MoMa keresési útmutató

Bevezető

A MoMa moldvai magyar beszéltnyelvi, nyelvjárási korpusz online lekérdezőfelülete szabadon elérhető bárki számára a https://corpus.moma.nytud.hu címen.

A lejegyzett, normalizált, morfológiailag elemzett korpuszt a NoSketchEngine (NoSkE) korpuszkezelő rendszerben tesszük elérhetővé. Napjainkban a NoSkE a legelterjedtebb, de facto standard korpuszkezelő rendszer, melyet lassan két évtizede folyamatosan fejlesztenek.

A megszokott „beírok egy szót, és kijön valami" keresési módszeren túl számos egyéb keresési lehetőséget nyújt, melyek a módszertanilag megfelelő korpusznyelvészeti kutatáshoz elengedhetetlenek. A klasszikus konkordanciák megjelenítésén túl a két alapvető feldolgozó művelethez – a gyakorisági listák készítéséhez és a találatok további lekérdezésekkel való szűréséhez – is kényelmes és rugalmas megoldást ad. Lehetőség a találati listák mentésére további feldolgozás céljára, véletlen mintavételre, kollokációs vizsgálatra. Kiemelendő a reguláris kifejezésekre épülő Corpus Query Language (CQL) formális lekérdezőnyelv: ez az az eszköz, mellyel a korpuszban rejlő információ teljességéhez hozzá lehet férni.

A rendszer a szöveges annotációs szintek kezelése mellett lehetőséget ad arra is, hogy az egyes szövegrészekhez hozzáillesszük a megfelelő hanganyagrészletet, így ezeket az egyes konkordanciasoroknál egy kattintással meghallgathatjuk.

Érdemes részletesen megismerkedni a NoSkE használatával, mert számos más korpusz lekérdezőfelülete is erre épül.

A rendszer ebben a dokumentumban nem tárgyalt aspektusairól, további funkciókról és részletekről az alábbi módokon tájékozódhatunk:

1. a felületen az egyes oldalakhoz tartozó ABOUT részben találunk leírást az adott oldalról;
2. a kattintható bekarikázott kérdőjelek segítségével: egyrészt az egyes részfunkcióknál, másrészt az oldal jobb felső sarkában;
3. a SketchEngine youtube csatornáján: https://www.youtube.com/@SketchEngine, érdemes megnézni a *Concordance BASIC*, a *Concordance ADVANCED*, a *Frequency* és a *CQL 1* videókat.

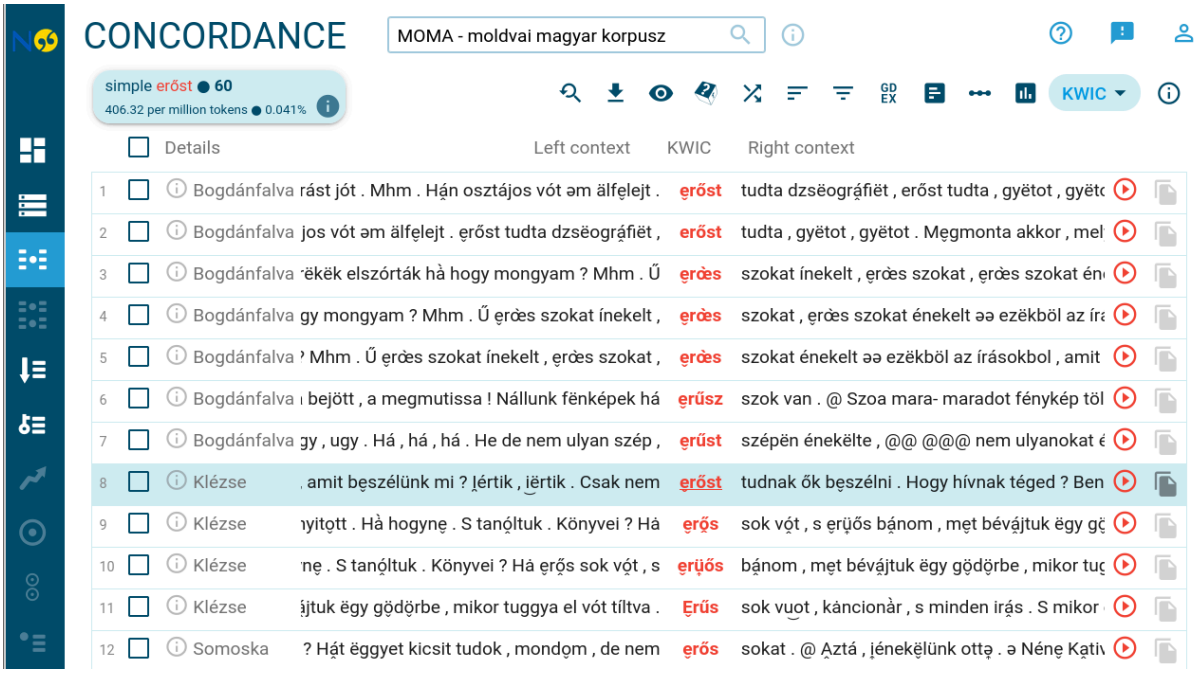
Gyakorlat

A felülettel való ismerkedéshez az alábbi gyakorlatokat érdemes lépésről lépésre végigcsinálni.

A keresés megkezdéséhez a nyitóoldalon az OPEN-t kell választani, majd a Concordance-t.

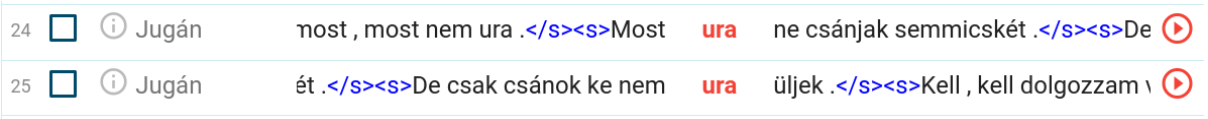
1. Egyszerű keresés. Az egyszerű keresésben (BASIC) egy vagy több szóalakot adhatunk meg. Az *erőst* alakot megadva az ábrán látható ún. konkordanciát kapjuk. Középen, egymás alatt pirossal láthatók a találati szavak (ún. KWIC-ek), a normalizálásnak köszönhetően az *erőst* összes lejegyzett variációját megkapjuk. Az eredeti/normalizált alak a felső menü harmadik gombjának használatával, ezen belül a Show attributes-ben a „eredeti"/„normalizált"

kiválasztásával jeleníthető meg. A kapcsolódó hanganyag az adott sor végén lévő piros lejátszás gombra kattintva hallgatható meg, a találat metaadatai pedig a sor elején látható információ gombra kattintva érhetőek el. A metaadatok tartalmazzák a találathoz tartozó interjú azonosítóját, a település nevét és koordinátáit, ahonnan az adat származik, illetve egy, a lejegyzések gyűjteményére mutató linket, amelyen keresztül az adott interjú teljes egészében elolvasható és folyamatában is meghallgatható.



Az *erőst* alak konkordanciája az eredeti lejegyzést megjelenítve.

Ha az *ura* szóra keresünk rá, a 24-25. találatban két rövid mondatban négy jellegzetes moldvai jelenségre is példát találunk: (1) *a (nem) ura vmit csináljon* speciális szerkezet, jelentése: '(nem) tud valamit csinálni'; (2) a *semmicskét* tipikus példa a gyakran használt kicsinyítésre; (3) a *ke* 'mert' egy szókincset érintő specialitás; a (4) *csán* pedig a *csinál* szokásos moldvai változata.



Jellegzetes moldvai nyelvi jelenségek az *ura* konkordanciájában. A szöveg kisebb egységeit `<s></s>` tagek jelölik, az ábrán a normalizált alak szerepel (ez a felső menü harmadik (szemecske) gombjának használatával, ezen belül a Show structures-ben állítható be).

A kiinduló keresési oldalra a felső menü első (nyilacskás nagyító) gombjával jutunk vissza.

2. Több szóra keresés és ragozott alakok. Az egyszerű keresésben több szóból álló kifejezést is meg lehet adni (pl. *Szűz Mária*), illetve fontos, hogy a megadott szót szóalakként és szótőként is keresi, azaz az összes ragozott alakot is meg fogjuk kapni a találatok között (pl. a *gyermek* szóra irányuló keresés keresés esetén: *gyermekek*, *gyermekeit* stb.).

3. Morfológiai jegyek. Lényegi tulajdonsága a MoMa korpusznak, hogy morfológiai jegyek (szófaj, esetrag, igemód stb.) alapján is tudunk keresni benne. Ehhez szükség van a morfológiai kódok és a CQL használatára. A beépített CQL BUILDER eszköz itt nagy segítségünkre van: lehetővé teszi a kívánt CQL lekérdezések összeállítását a CQL szintaxisának ismerete és a morfológiai kódok ismerete nélkül. Az összes igére például a következő módon kereshetünk rá:

- Válasszuk az egyszerű helyett most az összetett keresést (ADVANCED);
- kattintsunk a CQL-re, majd a CQL BUILDER-re, és indítás után válasszuk a „normal token”-t;
- az ige mint szófaj kiválasztásához a bal oldalon az Attribute-nál válasszuk az ELEMZÉS-t, a jobb oldalon a  menüben pedig görgessünk le az IGE bejegyzésig, majd pipa és USE THIS CQL;
- a keresőmezőben megjelenik az igék kikeresésére szolgáló CQL formula;
- kattintsunk a GO-ra, így megkapjuk az igék több mint 10000 elemű konkordanciáját, a morfológiai elemzésnek köszönhetően az összes ragozott alakot.

Hasonlóan működik a morfológiai alapú keresés tetszőleges morfológiai jegyre (pl. -val/-vel esetagos alak, feltételes mód, folyamatos múltidő stb.). A korpuszban használt morfológiai kódok teljes listája elérhető a https://nlp.nytud.hu/csango/morf_kod_lista linken.

4. Gyakorisági lista. Vizsgáljuk meg, hogy milyen gyakori az igék között ez a bizonyos *csán* tő!

Ehhez a fenti konkordanciából gyakorisági listát a felső menü kilencedik  gombjára, majd a KWIC alatti LEMMAS (szótő) gombra kattintva készíthetünk. LEMMAS helyett a WORD FORM (szóalak) választása esetén a igei szóalakok jelentősen több elemből álló gyakorisági listáját kapjuk (*mondja* és *mondják* külön bejegyzés lesz), az ADVANCED fülön az „eredeti” választása esetén pedig az eredeti lejegyzett alakok még több elemből álló gyakorisági listáját (*vót* és *volt* is külön bejegyzés lesz).

	Szótő ('lemma')	Frequency	Relative ?	
1	van	218	2,039.82	•••
2	tud	140	1,309.98	•••
3	mond	129	1,207.05	•••
4	megy	47	439.78	•••
5	imádkozik	37	346.21	•••
6	lesz	33	308.78	•••
7	csán	28	262.00	•••
8	jön	28	262.00	•••
9	énekel	22	205.85	•••
10	tanul	22	205.85	•••

Igetövek gyakorisági listája a MoMa korpuszban. Az anyag tematikáját jól mutatja az *imádkozik* és az *énekel* gyakorisága. A jellegzetes moldvai *csán* a hetedik helyen található.

5. Szűrés. A lekérdezések eredményét további lekérdezéssel szűkíthetjük, szűrhetjük. Erre a felső menü hetedik gombja szolgál.

Igealakot közvetlenül követő igekötőre (*jött le*, *vigye el*) például úgy kereshetünk, hogy az igék teljes listáját szűrjük azzal, hogy a találati szót (azaz az igét) követő pozícióban igekötő legyen.

- Ehhez induljunk ki a 3. pontban megkapott teljes igelistából;
- kattintsunk a felső menü hetedik  gombjára;
- válasszuk az ADVANCED lehetőséget;
- a CQL BUILDER segítségével válasszuk ki az igekötőket a 3. pontban leírtak szerint;
- a „Range”-nél állítsuk be, hogy az eredeti találati szóhoz (KWIC) képest hol helyezkedjen el a szűrésben meghatározott szó, ez mivel igét közvetlenül követő igekötőre keresünk, a KWIC-től jobbra eső „1” megjelölését jelenti;
- végül kattintsunk a GO-ra, így megkapjuk az igéket és az őket közvetlenül követő igekötők konkordanciáját.

A kutatások során a legtöbb esetben fontos, hogy kizárjuk a terepmunkás által mondottakat és csak az adatközlőkre korlátozzuk a keresést. Ezt a `[beszelo="ak"]` CQL kifejezéssel való KWIC-re vonatkoztatott szűrés révén érhetjük el.

6. Kollokációk. A rendszer kollokációkeresésre is alkalmas. Adjuk meg egyszerű keresésben a *szép* szót és kattintsunk a felső menü tizedik (három pötty) gombjára, majd a GO-ra. Megkapjuk a *szép* közvetlen környezetében előforduló legjellegzetesebb szavakat.

v1.0 – 2025.01.28.