Zpráva k úkolu č. 4 z PLIN033

Richard Holaj

Cílem této úlohy bylo nalézt kompozita mající jako svůj první člen číslovky a jako druhý člen adjektiva s časovým významem. U těchto kompozit jsme pak chtěli také zjistit, které číslovky se vážou s více než jedním adjektivem. V první řadě bylo nutno zvolit adjektiva s časovým významem, která chceme zkoumat, a následně definovat dotaz. Dotaz se skládá z vyhledání libovolné číslovky s volitelným sufixem a adjektiva, které má společnou počáteční část s číslovkou, tato část je následována jedním z námi vybraných adjektiv. Celý dotaz ukazuje Obrázek 1.

Obrázek 1

Výsledek tohoto dotazu (seřazený podle frekvence) nám ukazuje Tabulka 1. Zajímavostí tohoto výsledku je, že dotaz nepřegenerovává, což není příliš obvyklá situace. Mohli bychom diskutovat o tom, zda nepodgenerovává, ale dotaz je dostatečně obecný na to, aby pro námi zvolená adjektiva vracel adekvátní výsledek.

Co se týče množství adjektiv vázaných s jednotlivými číslovkami, všimněme si, že méně frekventované jevy často obsahují pouze jednu vazbu, ve výjimečných případech vazby dvě, a to obvykle se dvěma z následujících tří slov: denní, hodinový a minutový. Obvykle se jedná o základní číslovky větší než deset. Výjimku tvoří slova utvořená od číslovek mnoho, čtvrt, tři čtvrtě a překlepu pŮl, dále pak kompozita odvozená od rozsahu 2–3 a číslice 10. Jsou zde zastoupeny číslovky jak ve formě slov, tak ve formě číslic.

Situace u frekventovanějších kompozit (bráno do pozice 21 včetně) je zcela odlišná. Nacházíme zde prvních deset základních číslic, přičemž se všechny váží minimálně se třemi adjektivy, od číslovky dva, respektive od jejího genitivu, je derivováno dokonce pět slov. Pouze dvě číslovky se zde váží jen s jedním adjektivem, a to číslovky více (s adjektivem denní) a čtyřiadvacet (zde je typická frekventovaná vazba s adjektivem hodinový), a jen číslovka devadesát se váže se dvěma slovy (minutový a denní). Ostatní číslovky se váží minimálně se třemi adjektivy. Mimo již zmíněných se zde vyskytují číslovky, které se váží k frekventovaným časovým údajům (patnáct, třicet, dvanáct a čtrnáct), rovněž číslovky dvacet a čtyřicet, které se váží se čtyřmi adjektivy (vteřinový, minutový, hodinový a denní).

V neposlední řadě jsou zde zastoupena kompozita číslovky půl (jsou dokonce nejfrekventovanější) vázající se se třemi adjektivy (minutový, hodinový, denní, roční) a kompozita od neurčité číslovky několik, která sice nejsou nejfrekventovanější (nachází se na pátém místě), ale jsou, spolu s kompozity od číslovky dva, jako jediná utvořena od pěti adjektiv, a jsou tak pro tento slovotvorný proces nejproduktivnější.

1	půl (1023)	půldenní půlminutový půlroční půlhodinový (392)
2	dvou (29489)	dvouroční dvouvteřinový dvoudenní dvouminutový dvouhodinový (383)
3	jedno (20015)	jednominutový jednoroční jednodenní (350)
4	tří (11191)	tříroční tříminutový tříhodinový třívteřinový třídenní (322)
5	několika (25227)	několikaminutový několikahodinový několikaroční několikavteřinový několikadenní (239)
6	čtyř (5446)	čtyřminutový čtyřhodinový čtyřdenní čtyřroční (135)
7	deseti (7804)	desetidenní desetihodinový desetivteřinový desetiminutový (114)
8	šesti (6159)	šestivteřinový šestihodinový šestiminutový šestidenní (111)
9	pěti (10559)	pětivteřinový pětiminutový pětidenní pětihodinový (102)
10	čtrnácti (1376)	čtrnáctiminutový čtrnáctidenní čtrnáctihodinový (86)
11	třiceti (2703)	třicetidenní třicetivteřinový třicetiminutový (51)
12	sedmi (4136)	sedmiminutový sedmihodinový sedmidenní (50)
13	patnácti (2448)	patnáctiminutový patnáctidenní patnáctihodinový (43)
14	více (25453)	vícedenní (43)
15	osmi (3898)	osmihodinový osmidenní osmiminutový (33)
16	dvaceti (4640)	dvacetihodinový dvacetidenní dvacetivteřinový dvacetiminutový (27)
17	čtyřiceti (1578)	čtyřicetivteřinový čtyřicetihodinový čtyřiceti minutový čtyřicetidenní (21)
18	devadesáti (469)	devadesátiminutový devadesátidenní (21)
19	dvanácti (2112)	dvanáctihodinový dvanáctidenní dvanáctiminutový (20)
20	čtyřiadvaceti (326)	čtyřiadvacetihodinový (19)
21	devíti (2445)	devítidenní devítihodinový devítiminutový (17)
22	mnoha (15739)	mnohadenní mnohahodinový (15)
23	šedesáti (884)	šedesátidenní šedesátihodinový šedesátiminutový (13)
24	jedenácti (1217)	jedenáctihodinový jedenáctidenní jedenáctiminutový (9)
25	šestnácti (1105)	šestnáctihodinový šestnáctidenní (8)
26	padesáti (1979)	padesátidenní padesátiminutový (8)
27	třinácti (996)	třináctiminutový třináctidenní (6)
28	čtvrt (726)	čtvrtroční čtvrthodinový (6)
29	sto (2496)	stodenní stominutový (4)
30	pětačtyřiceti (167)	pětačtyřicetidenní pětačtyřiceti minutový (4)
31	30 (13678)	30minutový (4)
32	osmadvaceti (143)	osmadvaceti minutový osmadvacetidenní (3)
33	sedmdesáti (549)	sedmdesátiminutový sedmdesátidenní (3)
34	pětatřiceti (286)	pětatřiceti minutový pětatřicetihodinový (3)
35	pŮl (1)	pŮlroční pŮldenní (2)
36	mnoho (22644)	mnohodenní mnohohodinový (2)
37	dvaadvaceti (265)	dvaadvaceti minutový (2)
38	devatenácti (435)	devatenáctihodinový devatenáctidenní (2)
39	sedmnácti (754)	sedmnáctidenní sedmnáctihodinový (2)
40	osmdesáti (583)	osmdesátiminutový osmdesátihodinový (2)
41	jedenadvaceti (135)	jedenadvacetidenní (2)
42	pětadvaceti (696)	pětadvacetihodinový pětadvaceti minutový (2)
43	20 (20438)	20minutový (2)
44	dvaačtyřiceti (82)	dvaačtyřicetidenní (1)
45	sedmadvaceti (137)	sedmadvacetihodinový (1)
46	pětašedesáti (94)	pětašedesátiminutový (1)
47	osmnácti (1053)	osmnáctidenní (1)
48	dvaatřiceti (109)	dvaatřiceti minutový (1)
49	10 (19732)	10minutový (1)
50	15 (13879)	15minutový (1)
51	třičtvrtě (50)	třičtvrtěhodinový (1)

52	2-3 (281)	2-3roční (1)
53	čtyřiatřiceti (54)	čtyřiatřicetidenní (1)
54	40 (8090)	40minutový (1)
55	45 (2977)	45minutový (1)
56	60 (7879)	60vteřinový (1)
57	77 (1092)	77minutový (1)
58	90 (5779)	90minutový (1)

Tabulka 1

Zdroje

- Čermák, F. – Blatná, R. – Hlaváčová, J. – Klímová, J. – Kocek, J. – Kopřivová, M. – Křen, M. – Petkevič, V. – Schmiedtová, V. – Šulc, M.: SYN2000: žánrově vyvážený korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha 2000. Cit.04.10.2016, dostupný z WWW: <http://www.korpus.cz>.

- Jan Hajič: *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Vol. 1. Karolinum Charles University Press, Praha 2004.

- Tomáš Jelínek (2008): Nové značkování v Českém národním korpusu. In: *Naše řeč*, 91, 1, pp. 13-20.

- Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, Pavel Květoň: The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In:*Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. ACL 2007, Praha. pp. 67-74.

- Vladimír Petkevič (2006): Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In: *Insight into the Slovak and Czech Corpus Linguistics (Šimková M. ed.)*. Veda, Bratislava, pp. 26-44.