⇒ saný k2rpas 5 1

krýt (17 16):

⇒ krycí k2ucel 16 14
⇒ kryjící k2proc 12 10
⇒ kryt k2pas 17 15
⇒ krytí k1verb 16 14
⇒ krytý k2rpas 17 16

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

hrát (22 21):

⇒ hrán k2pas 0 0
⇒ hrávat k5freq 15 15
⇒ hrající k2proc 16 17
⇒ hraní k1verb 17 16
⇒ hraný k2rpas 16 16

Tvary sací generované rozhraním *Ajky*

Odvozené tvary ke slovu "sací"

Negace: Afirmace															
Stupeň: Nominativ															
Rod: Mužský životný				Rod: Mužský neživotný				Rod: Ženský				Rod: Střední			
Pád	Singulár	Pád	Plurál	Pád	Singulár	Pád	Plurál	Pád	Singulár	Pád	Plurál	Pád	Singulár	Pád	Plurál
1	sací	1	sací	1	sací	1	sací	1	sací	1	sací	1	sací	1	sací
2	sacího	2	sacích	2	sacího	2	sacích	2	sací	2	sacích	2	sacího	2	sacích
3	sacímu	3	sacím	3	sacímu	3	sacím	3	sací	3	sacím	3	sacímu	3	sacím
4	sacího	4	sací	4	sací	4	sací	4	sací	4	sací	4	sací	4	sací
5	sací	5	sací	5	sací	5	sací	5	sací	5	sací	5	sací	5	sací
6	sacím	6	sacích	6	sacím	6	sacích	6	sací	6	sacích	6	sacím	6	sacích
7	sacím	7	sacími	7	sacím	7	sacími	7	sací	7	sacími, sacíma	7	sacím	7	sacíma, sacími

lemma *sací*

lemma **sací** 51,939 (4.13 per million)



Details

Left context

KWIC

Right context

1	<input type="checkbox"/> wikipedia.org il. </s><s> Rozvod: OHV (řízený vačkovým kotoučem), dvouventilový (s jedním	sací sací/k2eAgInSc7d1	a jedním výfukovým ventilem na válec) </s><s> A.C.A.B. (z anglického "All Cop
2	<input type="checkbox"/> wikipedia.org řízení nosné plochy. </s><s> Některé série měly na levé straně motorového krytu	sací sací/k2eAgNnSc1d1	hrdlo vzduchu do karburátoru. </s><s> Letouny byly rovněž vybaveny pumovým
3	<input type="checkbox"/> wikipedia.org cky šlo v podstatě o stejný vůz, s malými modifikacemi brzdového posilovače a	sacího sací/k2eAgNnSc2d1	potrubí u některých verzí. </s><s> Největším trhem pro společnost Giulietta byl:
4	<input type="checkbox"/> wikipedia.org Základem stálé sbírky Vzdušného muzea jsou objekty, instalace, vzduchostroj,	sací sací/k2eAgInPc1d1	a foukací objekty, vzdušná typografie a videa. </s><s> Sběrka je doplněna o vyp
5	<input type="checkbox"/> wikipedia.org íchnutím je vhodné seříznout konec stonku do špičky, což se dělá kvůli zvýšení	sací sací/k2eAgFnSc2d1	schopnosti rostliny a snazší zapáchnutí. </s><s> Kromě aranžovací hmoty jako
6	<input type="checkbox"/> wikipedia.org ytonu. </s><s> Markův vidlicový osmiválec byl zvýšením komprese, zvětšením	sacíh sací/k2eAgMnPc2d1	ventilů, speciálním vačkovým hřídelem, novým sáním a zvětšením karburátorů
7	<input type="checkbox"/> wikipedia.org standardní V8 lišil zadním spoilerem, zakrytou maskou chladiče a kapotou bez	sacího sací/k2eAgInSc2d1	otvoru. </s><s> Měl také širší podběhy a boční prahy. </s><s> V průběhu výrob
8	<input type="checkbox"/> wikipedia.org zidel se běžně používá podtlakový posilovač brzd využívající sníženého tlaku v	sací sací/k2eAgNnSc6d1	potrubí motoru k usnadnění ovládnání brzdového pedálu řidičem. </s><s> U aut
9	<input type="checkbox"/> wikipedia.org s> Od 90. let jej nahradilo elektronicky řízené nízkotlaké vstřikování benzínu do	sacího sací/k2eAgNnSc2d1	potrubí v prostoru před sacím ventilem a to společnou tryskou (SPI = Single-poi
10	<input type="checkbox"/> wikipedia.org ronicky řízené nízkotlaké vstřikování benzínu do sacího potrubí v prostoru před	sací sací/k2eAgInSc7d1	ventilem a to společnou tryskou (SPI = Single-point injection) nebo tryskami zvl:

Pozor

- *bít* → *bil* → *bicí*
- *krýt* → *kryl* → *krycí*
- *sát* → *sál* → *sací*
- *hrát* → *hrál* → *hrací*
- Ověřte, zda platí, že pokud se tvar účelového adjektiva na *cí* tvoří od kmene minulého slovesa, které nemá alternaci ve tvarech od kmene minulého, a zároveň má alternaci ve tvarech účelového adjektiva, pak ji nástroj *Derivancze* nezná.

Která slovesa nemají alternaci kořenového vokálu?

- Kořenový vokál dlouhý infinitivu se nekrátí u sloves III. třídy *krýt*, pokud tímto vokálem je *á*.
- Seznam sloves: *hrát, hřát, okřát, přát, smát se, sát, vát, ...*

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

přát (20 19):

- ⇒ *přán* k2pas 1 0
- ⇒ *přání* k1verb 19 17
- ⇒ *přávat* k5freq 8 7
- ⇒ *přaný* k2rpa 5 1
- ⇒ *přející* k2proc 13 11

Zachycení paradigmatických derivací někde bez ohledu na významové posuny

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

hodný (23 23):

⇒ hodnost klprop 13 12
⇒ hodnota klprop 14 13
⇒ hodně k6a 19 18

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

hodnota (21 19):

⇒ hodnotový k2rel 15 13

Derivancze — *Derivational Analyser of Czech*

Word for analysis:

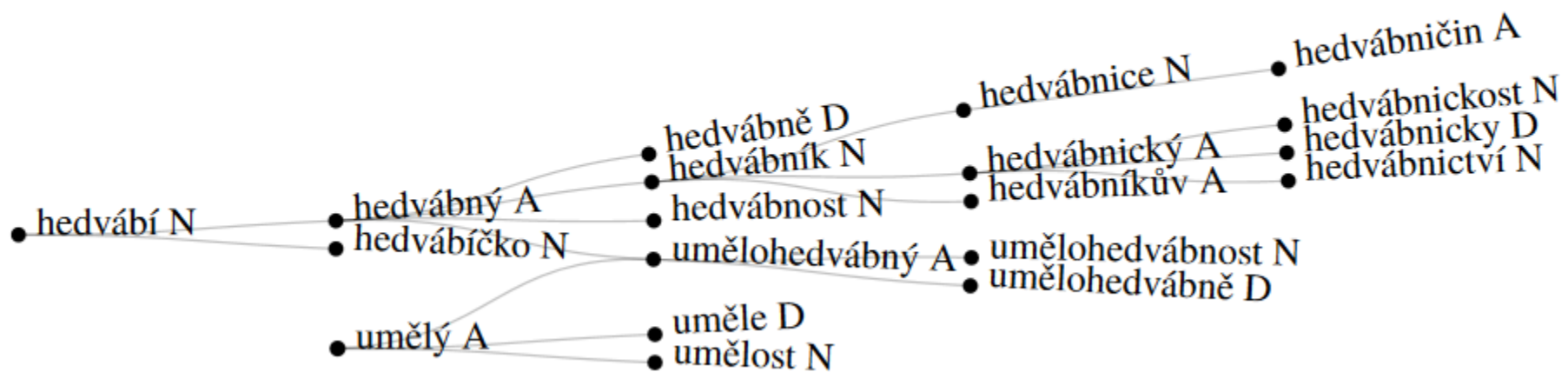
hodný (18 16):

⇒ hodnost klprop 15 14
⇒ hodně k6a 23 22

Derinet (<https://ufal.mff.cuni.cz/derinet>)

- **DeriNet 2.0 (<https://ufal.mff.cuni.cz/derimo2019/pdf-files/derimo2019-10.pdf>)**
- 1 milion lexemů (vzorek ze slovníku MorfFlex) propojených 808 tisíci derivačních vztahů a 600 odkazů z kompozit na základová slova;
- anotace morfologických kategorií (u všech lexémů – zajišťuje rozdílné derivační vztahy u homonym: *stát = V/N*, *tulení = N(v)/A(n)*, ...),
- identifikace kořenových morfů (u 250 lexemů),
- semantické labely (150 relací, 5 labelů),
- kompozita (600 lexemů)
- tzv. fiktivní lexémy (testování např. *-bízet*).

Lexikální síť - příklad

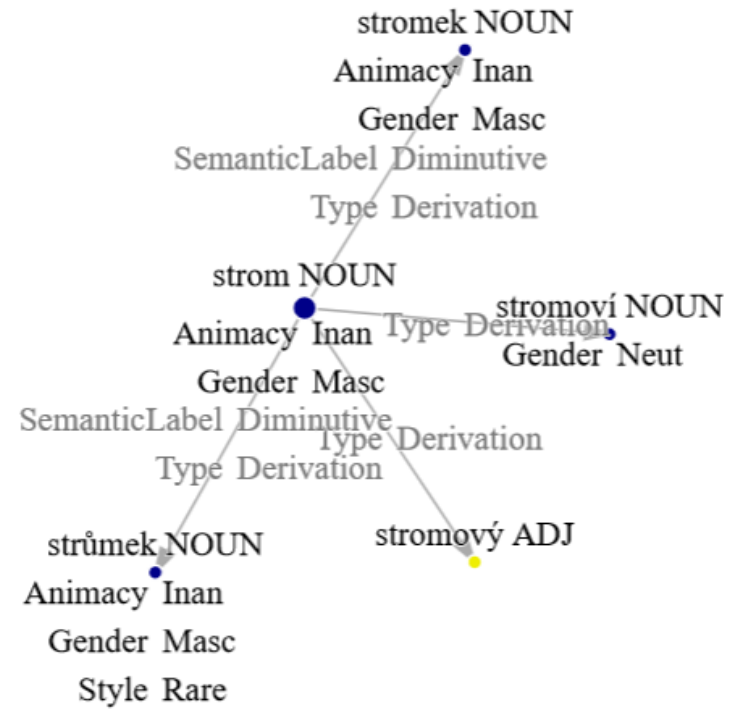


Sémantické labely

Label	Count
POSSESSIVE	88,718
FEMALE	29,023
ASPECT	15,439
ITERATIVE	11,886
DIMINUTIVE	5,939

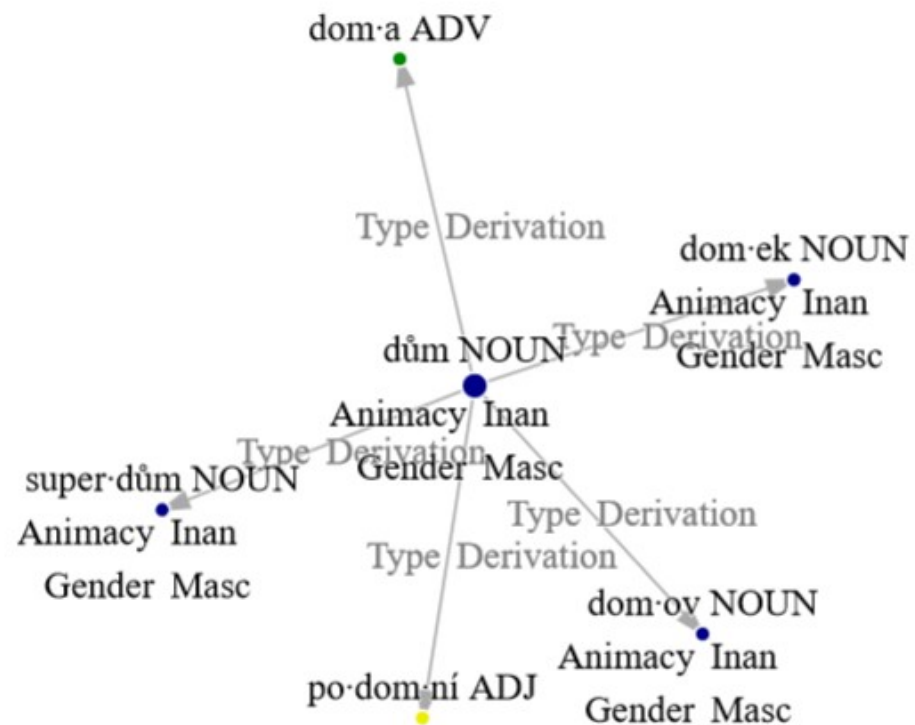
Table 3: Counts of the semantic labels in DeriNet 2.0 data.

strom



dům

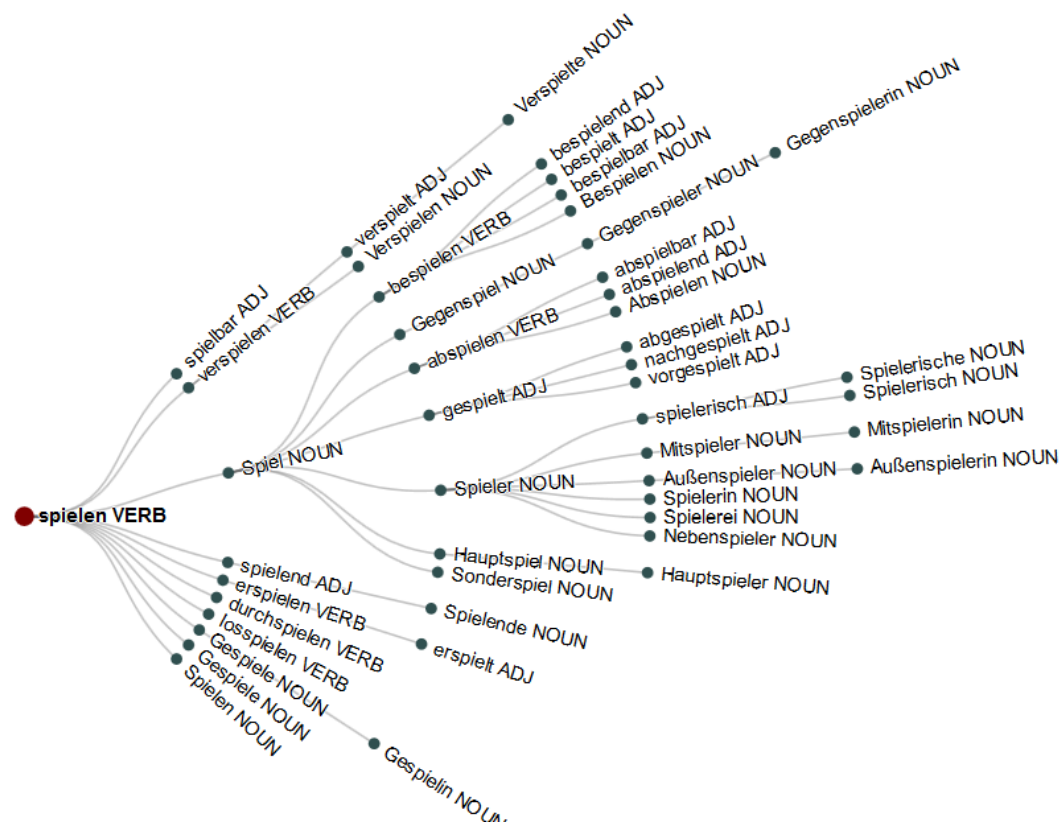
<https://quest.ms.mff.cuni.cz/derisearch2/v2/databases/Czech-DeriNet-2.0/dcql?ci=false&defA=lemma&limit=10&offset=0&q=d%C5%AFm&style=stretch>



Vyzkoušejte

- <https://quest.ms.mff.cuni.cz/derisearch2/v2/databases/Czech-DeriNet-2.0/dcql?ci=false&defA=lemma&limit=10&offset=0&q=d%C5%AFm&style=stretch>

Němčina



Četba: Dů – 9. 12. 2020

- Prostudujte [www stránky ÚFAL](#)
- články v IS_studijní materiály