A BUSZI-2 lekérdező használata

Sass Bálint
v0.7.2 – 2012. május

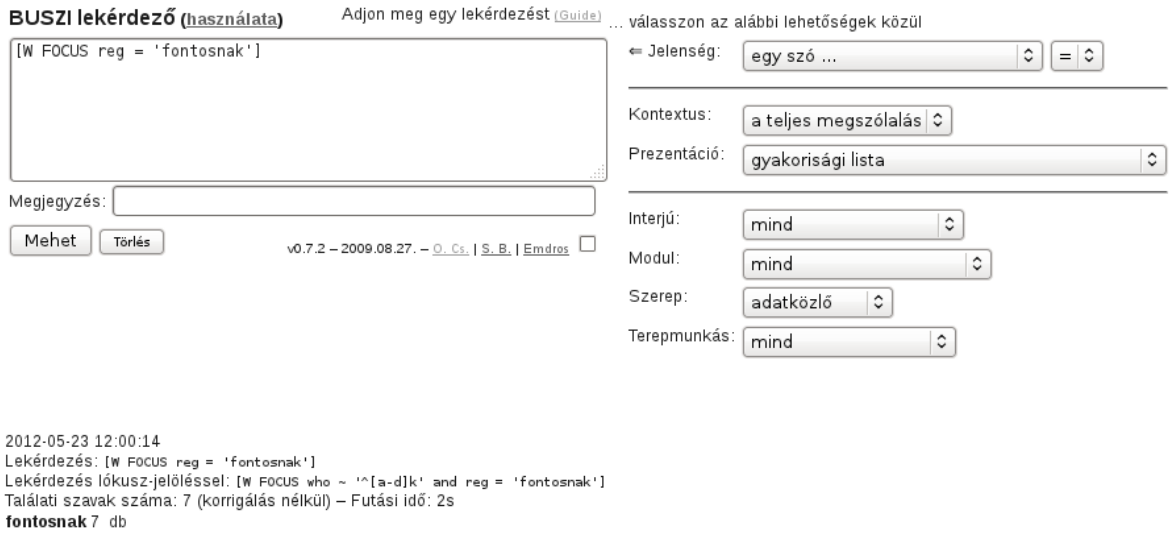
1. Első lépések

1.1. Bevezető példa

Arra vagyunk kíváncsiak, hogy a *fontosnak* szóalak hányszor fordul elő a BUSZI-2 korpuszban.

1. A `http://camel.nytud.hu/buszidemo` XXX oldalon a *Jelenség* menüből válasszuk ki a (legelső) *egy szó...* bejegyzést.
2. A megjelenő felületen a *Felszíni alak*-hoz írjuk be, hogy: *fontosnak*, majd nyomjuk meg az *OK*-t. Ekkor az imént megjelent felület eltűnik, és a megfelelő lekérdezés a bal oldali szövegmezőbe kerül.
3. Futtassuk le a lekérdezést a *Mehet* gomb megnyomásával.

A szűkszavú eredmény arról számol be, hogy a korpuszban hétszer szerepel a kérdezett szó.



1.2. A felület részei

A `http://camel.nytud.hu/buszidemo` XXX címen elérhető BUSZI-2 lekérdezőfelület felépítését az alábbi ábrán mutatjuk be.

1. A felület középső részén felül található legördülő menü tartalmazza az összes kereshető nyelvi jelenséget.
2. A bal oldali szövegmező (*lekérdezésmező*) a futtatandó lekérdezés összeállításának helye.
3. Középen, a nyelvi jelenségeket tartalmazó menü alatt kaptak helyet a megjelenítés beállításai...
4. ...és az alkorpuszok kiválasztására szolgáló felület.
5. A *Mehet* gomb futtatja le a lekérdezésmezőben található lekérdezést.
6. A jobb oldalon fent található *Összeállítás-vezérlő*ről később lesz szó (ld.: a 2.6. és a 2.7 részt).

1.3. A használat menete

A BUSZI-2 lekérdező használata alapesetben a fenti ábrán található a számoknak megfelelő sorrendben történik. A következő lépésekből áll:

1. A kereshető jelenségeket a felület középső részén felül található összetett menürendszerből választhatjuk ki, ennek segítségével állíthatjuk össze a lekérdezésünket,...
2. ...mely automatikusan megjelenik a lekérdezésmezőben. Itt a lekérdezés futtatás előtt szerkeszthető, illetve a lekérdezőnyelv ismeretében közvetlenül, a *Jelenség* menü használata nélkül is megfogalmazható itt egy lekérdezés.
3. A lekérdezés futtatása előtt beállíthatjuk megjelenítés paramétereit...

4. . . . valamint szükség esetén megadhatjuk, hogy mely alkorpuszra akarjuk korlátozni a lekérdezést.

5. Végül a *Mehet* gomb megnyomásával futtathatjuk le a lekérdezésmezőben lévő lekérdezést. (Azaz ha ez a szövegmező üres, akkor hiába van bármi beállítva a menüben, nem kapunk eredményt.)

A lekérdezés eredménye a képernyő alsó felében jelenik meg.

2. Részletes leírás

2.1. Jelenségek

A BUSZI-2 korpuszban lévő minden bekódolt, lekérdezhető, nyelvészetileg releváns információ a *jelenség* egységes fogalma alá van besorolva. A *Jelenség* menüben a következő kategóriákat találjuk:

- *egy szó:* itt egy adott felszíni alakú, szótövű stb. szóra kereshetünk rá, a keresett szó tulajdonságait részletesen meghatározhatjuk;
- *kihagyás:* ld.: a 2.6. részt.
- *annotációk pozícióval:* itt olyan bekódolt jelenségeket találunk, melyekből egy szóban több is lehet, ennek megfelelően a jelenség szóbeli pozíciójára is külön rákérdezhetünk;
- *annotációk:* itt olyan bekódolt jelenségeket találunk, melyek egy szóban maximum egyszer fordulhatnak elő;
- *önálló egységek:* itt az önálló szószintű egységekre (szünet, hezitáció stb.) kereshetünk rá;
- *egyéb / az összes megszólalás:* itt egy speciális lehetőség kapott helyet, mikor *nem* egy adott szóra, hanem (adott részkorpuszban) az összes megszólalásra keresünk rá, ez ad lehetőséget az interjúk szövegének folytatólagos olvasására.

A BUSZI-2 korpusz alapegysége a szó. A korpusz szavak (illetve szószintű egységek) sorozatának tekinthető. A lekérdezések eredménye – az összes megszólalásra irányuló lekérdezés kivételével – mindig a találati szavak listája.

Alább csak azokat a jelenségeket tárgyaljuk, melyeknél az adott jelenség tulajdonságait egy megjelenő kiegészítő felületen lehet megadni (e jelenségek neve a legördülő menüben három pontra végződik). Az ilyen kiegészítő felület kitöltése után mindig meg kell nyomni a hozzá tartozó *OK* gombot ahhoz, hogy a kívánt jelenség a lekérdezésmezőbe kerüljön!

2.1.1. Egy szó

Egy szót számos különféle jellemzőjük alapján kereshetjük.

A *Regularizált alak* az adatközlő által kimondott szó szokásos, kanonikus írott alakja. A *Felszíni alak* az elhangzott szó hangképéhez legközelebb álló írásos megjelenítés, amit a lejegyzők alkalmaztak. A *tát* felszíni alakhoz például a *tehát* regularizált alak tartozik.

A *Szótő*vet és a morfológiai *Elemzés*t a regularizált alak alapján automatikus nyelvi elemzés határozta meg. A morfológaiai elemzésben a Magyar Nemzeti Szövegtárban is használatos kódokat használtuk (`V` – ige, `N` – főnév, `A` – melléknév stb.). A kódrendszerről részletesen itt lehet tájékozódni:
`http://corpus.nytud.hu/mnsz/sugo_hun.html#msdrendszer`

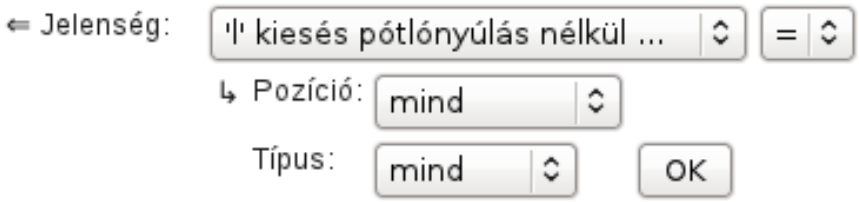
További kiegészítő jellemző a regularizált *Szótő CV-váz*a, a mássalhangzók jele a `C`; a magánhangzókat `V`-vel, illetve képzési hely szerint `B` (hátulképzett), `N` (semleges), `F` (elölképzett) kódokkal is jelölhetjük. Végül megadhatjuk az egy szóra irányuló keresést az elhangzott szóalak fonetikai reprezentációja (*Felszíni fono-váz*) alapján is. Itt minden hangnak egy egykarakteres jel felel meg. Az egy betűs hangok jele a megfelelő kisbetű, a további jelölések a következők:

hang(kapcsolat)	cs	dz	gy	ly	ny	sz	ty	zs	dzs	x	qu	ch	y (i-ként)	mgh	mssh
jel	F	D	G	J	N	S	T	Z	X	KS	KW	H	I	V	C

Egy szóra irányuló keresésnél a fenti jellemzőket kombinálni is lehet, megadható például a morfológiai elemzés és a fonetikai váz együttesen.

2.1.2. Annotációk pozícióval

Ahogy említettük, itt olyan bekódolt nyelvi jelenségeket találunk, melyek egy szóban többször is előfordulhatnak. A különféle hangkiesések tartoznak ide. A vizsgálatok szempontjából az is érdekes lehet, hogy az adaott kieső hang a szó mely részén illetve milyen környezetben volt, ezért a felület biztosítja az erre való rákérdezés lehetőségét.



A kiesés *Pozíció*ja lehet szóvégi; szóbelseji kiesés esetén pedig megadhatjuk, hogy magánhangzó/mássalhangzó követte illetve előzte meg az adott kiesést. A kiesés *Típusá*nál elkülöníthetjük azt az esetet, mikor hosszú mássalhangzó esik ki (*2 esik ki*), valamint mikor a hosszú mássalhangzó rövidül (*rövdidülés*).

2.2. Megjelenítés

A *Kontextus*nál beállíthatjuk, hogy mekkora szövegkörnyezettel – esetleg az egész megszólalással együtt – kérjük a találati szavakat.

A *Prezentáció*nál kiválaszthatjuk, hogy a találati adatokat milyen formában jelenítse meg a lekérdező. A *gyakoriségi lista* csak a találati szavakból készül, itt a bővebb kontextust figyelmen kívül hagyja a rendszer. Az *összesítés*ben egy táblázatot kapunk kvóták és modulok szerint a találati számokból. Ez a számszerű adatok összevetését könnyíti meg. *Rendezett konkordancia* esetén az egyes találatok sorra egymás alatt jelennek meg, és a pontos korpuszpozíció megjelölésével, a kért kontextussal, az összes bekódolt jelenség feltüntetésével.

2.3. Alkorpuszok

A keresést három független dimenzió szerint szűkíthetjük alkorpuszra. A BUSZI-2 50 interjúja közül bármelyiket külön is vizsgálhatjuk, illetve lehetőség van adott adatközlő-csoport (ún. *kvóta*: tanárok, egyetemisták, bolti eladók, gyári munkások, szakmunkástanulók) 10 interjújának egyben való vizsgálatára (*Interjú*). Szűkíthetjük a keresést adott *Modul*ra is, azaz az interjúknak csak azon részeire, ahol bizonyos a terepmunkások által kötelezően érintett témákról esik szó. Végül megadhatjuk, hogy az adatközlő és/vagy a terepmunkás által mondottakra vonatkoztatjuk a lekérdezést (*Szerep*). Az alapbeállítás itt a terepmunkást kizárja, azaz nem a teljes korpuszra, hanem csak az adatközlők nyelvi produkciójára vonatkozik.

2.4. A konkordancia elemei

A 2.2. részben említett prezentációs lehetőségek közül csak a konkordancia igényel részletes magyarázatot. Az alábbi ábrán az 1.1. részben említett lekérdezés eredménye látható, de most nem gyakorisági listaként, hanem konkordanciaként.


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]*_vvallás 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 .

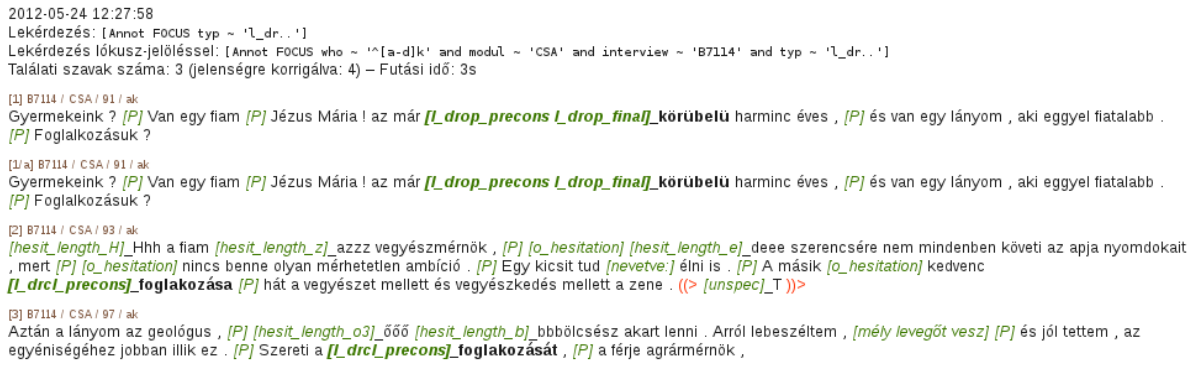
[7] B7515 / NYE / 642 / ak
Hát **fontosnak** fontos , *[P]* persze .

A fejlécben szerepel a találati szám, majd a találatok következnek szövegkörnyezettel (az ábrán a teljes megszólalással) együtt. Az egyes találatok fejlécében található négy adat pontosan megadja az adott nyelvi adat korpuszbeli pozícióját. Ezek: az interjú azonosítója, a modul azonosítója, a megszólalás interjún belüli sorszáma, valamint, hogy adatközlőtől vagy terepmunkástól származik az adat. A szövegben találati szó félkövérrel van kiemelve. A szavakhoz kapcsolt illetve önálló zöld kódok (pl.: `[hesit_length_n]` – hezitációs *n*-nyúlás; `[t_drop_final]` – szóvégi *t*-kiesés; `[o_hesitation]` – hezitáció (*öööö*); `[P]` – szünet stb.) a bekódolt nyelvi jelenségeket jelenítik meg (ld. még: 2.1. rész, *Jelenség* menü). A narancssárga kódok az egyszerre elhangzó beszéd szakaszait jelölik meg.

2.5. Jelenségre korrigálás

A 2.1. részben említettük, hogy a korpusz alapegysége a szó, a lekérdezések adott tulajdonságú szavakat adnak vissza, a lekérdezések eredménye a találati szavak listája. Ezek szerint minden szó csak *egy* találatot jeleníthet meg. Ez problémát okoz azoknál a jelenségeknél, melyek egy szóban többször is előfordulhatnak (ld. *annotációk pozícióval* a 2.1. részben), ugyanis nyilván érdekes lehet ezek összesített száma. A megoldást egy korrigáló lépés jelenti, melynek eredményeképpen ilyen esetekben ha egy jelenség egy szóban kétszer/többször szerepel, akkor az adott szó kétszer/többször fog megjelenni a találati listán is. Ilyenkor a fejlécben a *Találati szavak száma* mellett megjelenik a jelenségek száma is – az ún. *jelenségre korrigált* érték –, lehetővé téve azt, hogy a felhasználó a számára szükséges értékkel számolhasson.

Ha a B7114 interjú család (CSA) moduljában keressük meg az *l*-kieséseket, akkor 3 találati szón 4 darab találatot kapunk, mivel a *körülbelül* szóban két független *l*-kiesés történt:



2.6. Több szóra kiterjedő lekérdezés

Eddig mindvégig egy szóra, illetve az egy szóban lévő valamilyen jelenségre kerestünk rá. Természetes az igény a bonyolultabb, több szóból álló, több szóra kiterjedő lekérdezésekre.

A több egységből álló lekérdezések összeállítását teszi lehetővé a felületen a *Jelenség* menü mellett látható *összeállítás-vezérlő* elem (ld.: 1.2. rész, ábra XXX, 6.). Ha ez (alapbeállítás szerint) ′=′-re van állítva, akkor – amint ezt eddig mindig láttuk –, az aktuálisan megadott szó/jelenségre vonatkozó beállítás egészében felülírja a korábbi lekérdezést (a lekérdezésben), azaz ezáltal ugye egy új lekérdezést adhatunk meg. Ha viszont az összeállítás-vezérlőt ′+′-ra állítjuk, akkor kiegészíti a lekérdezésmezőben már korábban meglévő lekérdezés-részletet egy újabbal. Ezen a módon tehát több egységből álló lekérdezéseket tudunk felépíteni.

Olyan több elemű lekérdezés esetén, melyben a megadott elemek nem közvetlenül érintkeznek, hanem közöttük egyéb tetszőleges elem(ek) fordulhat(nak) elő, szükséges a *Jelenség* menüben található speciális *kihagyás* lehetőség használata.

2.7. Adott jelenség adott szón

Az eddigiekben vagy egy adott szóra, vagy egy adott jelenségre (mely természetesen mindig egy szón jelenik meg) kerestünk rá. Arra is van lehetőség, hogy egy jelenségnek csak egy adott (tulajdonságú) szón való előfordulását keressük. Ehhez először meg kell adnunk a jelenséget (valamilyen *annotációt*), majd az összeállítás-vezérlőt ′«′-re állítva a szót (az *egy szó...* segítségével). A lekérdezésmezőben a kombinált lekérdezés fog megjelenni, és eredményül a kívánt jelenségnek azon előfordulásait kapjuk, melyekben a kívánt jelenség a kívánt szón fordul elő.

2.8. Összefoglaló példa

Arra vagyunk kíváncsiak, hogy milyen konfigurációban fordul elő egymást követően egy hezitációs hangzónyúlást tartalmazó *hogy* szó, és egy önnálló hezitáció (*öööö*).

A lekérdezést a következőképpen építjük fel:

1. Összeállítás-vezérlő: *'='*.
2. *Jelenség:* hezitációs hangzónyúlás.
3. Összeállítás-vezérlő: *'«'*.
4. *Jelenség/egy szó…/Regularizált alak: hogy*; utána *OK*.
5. Összeállítás-vezérlő: *'+'*.
6. *Jelenség/kihagyás:* minimum 0, maximum 3 szó; utána *OK*.
7. *Jelenség:* hezitáció.

Ennek eredményeképpen a lekérdezőmezőben a következő lekérdezés áll elő:

```
[Annot FOCUS typ ~ 'hesit_length'
  [W FOCUS reg = 'hogy']
]
.. BETWEEN 0 AND 3
[Vocal FOCUS]
```

Ezt lefuttatva 12 találatot kapunk.