Základy využití korpusu v praxi **cjbb75**

**pondělí: 8.00-10.00 G13**

**11. 5. 2020**

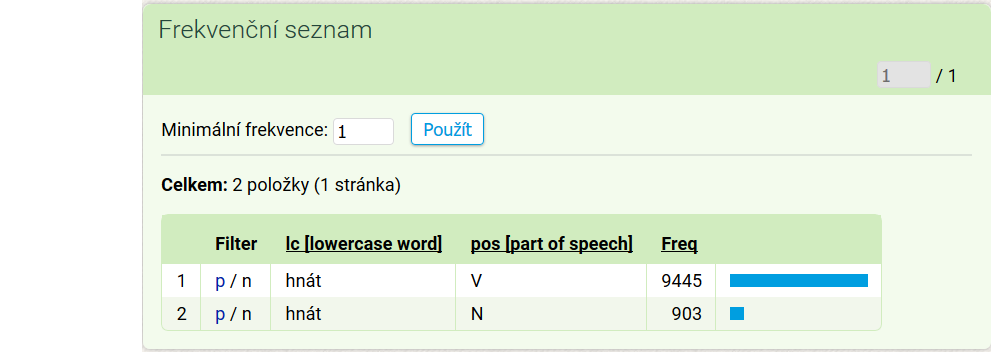
**Korpus jako zdroj dat pro výzkum syntaxe.**

Syntaktické kritérium určení slovnědruhové platnosti slov. (lingvistické termíny: slovní druh, větný člen, primární a sekundární větněčlenské funkce slovních druhů, slovnědruhový přechod/transpozice, substantivizace, prepozicionalizace, slovosled uvnitř jmenné skupiny, gramatická shoda…; termíny pro formální zadání: řetězec, token, lemma, morfologická značka/tag, pravý a levý kontext, pozice vpravo a vlevo od KWIC, filtr, …).

**Jakou slovnědruhovou platnost může mít slovo *hnát, popravčí, pila*?**

 **Korpus:**syn v8

  | **Dotaz**: hnát (10 348 výskytů)

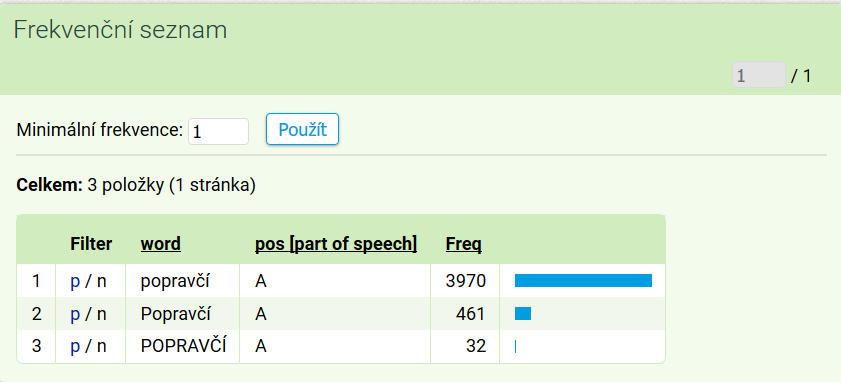






 **Korpus:**syn v8

  | **Dotaz**: popravčí (4 463 výskytů)

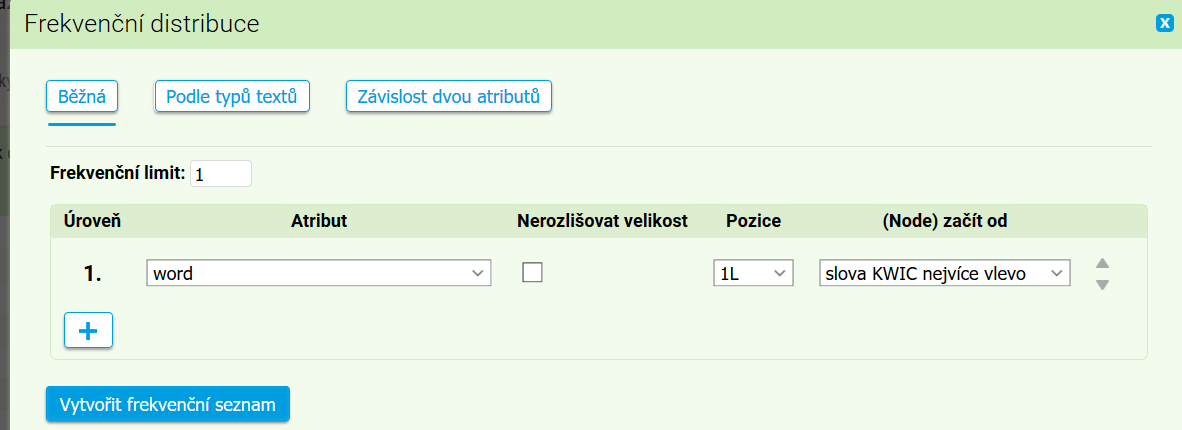


 **Korpus:**syn v8

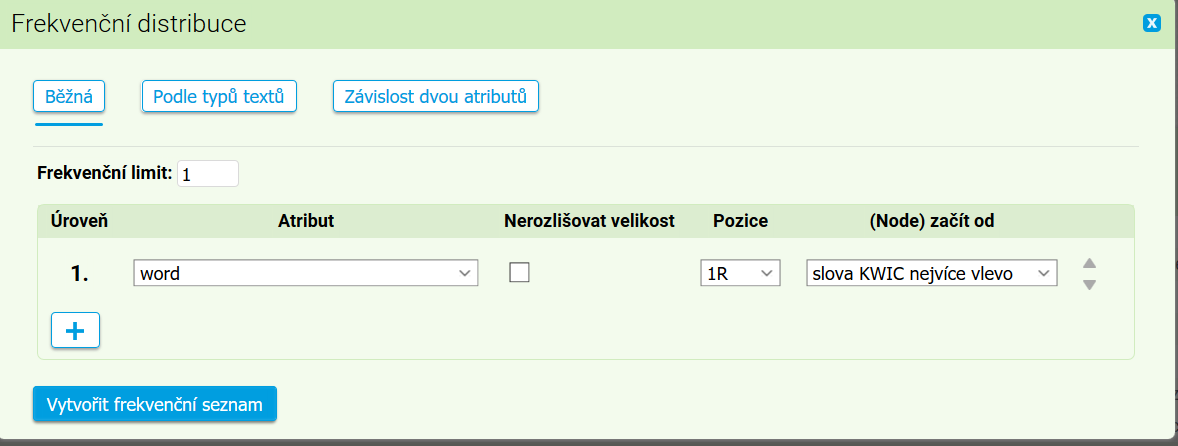
  | **Dotaz**: pila (24 179 výskytů



Podívejte se na frekvence slov nejvíce vlevo

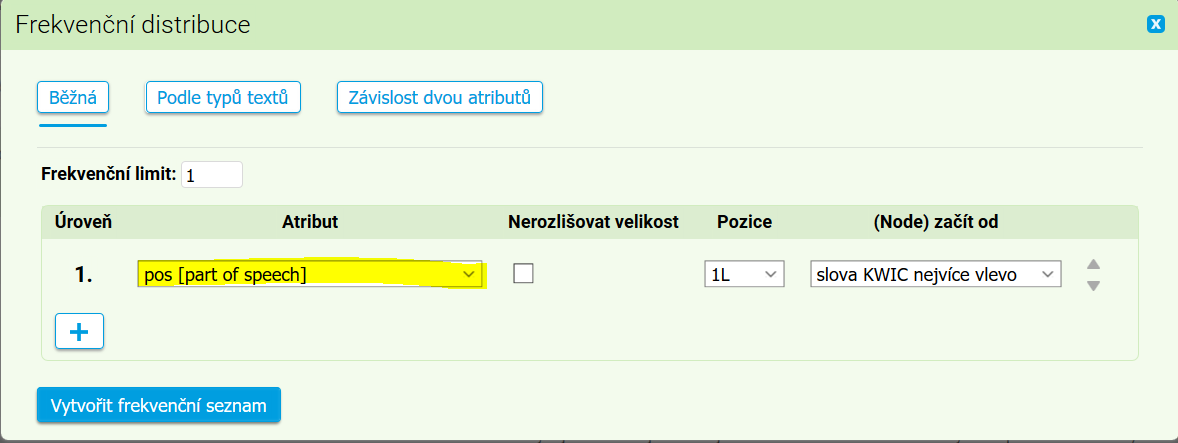


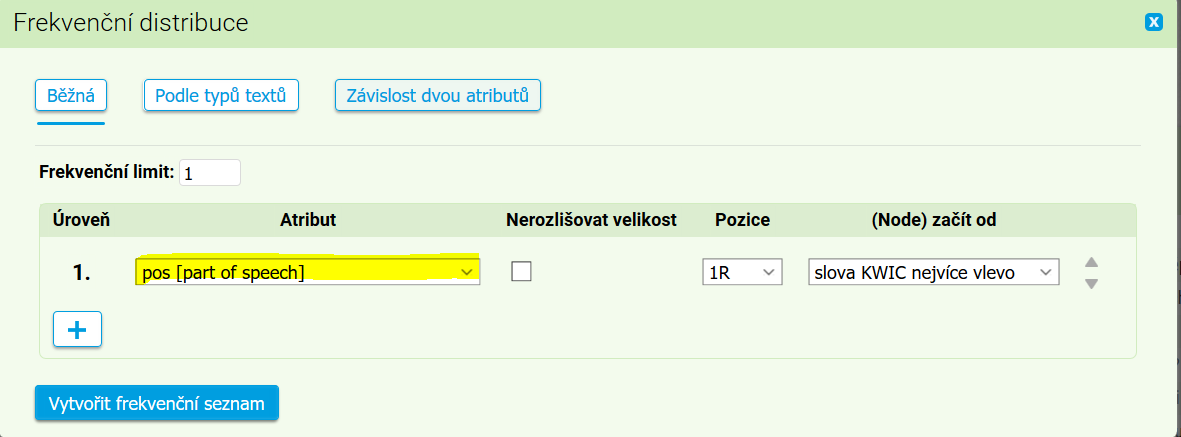
/vpravo



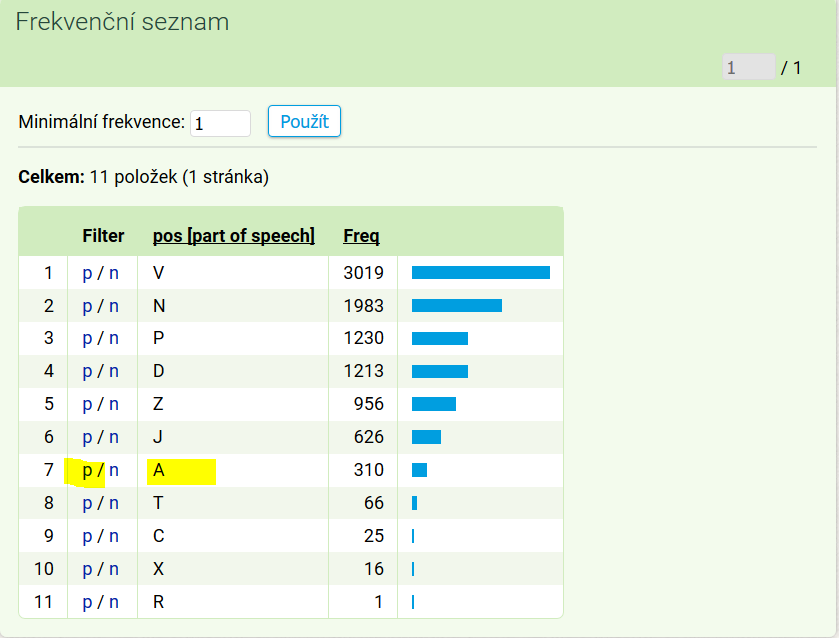
od klíčového slova.

Vyzkoušejte frekvenci slovnědruhového obsazení:





Pomocí p-filtru vyberte v levém kontextu tvaru *hnát* označkovaného jako POS=V (sloveso) řádky, na nichž je adjektivum (POS=A).





Pomocí p-filtru vyberte v pravém kontextu tvaru *hnát* označkovaného jako POS=N (substantivum) řádky, na nichž je zájmeno (POS=P).



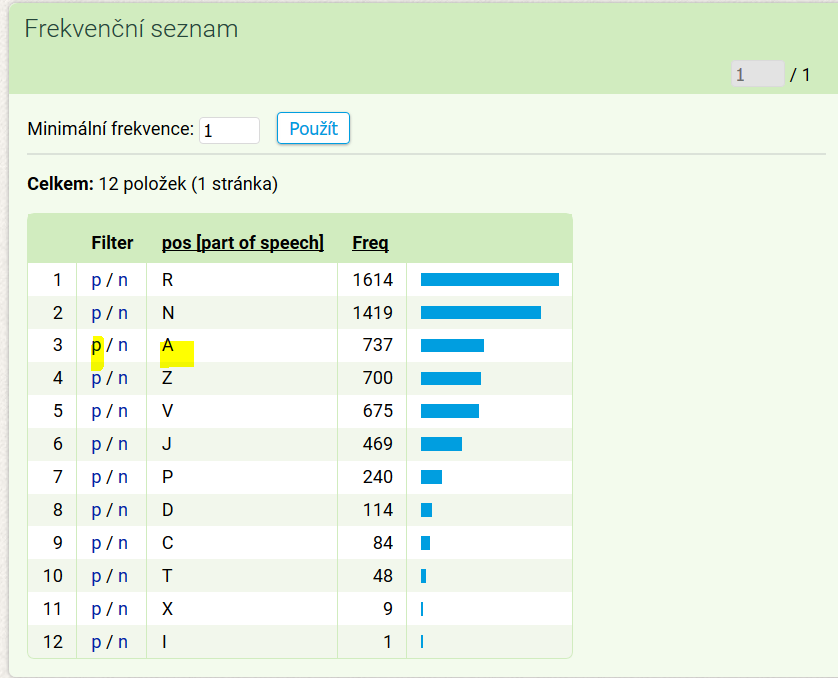
Má-li tvar, který lze interpretovat jako substantivum v levém kontextu tvar adjektiva, pak může, ale nemusí jít o tvar substantiva.

Má-li tvar, který lze interpretovat jako sloveso (může jít o zvratné sloveso) v pravém kontextu tvar zájmena, pak může, ale nemusí jít o tvar slovesa.

Odhalili jsme chyby v desambiguaci, ale nenašli jsme ani nutnou, ani postačující podmínku pro desambiguaci.

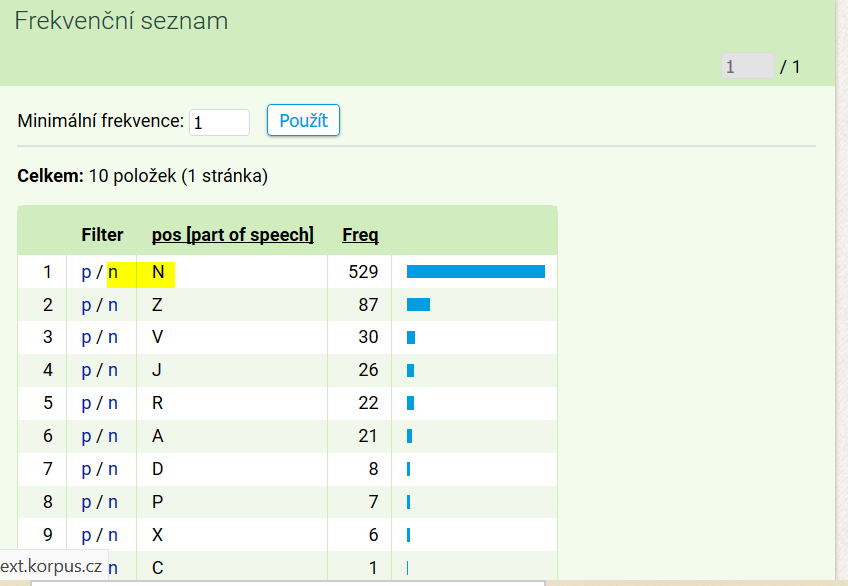
Podívejme se nyní na lemma=popravčí

Vyhledejme v levém kontextu lemmatu *popravčí* (je vždy označkována jako POS=A) adjektiva (POS=A).





Z uvedené konkordance vybereme řádky, na nichž vpravo od klíčového slova popravčí není substantivum:





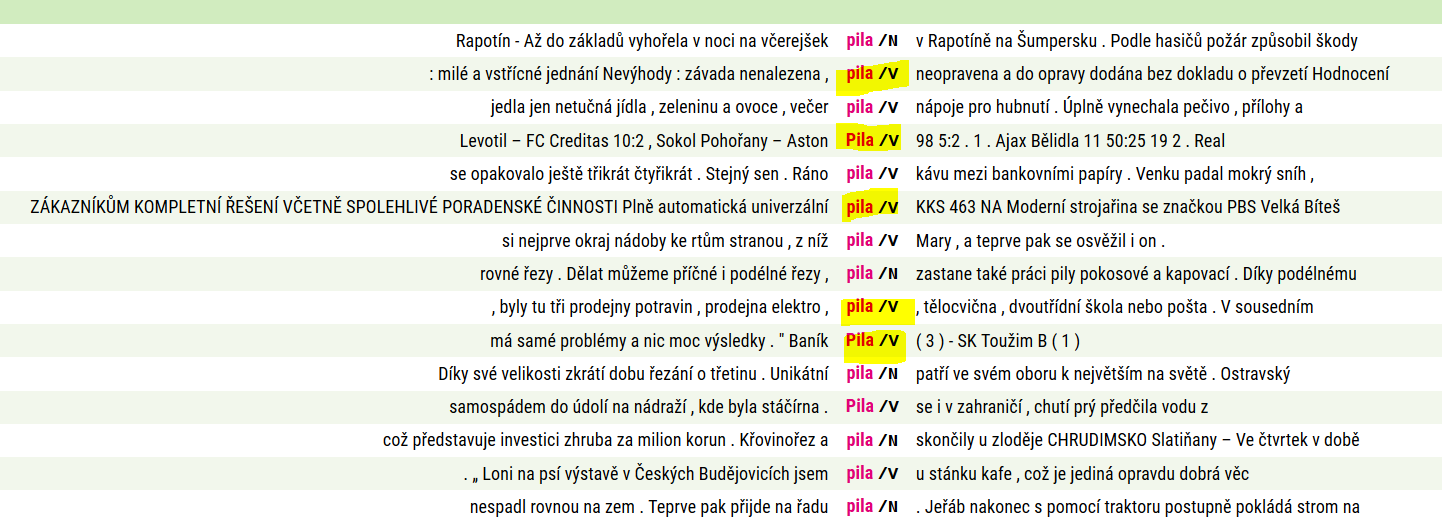
POZOR:

*… Například o lupičské skupině Erharta Pušky nás podrobně informují Jihlavské <****popravčí> a psanecké zápisy****.*

Má-li mít lemma *popravčí* interpretaci POS=N (s ní se ve značkování nepočítá, není zřejmě uložena ve slovníku automatického analyzátoru), pak v levém kontextu nechť stojí adjektivum a v pravém nechť nestojí substantivum. Nejde ovšem /jak jsme viděli ani o podmínku nutnou, ani dostačující.

Zatím jsme tedy odhalili jisté pravděpodobnosti, na jejichž základě dokážeme hledat chyby v desambiguaci.

Podívejme se nyní na interpretace tvaru *pila*.



Všimněme si opět chyb: propria, případy těžko definovatelné kontextovými pravidly.

**Které syntaktické vlastnosti můžeme využít, abychom v korpusu našli chyby v disambiguaci?**

**Můžeme využít minimalistická slovosledná pravidla. Většinou se z nich vychází jako z pravděpodobných kombinací (bigramy, trigramy) značek ve statistických desambiguátorech.**

**Úkol na 18. 5. *„Místo vašich schůzek sis měl lépe vybírat.“***

**Sleduj v korpusu slovnědruhové značkování jednotek (lc=kolem, lc=místo, lemma=výchovné), které lze interpretovat vícero způsoby a pokus se navrhnout, jak postupovat při odhalení chyb v disambiguaci.**