Tokenizálás

A különbség okai:

- zárójelben (() vagy []) lévő mondat utolsó szava utáni pontot az e-magyar a szó részének veszi (nem külön token a pont), a huspacy pedig nem 153
- (1) : huspacy szerint 3 token, e-magyar szerint 2: (és 1) 7
- hármas kötőjeles szó: e-magyar szerint a kötőjelek és a szavak is külön tokenek (5 token), huspacy szerint pedig az egész 1 db token 5
- szóközzel tagolt számok: huspacy külön tokennek veszi őket szóköz szerint, az e-magyar viszont egyben kezeli a számokat, 1 db tokent készít 5
- ?! huspacynél 2 db token, e-magyarnál 1 db 2
- sorszám (13., 2022.): e-magyar szerint külön token a pont, huspacy szerint nem 3
- regex (footnoteRef:1): e-magyar szerint külön token a : és az 1, huspacy szerint nem 1
- h) alpont típusú szavak: huspacy szerint 2 token, e-magyar szerint 1 db 2
- kisbetűs mondatkezdés előtti utolsó szó: e-magyar a pontot a szóhoz veszi, huspacy külön tokennek 1
- Zrt. mondat közben: huspacy felismeri, hogy rövidítés, 1 tokennek veszi, e-magyar viszont különszedi 1
- ?, (alkalmi összekerülés): huspacy külön, e-magyar egyben számítja 2
- mondatvégi szám: a huspacy sorszámnak veszi (1 token), az e-magyar viszont külön 1
- 1-3. típusú sorszámok: huspacynél 1 token, e-magyarnál a pont külön token 1
- || (két elválasztó célú | véletlenül egymás mellett): a huspacynél 1 db token, az e-magyarnál 2 db 1
- trükkös e-mail-címek (pl. webmaster(kukac)parlament.hu): huspacy a zárójelek mentét különveszi (5 token), e-magyar nem (1 token) 3

szempont	huspacy	emagyar	db előfordulás
zárójelben lévő mondat utolsó szava	a pont külön token	a pont nem külön token	153
(1)	(, 1,)	(, 1)	7
hármas kötőjeles összetétel	az egész 1 token	a kötőjelek és a szavak is külön tokenek (5 db token)	5
szóközzel tagolt számok	külön tokenek	egy token	5
?!	2 token	1 token	2
sorszám	a pont nem külön token	a pont külön token	3

regex (footnoteRef:1)	1 db token	külön token a : és az 1	1
h) alpont típusúak	2 token	1 token	2
kisbetűs mondatkezdés előtti utolsó szó	a pont külön token	a pont a token része	1
Zrt. mondat közben	1 token	2 token	1
?, alkalmi kapcsolat	külön tokenek	1 token	2
mondatvégi szám	1 token	2 token	1
1-3. típusú sorszámok	1 token	a pont külön token	1
|| (két elválasztó célú | véletlenül egymás mellett)	1 token	2 token	1
trükkös e-mail-címek (pl. webmaster(kukac)parlament.hu)	zárójelek mentén külön tokenek (5 db)	1 token az egész	3

Súlyozás nélkül:

15 szempont
huspacy jobb: 6 db
emagyar jobb: 5 db
véleményes: 4 db
huspacy a jobb
össz eset 40%-a
(emagyar: össz eset 33,33%-a)
egyértelmű eset 54,54%-a

Súlyozással:

188 hibahely volt a szövegben
huspacy jobb: 169 db
emagyar jobb: 11 db
véleményes: 8 db
huspacy a jobb
össz eset 89,89%-a
(emagyar: össz eset 5,85%-a)
egyértelmű eset 93,89%-a