KERESÉS AZ IRÁNYÍTOTT BESZÉLGETÉSEKBEN

A BUSZI-2 KORPUSZLEKÉRDEZŐ

Sass Bálint
sass.balint@nytud.mta.hu

BUSZI-2 bemutató
Budapest, 2012. október 2.

Beszélt nyelvi korpusz

- A korpusz adatbázis = a hasznos információ expliciten, egyértelműen, egységes szerkezetben, számítógéppel hatékonyan feldolgozható formában van tárolva.
- BUSZI-2: részletes lejegyzés, a jelenségek széles köre.
  Automatikus módon hozzáadott információ:
  - regularizált alak
  - egyértelműsített morfológiai elemzés, szótő
  - regularizált szótő CV váza
  - elhangzott szóalak fonetikai reprezentációja
- *Cél:* számítógéppel segített elemzés
  A nyelvi adatbázis/lekérdező adatot szolgáltat ehhez:
  - az összes releváns adatot visszaadja
    → statisztikai elemzés lehetősége
  - összetett, részletes lekérdezőfelület

KORPUSZ: EGYSÉGEK SOROZATA

PÉLDA

. . . bizonyos dógokban □ mmm tát, hogy öööö lustább annál, mint amilyennek elkép*zel*tem, . . .

- *egység:* szó, szünet, hezitáció stb.
- korpusz = egységek sorozata
- *lekérdezés:* adott tulajdonságú egységeket keresünk
  pl.: *asztal* összes alakja, *t*-végű szó, szünet, *öööö*, *l*-kiesést tartalmazó szó, *-bAn* helyett *-bA* . . .
- *eredmény:* a találati egységek listája

EGYSZERŰ LEKÉRDEZÉS: *fontosnak*

Keressük meg a *fontosnak* szó elő fordulásait a korpuszban!

2012-05-24 10:59:14
Lekérdezés: `[W FOCUS surface = 'fontosnak']`
Lekérdezés lókusz-jelöléssel: `[W FOCUS who ~ '^[a-d]k' and surface = 'fontosnak']`
Találati szavak száma: 7 (korrigálás nélkül) – Futási idő: 2s

[1] B7102 / MUN / 78 / ak
*[hesit_length_n]*_nnnamost azért nem jele nem jelentett nehézséget , *[t_drop_final]*_mer mint mondottam eléggé válogatott *[o_hesitation]* gyerekek *[o_hesitation]* jöttek ide . *[P]* És ezeknek a gyereknek a nagy része *[P]* *[o_hesitation]* úgy jött ide , mint ahogy most a gyerekek gimnáziumba mennek , *[P]* hogy *[o_hesitation]* *[P]* *[o_hesitation]* *[P]* nevezetesen azzal a szándékkal , *[P]* hogy tovább *[o_hesitation]* szeretnének tanulni majd egyetemen . *[P]* Tehát ők *[o_hesitation]* *[P]* **fontosnak** tartot*[P]*ták az olyan tárgyak tanulását is , *[P]* amelyek majd az egyetemen , vagy főiskolán *[hesit_length_m]*_nemmm lesznek *[o_hesitation]* *[o_hesitation]* már *[hesit_length_m]*_nemmm nem léteznek , és és *[o_hesitation]* és *[d_drop_final]*_maj nem *[P]* már *[o_hesitation]* már nem élnek , *[t_drop_final]*_min nevezetesen a magyar nyelv és irodalom *[P]* tehát az általános műveltségű *[o_hesitation]* műveltségükhöz hozzátartozik . És *[P]* hát *[o_hesitation]* vö voltak annyira érelmesek , hogy *[P]* hogy *[o_hesitation]* *[P]* megértették azt , hogy *[o_hesitation]* *[P]* *[o_hesitation]* *[o_hesitation]* nagyon fontos az anyanyelv ismerete , *[P]* mert anélkül nem lehet fejlődni és tanulni . *[P]* Nagyon fontos az anyanyelv ismerete és idegen nyelv ismerete is nagyon fontos . *[P]* A természetesen a *[P]* *[hesit_length_a]*_aaa *[o_hesitation]* *[hesit_length_o3]*_elsőőő helyre teszem a mamtematikát , *[P]* a matematika , a fizika és a műszaki tárgyak ismeretén kívül és ezek mellett .

[2] B7114 / VAL / 613 / ak
Hát mit mondjak , ha az ember nem *[o_hesitation]* *[P]* *[o_hesitation]* teljes mértékben vagy százszázalékban a *[P]* *[hesit_length_v]*_wallás alapján hívő , *[P]* akkor az első *[o_hesitation]* nem tudom egy vagy kettőt aligha fogja *[o_hesitation]* *[m_hesitation]* **fontosnak** tartani , *[P]* de mindazt amim ami emberi és humánus benne

[3] B7207 / VAL / 251 / ak
Én **fontosnak** tartom ((> (azt)))> .

[4] B7301 / VAL / 289 / ak
<((A vallást <)) nem tartom **fontosnak** .

[5] B7303 / VAL / 231 / ak
Igen , ((> **fontosnak** .))>

[6] B7512 / QQQ / 961 / ak
<((Hát ilyen <)) szereléseket , azt tartom **fontosnak** , de azt el is mondtam .

[7] B7515 / NYE / 642 / ak
Hát **fontosnak** fontos , *[P]* persze .

EGYSZERŰ LEKÉRDEZÉS: *fontosnak*

[1] B7102 / MUN / 78 / ak

*[hesit_length_n]*_nnnamost azért nem jele nem jelentett nehézség jöttek ide . *[P]* És ezeknek a gyereknek a nagy része *[P]* *[o_hesit* *[P]* *[o_hesitation]* *[P]* nevezetesen azzal a szándékkal , *[P]* hogy **fontosnak** tartot*[P]*ták az olyan tárgyak tanulását is , *[P]* amelyek *[o_hesitation]* már *[hesit_length_m]*_nemmm nem léteznek , és é *[t_drop_final]*_min nevezetesen a magyar nyelv és irodalom *[P]* te *[o_hesitation]* vö voltak annyira érelmesek , hogy *[P]* hogy *[o_hes* anyanyelv ismerete , *[P]* mert anélkül nem lehet fejlődni és tanulni természetesen a *[P]* *[hesit_length_a]*_aaa *[o_hesitation]* *[hesit_le* tárgyak ismeretén kívül és ezek mellett .

[2] B7114 / VAL / 613 / ak

Hát mit mondjak , ha az ember nem *[o_hesitation]* *[P]* *[o_hesitatio* az első *[o_hesitation]* nem tudom egy vagy kettőt aligha fogja *[o_*

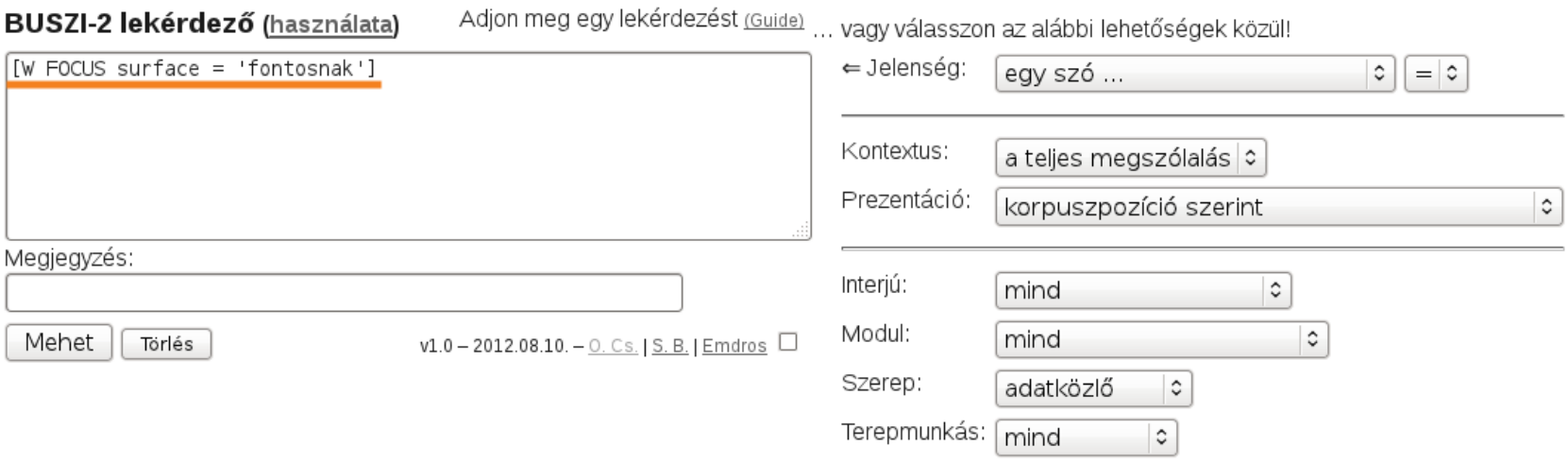
A LEKÉRDEZŐFELÜLET RÉSZEI

1. kereshető nyelvi jelenségek menüben
2. lekérdezésmező (szerkeszthető!)
3. megjelenítés beállításai
4. alkorpuszok kiválasztása
5. a lekérdezésmezőben található lekérdezés futtatása
6. *összeállítás-vezérlő*

Előző példa


BUSZI-2 lekérdező (használata)
Adjon meg egy lekérdezést (Guide)
Megjegyzés:
Mehet
Törlés
v1.0 – 2012.08.10. – O. Cs. | S. B. | Emdros
… vagy válasszon az alábbi lehetőségek közül!
⇐ Jelenség:
egy szó …
=
↳ Regularizált alak:
Felszíni alak (teljes):
fontosnak
Szótő (teljes):
Elemzés:
Szótő CV-váz (teljes):
Felszíni fono-váz: (teljes):
OK
Kontextus:
a teljes megszólalás
Prezentáció:
korpuszpozíció szerint

Előző példa


BUSZI-2 lekérdező (használata)
Adjon meg egy lekérdezést (Guide)
[W FOCUS surface = 'fontosnak']
Megjegyzés:
Mehet
Törlés
v1.0 – 2012.08.10. – O. Cs. | S. B. | Emdros
… vagy válasszon az alábbi lehetőségek közül!
⇐ Jelenség:
egy szó …
=
Kontextus:
a teljes megszólalás
Prezentáció:
korpuszpozíció szerint
Interjú:
mind
Modul:
mind
Szerep:
adatközlő
Terepmunkás:
mind

Előző példa

2012-05-24 10:59:14
Lekérdezés: `[W FOCUS surface = 'fontosnak']`
Lekérdezés lókusz-jelöléssel: `[W FOCUS who ~ '^[a-d]k' and surface = 'fontosnak']`
Találati szavak száma: 7 (korrigálás nélkül) – Futási idő: 2s

[1] B7102 / MUN / 78 / ak
*[hesit_length_n]*_nnnamost azért nem jele nem jelentett nehézséget , *[t_drop_final]*_mer mint mondottam eléggé válogatott *[o_hesitation]* gyerekek *[o_hesitation]* jöttek ide . *[P]* És ezeknek a gyereknek a nagy része *[P]* *[o_hesitation]* úgy jött ide , mint ahogy most a gyerekek gimnáziumba mennek , *[P]* hogy *[o_hesitation]* *[P]* *[o_hesitation]* *[P]* nevezetesen azzal a szándékkal , *[P]* hogy tovább *[o_hesitation]* szeretnének tanulni majd egyetemen . *[P]* Tehát ők *[o_hesitation]* *[P]* **fontosnak** tartot*[P]*ták az olyan tárgyak tanulását is , *[P]* amelyek majd az egyetemen , vagy főiskolán *[hesit_length_m]*_nemmm lesznek *[o_hesitation]* *[o_hesitation]* már *[hesit_length_m]*_nemmm nem léteznek , és és *[o_hesitation]* és *[d_drop_final]*_maj nem *[P]* már *[o_hesitation]* már nem élnek , *[t_drop_final]*_min nevezetesen a magyar nyelv és irodalom *[P]* tehát az általános műveltségű *[o_hesitation]* műveltségükhöz hozzátartozik . És *[P]* hát *[o_hesitation]* vö voltak annyira érelmesek , hogy *[P]* hogy *[o_hesitation]* *[P]* megértették azt , hogy *[o_hesitation]* *[P]* *[o_hesitation]* *[o_hesitation]* nagyon fontos az anyanyelv ismerete , *[P]* mert anélkül nem lehet fejlődni és tanulni . *[P]* Nagyon fontos az anyanyelv ismerete és idegen nyelv ismerete is nagyon fontos . *[P]* A természetesen a *[P]* *[hesit_length_a]*_aaa *[o_hesitation]* *[hesit_length_o3]*_elsőőő helyre teszem a mamtematikát , *[P]* a matematika , a fizika és a műszaki tárgyak ismeretén kívül és ezek mellett .

[2] B7114 / VAL / 613 / ak
Hát mit mondjak , ha az ember nem *[o_hesitation]* *[P]* *[o_hesitation]* teljes mértékben vagy százszázalékban a *[P]* *[hesit_length_v]*_wallás alapján hívő , *[P]* akkor az első *[o_hesitation]* nem tudom egy vagy kettőt aligha fogja *[o_hesitation]* *[m_hesitation]* **fontosnak** tartani , *[P]* de mindazt amim ami emberi és humánus benne

[3] B7207 / VAL / 251 / ak
Én **fontosnak** tartom ((> (azt)))> .

[4] B7301 / VAL / 289 / ak
<((A vallást <)) nem tartom **fontosnak** .

[5] B7303 / VAL / 231 / ak
Igen , ((> **fontosnak** .))>

[6] B7512 / QQQ / 961 / ak
<((Hát ilyen <)) szereléseket , azt tartom **fontosnak** , de azt el is mondtam .

[7] B7515 / NYE / 642 / ak
Hát **fontosnak** fontos , *[P]* persze .

Jelenségek listája

- egy szó
  regularizált alak (*tát* → *tehát*), felszíni alak, szótő, elemzés, szótő CV-váza (BNF), felszíni alak fonetikai reprezentációja
- szón belüli pozícióval bíró jelenségek = kiesések (*l*, *t*, *d*, *ly*)
  pozíciók: szóvégi, V•V, C•V, V•C, C•C
- pozíció nélküli jelenségek
  pl.: *-bAn* helyett *-bA*, nem állítmányi *-e* . . .
- önálló egységek
  pl.: szünet, megjegyzés, hezitáció . . .
- egyéb: az összes megszólalás

MEGJELENÍTÉS

1. (rendezett) konkordancia: a találati megszólalások listája
2. gyakorisági lista
3. összesítő táblázat

MEGJELENÍTÉS: GYAKORISÁGI LISTA

Lekérdezés: `[W FOCUS reg = 'majdnem']`
Lekérdezés lókusz-jelöléssel: `[W FOCUS who ~ '^[a-d]k' and reg = 'majdnem']`
Találati szavak száma: 60 (korrigálás nélkül) – Futási idő: 2s

*[d_drop_ic]*_majnem	38 db
majdnem	10 db
*[d_drop_ic]*_Majnem	4 db
Majdnem	3 db
*[hesit_length_m]*_mmmajdnem	2 db
*[d_drop_ic hesit_length_m]*_majnemmm	1 db
*[d_drop_ic hesit_length_M]*_Mmmajnem	1 db
*[d_drop_ic hesit_length_m]*_mmmajnem	1 db

MEGJELENÍTÉS: GYAKORISÁGI LISTA

Lekérdezés: `[W FOCUS reg = 'majdnem']`
Lekérdezés lókusz-jelöléssel: `[W FOCUS who ~ '^[a-d]k' and reg = 'majdnem']`
Találati szavak száma: 60 (korrigálás nélkül) – Futási idő: 2s

*[d_drop_ic]*_**majnem**	38 db
majdnem	10 db
*[d_drop_ic]*_**Majnem**	4 db
Majdnem	3 db
*[hesit_length_m]*_**mmmajdnem**	2 db
*[d_drop_ic hesit_length_m]*_**majnemmm**	1 db
*[d_drop_ic hesit_length_M]*_**Mmmajnem**	1 db
*[d_drop_ic hesit_length_m]*_**mmmajnem**	1 db

Majnem mindig kiesik a *d*.
d-kiesés 75% ↔ *d* megvan 25%

MEGJELENÍTÉS

1. (rendezett) konkordancia: a találati megszólalások listája
2. gyakorisági lista
3. összesítő táblázat

MEGJELENÍTÉS: ÖSSZESÍTŐ TÁBLÁZAT

Lekérdezés lókusz-jelöléssel: [Annot FOCUS who ~ '^[a-d]k' and typ = 'd_drop']
Találati szavak száma: 421 (jelenségre korrigálva: 421) – Futási idő: 4s

	tanárok	egyetemisták	bolti eladók	gyári munkások	szakmunkástanulók	∑
ABO	–	–	–	3	–	3
ÁLM	–	–	–	–	1	1
ATO	–	–	–	–	1	1
BAR	–	–	–	–	8	8
BIO	1	11	5	5	4	26
BŰN	2	–	–	–	1	3
CMÖ	1	5	6	4	13	29
CSA	3	1	4	5	6	19
DEM	–	–	–	–	–	–
ETN	4	11	4	2	1	22
FÉL	–	–	3	–	–	3
GYE	–	–	–	–	–	–
HAL	4	4	4	1	1	14
HÁZ	–	–	9	4	3	16
HUM	1	–	1	–	5	7
ISK	2	15	1	1	10	29
JÁT	–	–	–	2	6	8
KIS	1	5	2	4	2	14
MOZ	2	3	2	1	6	14
MUN	6	5	26	10	7	54
NFK	3	6	4	–	12	25
NYE	3	4	9	3	9	28
QQQ	–	1	2	6	9	18
RIP	–	–	–	–	–	–
SZE	1	–	1	1	–	3
SZI	1	4	10	–	23	38
TEM	2	6	3	3	2	16
VAL	1	–	2	7	3	13
VER	–	–	–	3	6	9
∑	38	81	98	65	139	421

Alkorpuszok

3 független dimenzió:

1. interjú
2. modul
3. szerep: adatközlő vagy terepmunkás

Alkorpuszok

3 független dimenzió:

1. interjú – adott interjú, kvóta, terepmunkás személye
2. modul
3. szerep: adatközlő vagy terepmunkás

Szóalapú – jelenségalapú

A korpusz *szóalapú* (egységalapú)
→ a lekérdezések adott tulajdonságú egységeket adnak vissza.

Gond:

A pozícióval bíró jelenségekből több is lehet egy szóban.
Nyilván szeretnénk tudni az ilyenek összesített számát.

A megoldás: *korrigálás*

KORRIGÁLÁS

Lekérdezés: l-kiesés a B7114-es interjú család moduljában

2012-05-24 12:27:58
Lekérdezés: `[Annot FOCUS typ ~ 'l_dr..']`
Lekérdezés lókusz-jelöléssel: `[Annot FOCUS who ~ '^[a-d]k' and modul ~ 'CSA' and interview ~ 'B7114' and typ ~ 'l_dr..']`
Találati szavak száma: 3 (jelenségre korrigálva: 4) – Futási idő: 3s

[1] B7114 / CSA / 91 / ak
Gyermekeink ? *[P]* Van egy fiam *[P]* Jézus Mária ! az már ***[l_drop_precons l_drop_final]*_körübelü** harminc éves , *[P]* és van egy lányom , aki eggyel fiatalabb . *[P]* Foglalkozásuk ?

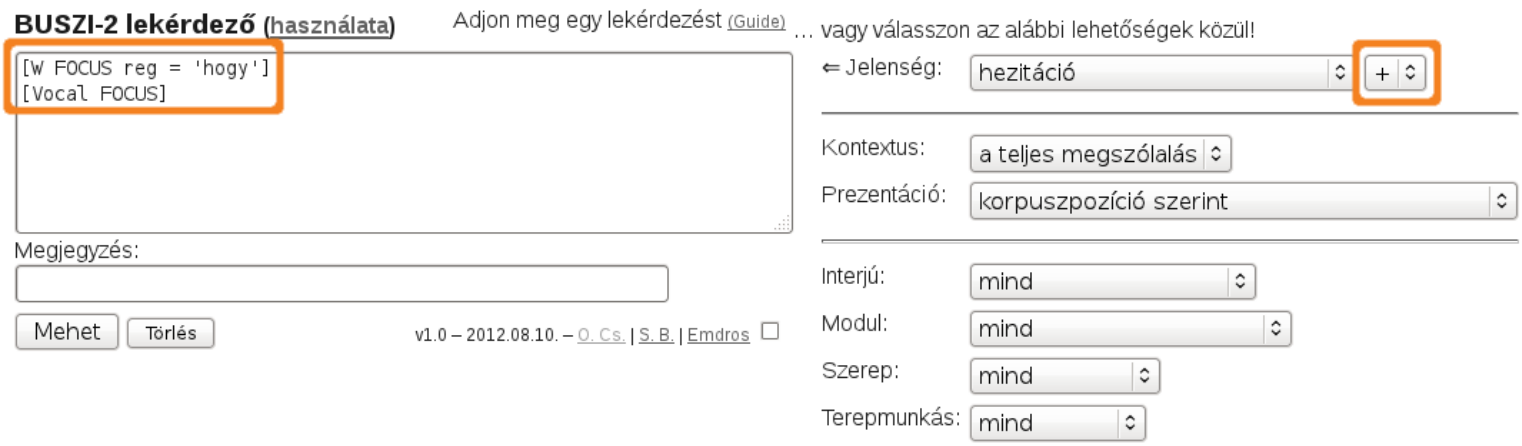
[1/a] B7114 / CSA / 91 / ak
Gyermekeink ? *[P]* Van egy fiam *[P]* Jézus Mária ! az már ***[l_drop_precons l_drop_final]*_körübelü** harminc éves , *[P]* és van egy lányom , aki eggyel fiatalabb . *[P]* Foglalkozásuk ?

[2] B7114 / CSA / 93 / ak
*[hesit_length_H]*_Hhh a fiam *[hesit_length_z]*_azzz vegyészmérnök , *[P]* *[o_hesitation]* *[hesit_length_e]*_deee szerencsére nem mindenben követi az apja nyomdokait , mert *[P]* *[o_hesitation]* nincs benne olyan mérhetetlen ambíció . *[P]* Egy kicsit tud *[nevetve:]* élni is . *[P]* A másik *[o_hesitation]* kedvenc ***[l_drcl_precons]*_foglakozása** *[P]* hát a vegyészet mellett és vegyészkedés mellett a zene . ((> *[unspec]*_T))>

[3] B7114 / CSA / 97 / ak
Aztán a lányom az geológus , *[P]* *[hesit_length_o3]*_őőő *[hesit_length_b]*_bbbölcsész akart lenni . Arról lebeszéltem , *[mély levegőt vesz]* *[P]* és jól tettem , az egyéniségéhez jobban illik ez . *[P]* Szereti a ***[l_drcl_precons]*_foglakozását** , *[P]* a férje agrármérnök ,

TÖBB EGYSÉGRE KITERJEDŐ LEKÉRDEZÉS: *hogy + öööö*

jelenségek *sorozatára* irányuló lekérdezés
összeállítás-vezérlő: ’+’

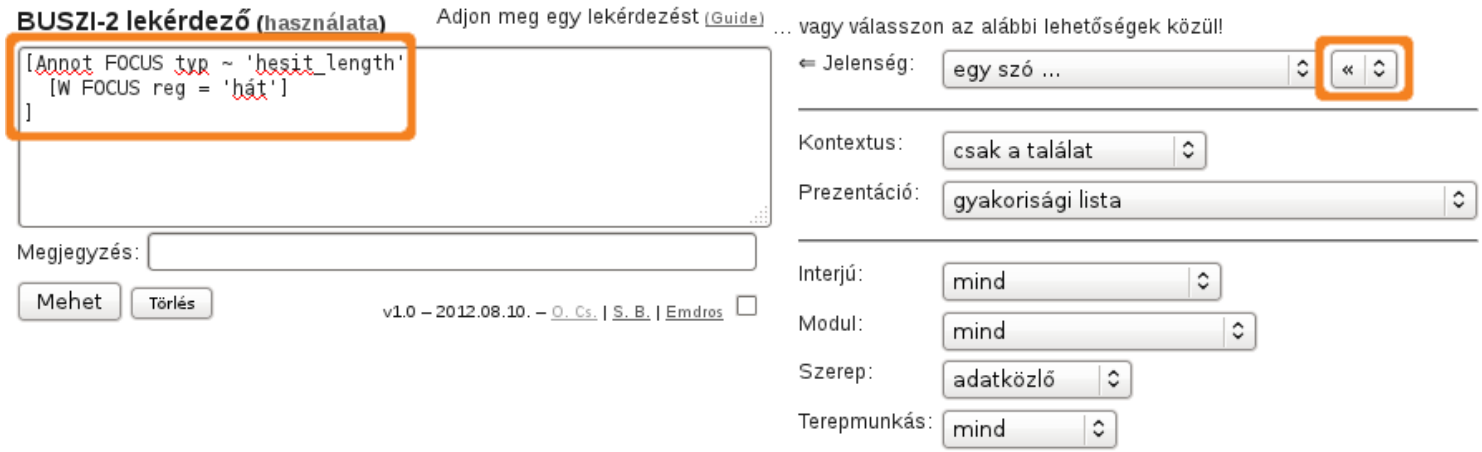


2012-10-02 12:13:17
Lekérdezés: `[W FOCUS reg = 'hogy'] [Vocal FOCUS]`
Találati szavak száma: 717 (korrigálás nélkül) – Futási idő: 4s

[1] B7102 / MUN / 2 / ak
Én 1957 szeptemberétől tanítok , *[sóhajt]* mivel **hogy** ***[o_hesitation]*** 1957-ben végeztem az ### ### Tudományegyetem *[P]* *[hesit_length_sz]*_Bölcsésssz *[P]* karán *[P]* a magyar-orosz szakot .

ADOTT JELENSÉG ADOTT SZÓN: *hááát*

1. jelenség: hezitációs hangzónyúlás
2. összeállítás-vezérlő: ’«’
3. szó: *hát* (regularizált alak)



2012-10-02 12:56:19
Lekérdezés: `[Annot FOCUS typ ~ 'hesit_length' [W FOCUS reg = 'hát'] ]`
Lekérdezés lókusz-jelöléssel: `[Annot FOCUS who ~ '^[a-d]k' and typ ~ 'hesit_length' [W FOCUS who ~ '^[a-d]k' and reg = 'hát'] ]`
Találati szavak száma: 197 (korrigálás nélkül) – Futási idő: 2s

*[hesit_length_tj]*_Háttt	75 db
*[hesit_length_H]*_Hhhát	48 db
*[hesit_length_tj]*_háttt	30 db
*[hesit_length_h]*_hhhát	21 db
*[hesit_length_a1]*_Hááát	12 db
*[t_drop_final hesit_length_H]*_Hhhá	2 db
*[t_drop_final hesit_length_a1]*_Hááá	2 db
*[hesit_length_a1]*_hááát	2 db
*[hesit_length_T unspec]*_Ttt	1 db
*[hesit_length_tj]*_*Háttt	1 db
*[hesit_length_a1]*_*Hááát	1 db
*[hesit_length_a1 hesit_length_tj]*_Háááttt	1 db

Összefoglalás

- Hozzáférhető és kutatható a BUSZI-2 anonimizált irányított beszélgetéseit tartalmazó 270000 szavas korpusz.
- A korpuszkezelő rendszer segítségével nyelvészetileg releváns lekérdezéseket fogalmazhatunk meg a hozzá tartozó lekérdezőfelülettel.
- Kereső: `http://buszi.nytud.hu/keresoprogramok`
- Részletes leírás a lekérdező használatáról: `http://corpus.nytud.hu/buszi/buszilekerdezo.html`

Összefoglalás

- Hozzáférhető és kutatható a BUSZI-2 anonimizált irányított beszélgetéseit tartalmazó 270000 szavas korpusz.
- A korpuszkezelő rendszer segítségével nyelvészetileg releváns lekérdezéseket fogalmazhatunk meg a hozzá tartozó lekérdezőfelülettel.
- Kereső: `http://buszi.nytud.hu/keresoprogramok`
- Részletes leírás a lekérdező használatáról: `http://corpus.nytud.hu/buszi/buszilekerdezo.html`

Köszönöm a figyelmet!

`http://buszi.nytud.hu`