„*A fatens felelt pedig…*” – A Történeti Magánéleti Korpusz igei szerkezeteinek mozaik n-gram alapú feldolgozása

Bajzát Tímea Borbála[1], Indig Balázs[123], Kalivoda Ágnes[4]

[1] Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar
Nyelvtudományi Doktori Iskola
bajzat.timi9696@gmail.com

[2] Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar
TI Digitális Bölcsészet Tanszék
indig.balazs@btk.elte.hu

[3] Digitális Örökség Nemzeti Laboratórium

[4] HUN-REN Nyelvtudományi Kutatóközpont
kalivoda.agnes@nytud.hun-ren.hu

Kivonat: A jelen tanulmány bemutatja a mozaik n-gram módszer alkalmazásának első eredményeit a Történeti Magánéleti Korpusz nyelvi anyagán. Az esettanulmány célja a finit igét tartalmazó elemi mondatok mintázatainak feltérképezése és konstrukciójelöltjeinek kinyerése. A funkcionális konstrukciós nyelvtan módszeréhez illeszkedő eljárás korábban alkalmazhatónak bizonyult infinitívuszos szerkezetek nyelvi mintázatainak azonosítására mai magyar nyelvi adatok elemzésénél. A konstrukciójelöltek félautomatikus módszerrel történő feltárása szükségszerű, mivel így kevésbé szükséges a nyelvész intuíciójára hagyatkozni. A mozaik n-gramok esetében a kinyert sorozatok elemei eltérő reprezentációs szinteken jelennek meg (szóalak, lemma, POS-tag), amely lehetővé teszi a nyelvi mintázatok eltérő absztrakciós szinteken történő azonosítását. Az eljárás diakrón korpuszokon történő alkalmazása segíti azt, hogy a konstrukciójelöltek feltárásán és egységes osztályozásán keresztül képesek legyünk hozzájárulni a magyar igei szerkezetek grammatikalizációs ösvényeinek adatalapú vizsgálatához. A tanulmány problémacentrikusan mutatja be az adaptációs folyamat kihívásaira (pl. annotációs séma egységesítése, elemi mondatok kinyerése) adott megoldásokat és a kapott eredményeket.

1 Bevezetés[1][2]

A *konstrukciós nyelvtan*ok (vö. Goldberg, 1995; Croft, 2001; Diessel, 2015) és *a mintázatnyelvtan* (vö. Hunston és Francis, 2000) felismerték, és a nyelvészeti diskurzus fókuszába helyezték, hogy az emberi nyelv mint tudás szimbolikus forma-jelentés párokból álló (Tolcsvai Nagy, 2017; Langacker, 1987), sémaalapú szerkezeti mintázatok hálózata. Ezek a *mentális konstrukciók* eltérő absztrakciós szinteken modellálhatók, ugyanakkor a használat során jönnek létre, vagy erősítik meg a saját mentális sémáik reprezentációit (Bybee, 2010; Tolcsvai Nagy, 2017: 57). Ezen felismerés egyértelművé teszi, hogy a funkcionális nyelvleírásban a gyakorisági értékeket is figyelembe vevő adatközpontú vizsgálati módszerek alkalmazása megkerülhetetlen kihívás napjainkban. Számos, az empirikus megismerést támogató eljárás született és került alkalmazásra korpuszalapú és korpuszvezérelt módszereket használva a nyelvi mintázatok azonosítására a funkcionális szemléletmódot érvényesítő kutatások körében (lásd például Luodonpää-Manni és mtsai, 2017 szerk.; Glynn és Robinson, 2014 szerk.; Hilpert és Flach, 2022 szerk.; Gries és Stefanowitsch, 2007; Simon, 2018). Azonban a nyelvi mintázatok eltérő absztrakciós szinteken történő egyidejű kinyerése nem problémamentes feladat. A megfelelő nyelvtechnológiai apparátus nélkül a nyelvész intuíciói által lesz irányított az adott lekérdezési és adatelemzési eljárás, amely gyengítheti az adatalapú kritériumnak való megfelelést. Ezen kívül az elmélet felől feltételezett mintázatoknak korpuszvezérelt eljárásokkal való visszaigazolhatósága szintén időszerű kérdéssé vált. A funkcionális kognitív nyelvészetben a (fél)automatikus módszerrel végzett *grammatikalizációs* kutatások (Bybee, 2010; Dér, 2008) szintúgy az érdeklődés előterébe kerültek (vö. például Heine és Narrog, 2021; Hilpert és Correia Saveedra, 2017; Kalivoda, 2021). Ez az igény olyan módszertanokat kíván meg, amelyek lehetővé teszik a különböző korpuszokból vett nyelvi adatok egységes kezelési módját és a belőlük kinyert mintázatok modellálását.

A jelen tanulmány bemutatja a *mozaik n-gram* alapú mintázatazonosítás koncepciójának leíró nyelvészeti alkalmazását az igés szerkezeti mintázatok korpuszvezérelt kinyerésén keresztül. A mozaik n-gram gondolata természetesen nem új (Indig és mtsai, 2016; Indig, 2017), azonban a nyelvleírásban még nem hasznosították szélesebb körben. A mozaik n-gramok esetében a szekvencia elemei eltérő reprezentációs szinteken jeleníthetők meg (pl. POS-tag, lemma, szóalak) attól függően, hogy milyen nyelvtechnológiai apparátust alkalmazunk a feldolgozás során (példaként szolgál a trigramokra az alábbi két mintázat: „nem lemma:tud [/V][Inf]” és „[/Prev] tudják [/V][Inf]”). A különböző absztrakciós szintek egyidejű jelenléte lehetőséget teremt az egymáshoz hasonló példányok csoportosítására, ezzel feltárva a forma–funkció felől a

[1] „Bajzát Tímea Borbála kutatása a Kulturális és Innovációs Minisztérium ÚNKP-23-3 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.”

[2] Kalivoda Ágnes kutatása az 142317 számú OTKA PD projekt keretében, az Innovációs és Technológiai Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a PD 22 pályázati program finanszírozásában valósult meg.

konstrukciójelöltet a nyelvi adatokból kiindulva, korpuszvezérelten[3]. Azonban ezt az eszközt eddig csak mai magyar nyelvi anyagon tesztelték (Indig és Bajzát, 2023). A jelen vizsgálat a módszer történeti korpuszokra való adaptációját kezdeményezi, elsőként a *Történeti Magánéleti Korpusz* (Novák és mtsai, 2018; Dömötör és mtsai, 2017) nyelvi anyagát dolgoztuk fel és állítottuk elő a benne előforduló, valamilyen finit igét tartalmazó *elemi mondat*ok (vö. Imrényi, 2017) mozaikjait. Ezzel célunk annak vizsgálata, hogy az adott nyelvi anyagra milyen gyakori igés mintázatok jellemzőek a különböző hosszúságú szekvenciákban. Másrészt a már meglévő eszköz és eljárás fejlesztése, valamint a nyelvtörténeti vizsgálatokban hasznosítható szerepének tárgyalása szintén a célkitűzéseink közé tartozik.

2 A mozaik n-gramok mint lehetséges konstrukciójelöltek

A mozaik n-gram módszernek abból származik a konstrukcióazonosításban betöltött lehetséges szerepe, hogy képes több ezer, vagy több százezer példányból álló minták (fél)automatikus elemzését elvégezni, másrészt pedig különböző absztrakciós szintek egyidejű megjelenítésével feltárja a potenciális *konstrukciójelölt*eket (Indig, 2017). Ezek az absztrakciós szintek a szóalak, szótő (lemma) és a POS-tag. Az, hogy a bemeneti adatokból a mozaik n-gramokat előállító algoritmus minden példány összes lehetséges mozaik sorozatát előállítja, lehetőséget teremt arra, hogy a valamiért egymással összetartozó kifejezésmódok csoportokként váljanak kezelhetővé (Indig, 2017), ezzel pedig forma–funkció párokként férhetünk hozzá a konstrukciójelöltekhez. Ez az adatfeldolgozási eljárás összhangban van a funkcionális nyelvészet által felállított elméleti kiindulópontokkal, például azzal, hogy a nyelvi forma és a funkció (jelentés) között *szimbolikus viszony* feltételezhető (Langacker, 2005: 103–107, 2008: 15–18; 2009; Tolcsvai Nagy, 2017: 52–53), tehát a formai oldalon tapasztalható eltérések és hasonlóságok valamilyen funkcionális összetartozást vagy elkülönbözést jelölnek, s ezt a mozaik n-gramok szisztematikus elemzésével képesek vagyunk azonosítani. Az eltérő reprezentációs szintek egyidejű alkalmazása továbbá abban segít minket, hogy míg a csak POS-tageken alapuló módszerrel a morfológiai és szófaji mintázatok egyszerűen kinyerhetők, addig a mozaik n-gram módszerrel az olyan szerkezeti absztrakciókhoz is eljuthatunk, amelyekben nem csupán a gyakran visszatérő morfológiai kidolgozások teremtik meg az összetartozást, hanem például bizonyos szótövek nagyobb mértékű asszociációja az adott konstrukciós mintázattal (Bajzát és Indig, 2023). A mintázatok sematikusságának skalaritása (Traugott, 2008; Brdar, 2020), tehát az, hogy milyen absztrakciós szinteken vagyunk képesek előhívni nyelvi mintázatokat, szintén megteremti az igényt a különböző sematikussági szintek egyenrangú kezelésére. Az 1. táblázat egy konstruált példán keresztül szemlélteti a mozaik n-gram módszerrel előállított kimenetet (a táblázat nem tartalmazza a példából előállítható összes mozaikot).

[3] A konstrukciójelöltek forma–funkció párokként való kezelése arra utal, hogy az így meghatározott sémák fonológiai és mentális reprezentációja egymástól elválaszthatatlan, emergensen létrejövő nyelvi egységet képeznek (Tolcsvai Nagy, 2017).

1. táblázat. Példa egy mozaik trigramra.

könyveket	Vettem	Tegnap
könyveket	lemma:vesz	[/N][_Tmp_Loc/Adv]
lemma:könyv	[/V][Pst.Def.1Sg]	[/N][_Tmp_Loc/Adv]
[/N][Pl][Acc]	Vettem	Tegnap

Az 1. táblázatban a *Könyveket vettem tegnap* elemi mondat (1. sor) mozaik n-gram módszerrel létrehozott lehetséges absztrakcióit láthatjuk (2–4. sor). Megfigyelhetjük, hogy a kimenetek megtartják a bemeneti példány szórendi elrendeződését, illetve három absztrakciós szintet alkalmazunk egyidejűleg (szóalak, lemma, POS-tag).

Mivel a funkcionális megközelítésben a konstrukció jelentése nem, vagy csak korlátozottan vezethető le a *komponens*ek jelentéséből, továbbá a vizsgált szekvenciákon belüli szórendi elrendeződés szintúgy szimbolikus relációként kezelhető, a komponensek sorrendi mintázatának megtartása indokolt, ahogyan az 1. táblázatban is megfigyelhettük. Ugyanazon komponensek különböző szórendi elrendeződése más-más közlési funkcióval asszociálódhat a feldolgozás során, illetve az aktív figyelemirányítás működtetése szintúgy kiaknázza a szórendi pozicionálás potenciálját a nyelvi konstruálásban (vö. Tolcsvai Nagy, 2017: 47; Tolcsvai Nagy és Kugler, 2017b: 451–457).

A példányalapú megközelítés módot alkalmazva fontos, hogy a sémák szerveződésére hatással van a *konstruálás* dinamikus természete, tehát újabb és újabb mintázatok jönnek létre a nyelvi változásban, valamint a szerkezetek újabb funkciókkal bővülnek. A konstrukciók a használatban példányalapon jönnek létre, nem pedig az absztrahált fölérendelt, makroszerkezeti sémákból levezetve (Hilpert, 2014: 65–67; Langacker, 2005: 102–103). Az elemi mondatként való megjelenítésre azért van szükségünk, mert a lexikon és a szintaxis nem határolható el élesen egymástól, így skalárisan, a *lexikon-szintaxis kontinuumon* kezeljük az adatok által felkínált lehetséges mintázatokat. Nem utolsó sorban pedig a mozaikokként történő reprezentálás nem kényszeríti a nyelvészt arra, hogy az adatok elemzése előtt döntést hozzon arról, hogy az általa vizsgált szerkezeti sémák intuitíven feltételezett keretkomponensei vagy a célkomponens előtérbe helyezésével kísérelje meg a korpuszból való adatkinyerést.

3 A mozaik n-gram módszer általános gyakorlati alkalmazásmódja a nyelvi mintázatazonosításban

A fejezet röviden összegzi a konstrukciójelöltek azonosításának mozaik n-gram alapú eljárását. Ezek a módszertani lépések a korpusztól függetlenül állandók, a 4. fejezetben ismertetjük a Történeti Magánéleti Korpusz nyelvi anyagának hatékony kezelésére bevezetett további eljárásokat.

A nyelvi mintázatazonosításban a mozaik n-gram gondolatának gyakorlatba ültetését elsőként a Magyar Nemzeti Szövegtár 2-ből (Oravecz és mtsai, 2014) és a Magyar Webkorpusz 2.0-ból (Nemeskey, 2020) kinyert konkordanciák feldolgozására tervezve hajtották végre (Indig és Bajzát, 2023). A szerzők a konkordanciákat egységesen újraelemeztették az e-magyar szoftverrel (Indig és mtsai, 2019), mivel egyfelől

szükségszerű volt az egységes lemmatizálás, másrészt pedig az emMorph (Novák és mtsai, 2016) elemzőmodul nyelvspecifikus és megbízható annotációval látja el a mai magyar nyelvi példányokat. A duplumok, fals pozitív találatok, illetve a hibás példányok szűrése automatizált módszerrel történik az eljárásban. A fals pozitív találatok kiválogatása két úton történik alapértelmezetten: először az automatikus szűréssel jól azonosítható, gyanús példányok törlésére kerül sor (egymás mellett több különleges karakter előfordulása, túl hosszú elemi mondatok, csak nagybetűt tartalmazó példák), a második szűrés során pedig az egyértelműen rossz POS-tag annotációval ellátott példányokat tudjuk eltávolítani. Lehetőség van továbbá a vizsgált szerkezeti mintázat elengedhetetlen komponensei közötti maximum távolsági ablak meghatározására is, azonban ez már a kutatási kérdésre specializált beállítás, nem pedig általános lépés (Indig és Bajzát, 2023).

A magyar nyelvre még nem érhető el olyan könnyen használható szoftver, ami dedikáltan és hatékonyan képes a tagmondatra-bontásra, ezért a szerzők igyekeztek a rendelkezésekre álló lehetőségek közül a legegyszerűbben hozzáférhető módszert felkínálni. A vizsgált szerkezetre vonatkozó valószínűsíthető sémák elrendeződése alapján (mik például a célkomponensek), illetve a tagmondatokra jellemző prototipikus tulajdonságok meghatározásával tudjuk ablakok mentén kinyerni a számunkra releváns elemi mondatokat az eszközzel.

Azonban korántsem elegendő a minták feljebb bemutatott előkészítése ahhoz, hogy a kimenetként kapott mozaikok nyelvészetileg, a konstrukciós összetartozás vizsgálatához megfelelően elemezhető formában álljanak elő. Az összes lehetséges mozaik legenerálása és kezelése egyrészt rendkívül erőforrás-igényes lehet, másrészt pedig létrejönnek duplum mozaikok is. Az e-magyar elemzőrendszer által előállított címkeszókincs nagyszámú típusjelölést tud létrehozni, amely radikálisan szétbontja a hasonló szerkezeti mintázatokat egymástól, amelyre a jelen kutatásban nincs szükség. Esetünkben célszerű a címkéket a formai és szemantikai összetartozások mentén csoportosítani, illetve, ha a morfológiai címke egyetlen szóalakot jelöl (például *-e* kérdőpartikula), akkor érdemes csak az egyik absztrakciós réteget megtartani, ezzel egyfelől csökkenthető a duplumok létrehozásának valószínűsége, másfelől pedig az adatok reprezentációja rendezettebbé válik, ami elősegíti a további elemzési munkákat (Indig és Bajzát, 2023: 2–3).

A kódban lehetőség van az adatok megjelenítésének további megváltoztatására is, például beállíthatunk küszöbértéket, hogy mekkora előfordulási érték felett jelenítődjenek meg a kimenetként kapott mozaikok, azonban ez a beállítás csupán az adatok megjelenítésére vonatkozik, a küszöbérték alatti mintázatok mögött álló példány más absztrakciós eloszlásban még bekerülhet más csoportba is (Indig és Bajzát, 2023).

Az algoritmus nem csupán létrehozza a mozaikokat, hanem azokat osztályozza is, valamint a kimenetben az absztraktabb, gyakori minták alá kerülnek a specifikusabb, alacsonyabb előfordulással azonosított sorozatok. Ez azokban az esetekben történik meg, amikor az absztrakció több elemet képes lefogni, mint a konkrétabb példányok. A konkrétabb példányok ekkor részhalmaz relációban vannak az általánosabb mintázattal. Azokból a mozaikokból, amelyek azonos gyakorisági értékkel rendelkeznek, és ugyanazon példányok alapján jönnek létre, csak a specifikus, konkrétabb mintázatot őrzi meg, ezzel elkerülhetővé téve a felesleges absztrakciókat és a duplum mozaikok keletkezését.

4 A módszer adaptálása a Történeti Magánéleti Korpusz nyelvi anyagára

A fejezet a Történeti Magánéleti Korpusz mozaik n-gram módszerrel történő feldolgozását mutatja be a finit igét tartalmazó elemi mondatok kinyerésén és az így előállított minta mozaik n-gram módszerrel való elemzésén keresztül. A vizsgált nyelvtörténeti időszak nyelvi anyagára (kései ó- és középmagyar kor) jellemzőek a nem finit igével megvalósított predikátumok előfordulásai is, tehát a határozói igenevek nem csupán határozói szerepben jelenhetnek meg a korpusz anyagában, hanem az igei állítmány funkcióját is betölthetik (Varga, 2019: 83–89; Horváth, 2003: 663–664). Mivel a célunk a finit igével előforduló tagmondatok azonosítása volt, a határozói igenévi predikátumok nem kerültek bele a vizsgálati mintába. Fontos megjegyezni, hogy mivel a morfológiai annotációból indultunk ki, tehát a forma volt a szűrések során a meghatározó, a melléknévi predikátumokat tartalmazó magok (*magmondat* vö. Imrényi, 2017: 673–677) szintén nem kerültek bele a mintába abban az esetben, ha egyes szám harmadik személyű, jelen idejű és kijelentő módú igei inflexiós toldalékkal fordultak elő.

4.1 A Történeti Magánéleti Korpusz

A Történeti Magánéleti Korpusz létrehozásának céljaként azt határozták meg, hogy az informális (a beszélt nyelvi regiszterhez feltehetően közelítő) kései ó- és középmagyar kori nyelvhasználati események írott szövegeinek kereshető adatbázisát elkészítsék, amelyben a szöveghű adatok, azok normalizált átiratai és a morfológiai elemzésük egyaránt elérhető (Dömötör és mtsai, 2017). A nyelvi adatok forrásai a célkitűzésnek megfelelően magánlevelek és bírósági jegyzőkönyvek. A korpusz anyagában nagyobb arányban jelennek meg a középmagyar kori szövegek, mivel az előbb említett szövegtípusok a 15. század végéről, illetve a 16. század első felétől datálhatók (Dömötör és mtsai, 2017: 88). A korpusz mérete körülbelül 1,5 millió tokenre tehető, az idegen nyelvű részek nélkül 1 044 893 darab szót tartalmaz. A korpusz magába foglalja a nyelvi anyagának metaadatait, azonban a kutatásunk jelenlegi korai szakaszában nem kezeltük ezeket. Fontos kiemelni, hogy a korpusz nyelvi elemzése és a normalizált szöveganyag kezelése megbízható, hiszen például a tagmondatra bontás manuálisan történt meg, grammatikai kategóriák figyelembevételével (állítmányi szerkezetek meghatározásával). A tagmondatok megbízható azonosíthatósága a fent bemutatott mozaik n-gram módszer számára kulcsfontosságú, ugyanúgy, ahogyan a különböző írásmódbeli variációk normalizálása is (vö. Dömötör és mtsai, 2017: 90–103). A normalizált szövegeket a Humor morfológiai elemző (Novák, 2003) ó- és középmagyar kori szövegekre adaptált változatával dolgozták fel, így a mai magyar nyelvben már nem megtalálható alaktani jelenségeket is hatékonyan kezelték, emellett kézi és automatikus egyértelműsítést is alkalmaztak (Dömötör és mtsai, 2017: 96–103), így elmondható, hogy a morfológiai annotáció szintén megbízhatónak bizonyul a mozaik n-gram módszer alkalmazásához.

4.2 A mozaik n-gram módszer adaptálása

A fent bemutatott eljárásmód alkalmazását és finomhangolását mutatjuk be a jelen fejezetben. A Történeti Magánéleti Korpusz nyelvi anyagának táblázatos formája képezte

a kiindulópontot a feldolgozáskor. A táblázatok rendre a következő oszlopokat tartalmazták: token ID mondatszinten, a szóalak eredeti betűhű formája, a szövegszó normalizált formája, a szövegszó normalizált formája punktuáció nélkül, a szövegszó lemmája normalizált formában és a szövegszó morfológiai elemzése (POS-tag). Az oszlopokból a mozaikok előállításához hármat őriztünk meg: a normalizált változatot punktuációval, a lemmákat tartalmazó oszlopot, illetve a morfológiai elemzést. Mivel a központozás, illetve az *-e* kérdőpartikula nem önálló elemként jelentek meg külön sorokba tördelve, ezért tokenizálni kellett az adatokat. A módszer elmélete szerint az így előállított bemeneten már képesek lettünk volna létrehozni a mozaik n-gramokat, azonban egyfelől az így előállított kimenetek formailag nem lettek volna összhangban más korpuszból kinyert minták feldolgozott eredményeivel, mivel eltérő annotációs sémában jelennek meg a POS-tagek. Másrészt valószínűsíthető volt az MNSZ2-ből és a Webkorpusz 2.0-ból származó adatok feldolgozásából következtetve, hogy az egyedi címkék magas számú egyszeres, vagy kétszeres előfordulása nem tenné lehetővé a funkcionálisan összetartozó példányokat jelölő mozaikok hatékony kinyerését.

Összesen 2384 darab különböző címke meglétét azonosítottuk a korpuszban, miután kinyertük a címkeszókincset a táblázatokban. Ebből 701 darab típus összesen csak egyszer jelenik meg a vizsgált nyelvi anyagban. Kézenfekvő megoldásnak tűnt, hogy a mai magyar nyelvi anyagra kialakított relációkat és címkeegyszerűsítéseket alkalmazzuk a jelen feladat megoldására is, azonban a morfológiai annotációs sémák különböztek. Emiatt manuálisan átalakítottuk a TMK által használt címkekészletet az emMorph címkéinek mintájára. A fordítási folyamat az esetek jelentős részében valóban fordítást jelentett, tehát megkerestük, hogy az adott morfológiai címkét a másik jelölési rendszerben minek tudjuk megfeleltetni (pl. „V.Past.S3” (*ment*) (TMK) -> „[/V][Pst.NDef.3Sg]”), azonban ez nem minden esetben volt egyértelmű. Kezelnünk kellett azokat a címkéket is, amelyek ugyan léteznek az emMorphban, viszont más a kategorizációs hatókörük. Például az *is* lexéma jelölésére az emMorph egységesen „[/Adv]” címkét alkalmaz (hasonlóan például a tagadószó kategorizálásához), a TMK pedig önálló taget tart fenn számára („Clit_is”). A probléma feloldása viszonylag egyszerű abban a kontextusban, hogy az „[/Adv]” címkét, annak a túláltalánosító tulajdonságából fakadóan, az eredeti eljárás törli, s csak az adott példány szóalakját tartja meg a mozaik n-gram generálás során, tehát ha besoroljuk az „[/Adv]” címke alá a „Clit_is” taget, az nem befolyásolná az eredményeket. Ennél bonyolultabb például a névelők helyzete. A TMK-n alkalmazott morfológiai annotáció nem tesz különbséget a határozott és határozatlan névelők között, mindegyik példány ugyanazt a címkét kapja meg („Det”), viszont az emMorph elvégzi a határozottság szerinti differenciálást („[/Det|Art.Def]” és „[/Det|Art.NDef]”). Mivel a Humor morfológiai elemző címkéje nagyobb halmazt fed le, ezért nem tudtuk fordítani a címkét, így a névelőkre a „[/Det]” jelölést alkalmaztuk a későbbiekben, ezzel explicitté téve, hogy potenciálisan lehet határozott, vagy határozatlan az adott példány. A harmadik probléma a ma archaikusnak számító morfológiai jelenségek kezelése volt. Abban az esetben, ha valamilyen alakvariánsról volt szó, de a funkciója megfeleltethető a mai alaknak, egyszerűen az adott jelentést kifejező címkével helyettesítettük (pl. „V.Cond.P1=nÓk.Def” (*akarnók*) -> „[/V][Cond.Def.1Pl]”). Ha nem tudtunk megfelelő címkét találni az emMorph listájában (például az elbeszélői múlt idő esetén (*érkezék*)), akkor a jelölési mintázat alapján új címkét hoztunk létre (pl. „[/V][Ipf.NDef.3Sg]”). A címkeszókincs egyenként való ellenőrzésével még tovább tudtuk tisztítani a korpuszunkat a mozaik n-gramok

generálása előtt. Fontos volt a nem egyértelmű annotációval rendelkező példányok, töredékes adatok és a latin nyelvű szekvenciák POS-tag alapján történő eltávolítása. Ez természetesen adatveszteséggel is járhat, azonban az elsődleges szempont a minél jobb minőségű bemenet előállítása volt. A címkék fordításánál minden esetben kértünk le példányokat a korpuszból, ezzel ellenőrizve azt, hogy az adott jelölés pontosan milyen típusú példányokat fed le. Az átírásokat a későbbi visszaellenőrizhetőség miatt táblázatkezelőben dokumentáltuk. Az alábbi oszlopokba rendeztük a szabályokat: 1. a címke jelentése; 2. példa a korpuszból; 3. megfeleltethetőség esetén az emMorphból származó tag; 4. a további feldolgozás során végrehajtott művelet; 5. a javasolt átírás.

Az újonnan keletkező címkék és a többféle típus előfordulása miatt az eredeti eljárásban kialakított címkeegyszerűsítő szabályrendszert is ki kellett egészítenünk, hogy koherens maradjon a módszer. Összesen 210 új relációt vezettünk be. A duplumok és a vélhetően hibás példányok törlése után a részkorpuszban 1 088 129 token maradt.

Mivel a jelen vizsgálatban a finit igealakot tartalmazó elemi mondatokból törekedtünk a mozaik n-gramok létrehozására, a következő lépést a megfelelő módon történő tagmondat-kinyerés jelentette. A normalizált szövegrétegből vettük ki az első absztrakciós szintet, így támaszkodhattunk a tagmondatok között megjelenő szisztematikusan kiosztott központozásra. Az elemi mondatokat úgy formalizáltuk, hogy minden esetben a finit igei példányokra illeszkedő morfológiai címkéket és környezetüket azonosítottuk három pozícióban: (1.) a mondat elején az első finit igét követő első írásjelig, (2) a két írásjel között elhelyezkedő eseteknél a baloldali írásjeltől az első finit igét követő írásjelig, (3) a mondat végén pedig az utolsó finit igét megelőző írásjeltől a mondatzáró írásjelig. Ennek a módszernek az alkalmazhatósága limitált, hiszen csak normalizált szövegek feldolgozásában működtethető megbízhatóan, s még ebben az esetben is korlátozhatja bizonyos kutatási kérdések megválaszolását. Például, ha egy vizsgálat szempontjából releváns a vesszővel elválasztott mellérendelői mondatrészek (pl. *Vettem almát, barackot és banánt*) kiterjedésének mérése, nem alkalmazható.

A mozaik n-gramok megjelenítésének küszöbértékét két példányban határoztuk meg. A megjelenítésre vonatkozó döntésünk oka, hogy alapvetően a korpusz mérete nem tekinthető nagynak (szemben mondjuk a Magyar Webkorpusz 2.0 csaknem 10 milliárd szavas terjedelmével), továbbá nem alkalmazunk fölérendelt finit ige címkét a mozaikok létrehozásánál, így a program az igei morfológiai mintázatok mentén, vagy a konkrét igei lemmák alapján osztályoz.

4.3 A kapott mozaikok megjelenítése

A mozaik n-gramokat összesen 122 536 elemi mondatból hoztuk létre (az előző alfejezetben bemutatott szűrések után fennmaradó mennyiség). A kettő, három, négy, öt, hat, hét, nyolc és kilenc tokenből álló elemi mondatok mozaikjait generáltuk le. A POS-tagek alapján elvégzett törlés 103, míg a duplumszűrés 15 365 elemi mondatot távolított el. A duplumszűrés alatt azt értjük, hogy a szó szerint egyező példányokból csak egyet tartottunk meg. A 2. táblázat a kimenetben látható osztályozást és megjelenítést szemlélteti.

2. táblázat. Példa mozaik 3-gramokra a TMK-ból.

46	[/V][Prs.NDef.3S]	[/Det]	[/N][Nom]	
	4	meggyógyul	[/Det]	[/N][Nom]
	4	lemma:meg|+van	[/Det]	[/N][Nom]
	4	Felel	[/Det]	[/N][Nom]
	3	Mond	[/Det]	[/N][Nom]
	2	[/V][Prs.NDef.3Sg][4]	[/Det]	lábad[5]
	2	[/V][Prs.NDef.3Sg]	[/Det]	Isten
	[…]	[…]	[…]	[…]

A 2. táblázat legfelső sora mutatja be az absztrakció legmagasabb szintjén azonosított mozaik n-gram mintázatát az egyes szám harmadik személyű, jelen idejű, kijelentő módú, határozatlan ragozású igealaknak, a névelőnek és az alanyesetű főnévnek Az első sorban szereplő gyakori mozaik n-gram által lefedett konkrét n-grammok további, ritkább absztrakciói, melyek az első sor által fedett konkrét n-grammok csak egy részhalmazát fedik le, másodrendű absztrakciót jelentenek az első sorban szerepeltetett mozaik n-gramhoz képest. A másodrendű absztrakciók elhagyásával ritkítható a számításba veendő mozaikok száma. A megjelenítési beállítások miatt csak azokat a trigramokat látjuk, amelyek legalább kétszer előfordultak a vizsgált mintában. Fontos kiemelni, hogy csak azoknak a soroknak a gyakorisági értékeit adhatjuk össze egymással, amelyek minden pozícióban ugyanazon absztrakciós megjelenítést alkalmazzák.

5 Eredmények

A fejezet az első eredmények áttekintő bemutatását teszi meg. A 2-, 3-, 4-, és 5-gram első tíz leggyakoribb mozaik n-gram mintázatát szemlélteti, összekapcsolva a nyelvészeti elemzéssel. A tíz leggyakoribb konstrukciójelölt közül a legmagasabb absztrakciós szinteket figyelhetjük meg a fejezet táblázataiban (2–5. táblázat). A táblázatok első oszlopa a gyakorisági listán elfoglalt pozíciót mutatja be, a második a mozaik előfordulási számát, míg az ezt követő oszlopokban a sorozat elemei láthatók.

[4] Mivel az e-magyar a mai magyar nyelvi adatok elemzésénél a kimenetben az igekötős példányokat a nem igekötős igék mintájára címkézte, így a TMK-ból vett mintát is ennek megfelelően kezeltük. A későbbiekben ezt az eljárást érdemes felülvizsgálni, azért, hogy a módszer szélesebb körben alkalmazhatóvá váljon a grammatikai, grammatikatörténeti vizsgálatokban.

[5] Azért, hogy a birtokviszonyok különböző grammatikai kidolgozásai ne sokszorozzák meg a csoportok számát, illetve ne aprózódjanak fel a mozaikok, a relációs viszonyok kialakításánál az [/N][Nom] alá rendeltük a birtokjellel ellátott példányokat is.

3. táblázat. A tíz leggyakoribb finit igés mozaik 2-gram.

Hely	db	1. elem	2. elem
1.	279	[/Cnj]	[/V][Pst.NDef.3Sg]
2.	270	[/Cnj]	[/V][Pst.Def.3Sg]
3.	241	[/Cnj]	[/V][Prs.NDef.3Sg]
4.	153	[/Cnj]	[/V][Prs.Def.3Sg]
5.	144	[/Cnj]	[/V][Prs.Def.1Sg]
6.	97	[/Cnj]	[/V][Prs.NDef.3Pl]
7.	94	[/Adv|Pro|Rel]	[/V][Pst.NDef.3Sg]
8.	86	[/V][Sbjv.Def.2Sg]	[/Prev]
9.	85	[/Cnj]	[/V][Sbjv.NDef.3Sg]
10.	83	[/Adv|Pro]	[/V][Pst.NDef.3Sg]

A 3. táblázatban megfigyelhetővé válnak a korpuszból azoknak a két szó hosszúságú elemi mondatoknak a leggyakoribb mozaikjai, amelyek valamilyen finit igés magot tartalmaznak. Az első négy mozaik igei komponensén egyes szám harmadik személyű *inflexiós toldalék*ot azonosítottunk. Az E/3. grammatikai jelöltség alapbeállításként kezelhető, mivel a megnyilatkozó (konstruáló szubjektum) a saját kiindulópontjából referál egy entitásra, amelyet az elemi mondat *elsődleges figurá*jaként (eltérő nyelvi kidolgozottságban) helyez a figyelem előterébe. Az E/3. bármilyen entitásra vonatkozhat, míg az E/1. és az E/2. kizárólag a *diskurzusszituáció*ban azonosított mindenkori beszélőre és hallgatóra referálhat (Tolcsvai Nagy és Kugler, 2017: 432). Ugyan általánosságban elmondható, hogy az E/3. személyjelölést valószínűsíthetjük a korpuszokban leggyakoribbként előforduló kidolgozásnak, azonban bizonyos konstrukciójelöltek esetében tapasztalhatunk elmozdulást az *alapbeállítás*tól (Bajzát és Indig, 2023). A grammatikai igeidő *deiktikusan* az ige által előhívott folyamatot a diskurzusszituáció időviszonyaihoz kapcsolja (Tolcsvai Nagy és Kugler, 2017b: 350). Alapbeállításként a grammatikai jelen idejű kidolgozást tartjuk, mivel a megértett beszédidővel egyidejű események kifejezését teszi lehetővé, s a formai oldalon ez jelöletlenséggel jár együtt.

A 3. táblázatban azonban megfigyelhető az ettől az alapbeállítástól való elmozdulás a grammatikai múlt idő felé. Ez valószínűsíthetően a korpuszt alkotó nyelvi anyag műfaj- és szövegtípusaira jellemző narratív keretekből következhet. A 8. leggyakoribb mozaik a tipikus felszólító mondatok sémáját valósítja meg azon keresztül, hogy a beszédpartnerre referál a felszólító módú igei inflexió, és az ige és igekötő kidolgozási viszonyában inverz szórendi mintázatot mutat az elemi mondat. A 9. sorban kötőmódú kidolgozásokat figyelhetünk meg a kötőszó és kötőmódú (formailag felszólító) E/3. személyjelölésű finit igealak együttállásában. A leggyakoribb kételemű mozaikok első pozíciójában jellemzően valamilyen kötőszót, határozói vonatkozói névmást, vagy határozóként funkcionáló egyéb névmási elemet azonosítottunk. Az utóbbi két kategória szintén tagmondatviszonyok közötti reláció nyelvi kidolgozását hajtja végre.

4. táblázat. A tíz leggyakoribb finit igés mozaik 3-gram.

Hely	db	1. elem	2. elem	3. elem
1.	206	[/N][Nom]	[/N][Nom]	[/V][Prs.Def.3Sg]
2.	186	[/N][Nom]	[/N][Nom]	lemma:vall
3.	148	[/Cnj]	[/N][Nom]	[/V][Prs.NDef.3Sg]
4.	118	[/Cnj]	[/N][Nom]	lemma:van
5.	85	[/Cnj]	[/Adj][Nom]	lemma:van
6.	77	[/Cnj]	[/Adv|Pro]	[/V][Pst.NDef.3Sg]
7.	76	[/Cnj]	[/N][Acc]	[/V][Pst.Def.3Sg]
8.	73	[/Cnj]	[/N][Nom]	[/V][Prs.Def.3Sg]
9.	71	[/Cnj]	[/Adv|Pro]	[/V][Prs.NDef.3Sg]
10.	69	[/Det]	[/N][Nom]	[/V][Pst.NDef.3Sg]

A 4. táblázat a tíz leggyakoribb háromelemű mozaik n-gramot mutatja be a vizsgált mintából. Az első és a második helyen lévő mozaik felhívja a figyelmünket arra, hogy a későbbiekben érdemes a névelemek automatikus azonosítását hozzákapcsolni a módszerhez (pl. a HuSpaCy alkalmazásával (Orosz és mtsai, 2023)), hiszen a jelenlegi eljárás a vezeték- és keresztnév komponenseket külön elemként azonosítja, holott az elemi mondat egy entitásra referál (pl. „[/N][Nom] Simonné vallja”). Mindemellett a második mozaik 3-gram bemutatja, hogy a korpuszra jellemzőek a *kontextualizáló főmondat*ok (Kugler, 2019), amelyek a prominens mellékmondat beszélői szándék szerinti feldolgozását segítik elő, ezen funkciók közül is a *személyhez kötés* részletezőbb kidolgozását valósítják meg (Imrényi, 2017: 747–748). Ez tipikusan a bírósági jegyzőkönyvek által alkalmazott konstruálási módokkal állhat összefüggésben, ahol az egyes eseményekről tudósító példányok explicit kifejtést adnak a *referenciális központ*ról és a *tudatosság szubjektum*áról (vö. Tátrai, 2017: 940–942; Sanders és Spooren, 1997) tehát arról, hogy a közölt információ kitől származik.

A 4. és 5. sorban a *van* lemmát tartalmazó mozaikok figyelhetők meg. A második elem szófaji annotációjából következtethetünk arra, hogy a 4. sor a *van* létigei funkcióját vagy azonosító állítmányi magban betöltött szerepét mutathatja (Imrényi, 2011: 129–134). Az 5. sor pedig a minősítő komplex magmondatokat (Imrényi, 2011: 119–124) tartalmazó 3-gramokat fedi fel, tehát képesek vagyunk meghatározni a *van* lemma különböző specifikus magmondatbeli előfordulásait.

5. táblázat. A tíz leggyakoribb finit igés mozaik 4-gram.

Hely	db	1. elem	2. elem	3. elem	4. elem
1.	65	[/Cnj]	[/Det]	[/N][Nom]	[/V][Pst.NDef.3Sg]
2.	62	[/Cnj]	[/Adj][Nom]	[/N][Nom]	lemma:van
3.	54	[/N][Nom]	[/N][Nom]	[/N][Acc]	[/V][Prs.Def.3Sg]
4.	50	[/Cnj]	[/Det]	[/N][Acc]	[/V][Pst.Def.3Sg]
5.	47	[/Cnj]	[/V][Prs.NDef.3Sg]	[/Det]	[/N][Nom]
6.	42	[/Cnj]	[/V][Pst.NDef.3Sg]	[/Det]	[/N][Nom]
7.	39	[/N][Nom]	[/N][Nom]	[/N][Acc]	lemma:vall
8.	38	[/Cnj]	[/V][Pst.NDef.3Sg]	[/Det]	[/N][Acc]
9.	37	[/Cnj]	[/Det]	[/N][Nom]	[/V][Prs.NDef.3Sg]
10.	36	[/N|Pro|Rel][Subl]	lemma:felel	[/N][Nom]	[/N][Nom]

Az 5. táblázat a tíz leggyakoribb mozaik 4-gramot jeleníti meg az elemzett mintából. Látható az első pozícióban az, hogy a mozaik által lefedett példányok összetett mondatba illeszkednek. A mozaik specifikusabb mintázataiban mellérendelői, illetve *hogy*-os tartalomkifejtő alárendelői mellékmondatok sémáját tudtuk azonosítani, a leggyakoribb mellérendelői kötőszó az *és* volt (11 előfordulással). Továbbá látjuk a szórendi mintázatból azt, hogy az alapbeállításnak, tehát a semleges pozitív kijelentő mondatnak megfelelő sémát aktiválja. Azonban az igeidő jelölés eltér a kijelentő módtól a múlt idő irányába hasonlóan, ahogyan a 2. táblázatban megfigyeltük. Az 5. táblázatban a személyjelölések tekintetében nem fedezhetünk fel változatosságot, úgy ahogyan a 4. táblázatban sem fordult elő más személyjelölési kidolgozás.

A második sorban a *van* lemmát tartalmazó elemi mondatok absztrakcióját látjuk. A szórendi elrendeződésből és a szófaji jelölésből következtethetünk arra, hogy ez a csoport vegyesen tartalmaz létigei funkciót betöltő *van* lemmájú kidolgozásokat („[/Cnj] [/Adj][Nom] [/N][Nom] van”), illetve minősítő összetett magmondatokat is („hogy bűbájos [/N][Nom] legyen”). A harmadik sor esetén ugyanazt a problémát tapasztaljuk, amelyet a 4. táblázat 1. és 2. mozaikjának értelmezésénél láttunk, tehát azt, hogy valójában egy entitásról beszélhetünk az alanyi pozícióban, de a kereszt- és vezetéknév külön történő azonosítása miatt a 4-gramok közé kerül a szerkezet. A negyedik sorban a kötőszóval jelölt tagmondatok közötti kapcsolat a mozaikok több mint felében kapcsolatos volt (*és* és *s* kötőszó), ha a specifikusabb mintákat is megnézzük. Ezen kívül szintén az alapbeállításnak megfelelő szórendi és prozódiai (erre korlátozottan következtethetünk) elrendeződést látunk múlt idejű kidolgozásban.

Az eddig megfigyeltetett összes mozaik n-gram utolsó pozíciójában azonosítottuk a finit igealakot, viszont az 5., 6., 8. és 10. mozaik esetében ettől eltérő szórendi konstruálási módot látunk. A kötőszót, vagy az alárendelői mellékmondatot bevezető vonatkozói névmást követi a finit igealak, amely valószínűsíthetően a konstruálás során működő figyelemirányítási szándékkal áll összefüggésben, tehát az igével kifejtett konkrét folyamat kerül a figyelem előterébe a feldolgozás során, nem pedig az elemi jelenet valamely szereplője.

6. táblázat. A tíz leggyakoribb finit igés mozaik 5-gram.

Hely	db	1. elem	2. elem	3. elem	4. elem	5. elem
1.	27	[/Cnj]	[/V][Sbjv.Def.3Sg]	[/Prev]	[/Det]	[/N][Acc]
2.	23	[/Cnj]	[/Det]	[/N][Nom]	[/N][Acc]	[/V][Pst.Def.3Sg]
3.	21	[/N|Pro|Rel][Nom]	[/Det]	[/N][Nom]	[/N][Acc]	[/V][Prs.Def.3Sg]
4.	21	[/N|Pro|Rel][Nom]	[/Det]	[/N][Nom]	[/N][Acc]	lemma:illet
5.	16	[/Cnj]	[/Det][Pro|Nom]	[/N][Nom]	nem	[/V][Pst.NDef.3Sg]
6.	16	[/Cnj]	[/Det][Pro|Nom]	[/N][Nom]	nem	lemma:van
7.	15	[/N][Nom]	[/N][Nom]	vallja	hite	után
8.	15	[/N][Acc]	is	lemma:hall	[/N][Nom]	[/N][Abl]
9.	14	[/Cnj]	[/Det]	[/N][Nom]	[/N][Nom]	[/V][Pst.NDef.3Sg]
10.	13	[/N][Acc]	is	[/V][Prs.Def.3Sg]	[/Det]	[/N][Nom]

A 6. táblázat a leggyakoribb mozaik 5-gramokat mutatja be a vizsgált részkorpuszból. Látható, hogy az egyes mintázatok gyakorisági értékei lecsökkentek. Egyrészt ennek oka lehet az, hogy a morfológiai annotációs réteg a komplex igei toldalékkal látja el a mintázatokat (grammatikai idő-, mód-, személy- és határozottság jelölése). Ezért sok, egymástól elkülönülő csoport jön létre. Továbbá a mozaik n-gram módszer érzékeny a

szórendre, amely egy öt elemű sorozatnál már jobban kimutatható hatással van a csoportok létrehozására.

Az alapbeállításnak megfelelő semleges pozitív kijelentő mondat közlési funkciójától eltérő csoportok váltak azonosíthatóvá a mozaik 5-gram gyakori csoportjaiban. Ilyenek például a *nem* felülíró kifejezéssel (vö. Imrényi, 2017: 726–730) kívülről módosított magmondatot tartalmazó szerkezetek, illetve az *is* kiterjesztő (vö. Imrényi, 2017: 720–725) kifejezéssel együtt megjelenő környezetek. Szintén feltűnően elkülönbözik az alapbeállítástól az első sorban látható, összetett mondat részeként feldolgozott E/3. személyjelöléssel megjelenő felszólító módú konstrukció, mivel nem vártuk, hogy formailag és funkcionálisan jelöltebb mintázat válik a csoport leggyakoribb mozaikjává.

A szám- és személybeli kidolgozást illetően a mozaik 4-gram gyakori eseteihez hasonlóan az öthosszúságú sorozatok sem mutattak változatosságot. Az E/3. gyakori megjelenése ugyan mindegyik hosszúságú mozaik esetén várható volt, azonban a korpusz anyagát alkotó műfaj- és szövegtípusok (magánlevelezés és peres eljárások jegyzőkönyvei) jellemzőiből fakadóan számítottunk a konstruáló szubjektumra vagy a beszédpartnerre referáló szerkezetek prominens előfordulásaira is. A jegyzőkönyveknek tulajdonított nyelvi sémái hozzáférhetővé váltak a gyakori mozaik n-gram mintákban, azonban a magánlevek esetében várható sémák (a beszédpartner és a megnyilatkozó nyelvi kidolgozása) kevésbé jelentek meg a vizsgált gyakorisági régióban.

A 6. fejezetben rögzítettük az első tíz leggyakoribb mozaik n-gram (2-, 3-, 4-, és 5-hosszúságban) feltűnő nyelvi mintázatainak tulajdonságait és az így látható tendenciákat. Előtérbe helyeztük a szórendi sémáknak, az inflexiós toldalékok típusainak, illetve a sematikusan megfigyelhető, tipikus mondatfunkcióknak a bemutatását.

6 Összefoglalás és kitekintés

A tanulmány a mozaik n-gram módszer adaptációját mutatta be a Történeti Magánéleti Korpusz finit igés mintázatainak azonosításán keresztül. Az eljárás a nyelvi mintázatazonosításban betöltött szerepével hozzájárulhat ahhoz, hogy előzetes nyelvi intuíciók nélkül nyerjünk ki konstrukciójelölteket nagy nyelvi mintákból, és a nyelvész számára elemezhetővé tegyük őket, valamint a TMK-ra nincs olyan függőségi elemző, amely jól teljesítene, a mozaikokkal viszont a szintaktikai szerkezeteket is tudjuk vizsgálni. A folyamatban lévő kutatás további célkitűzése, hogy a módszer alkalmassá váljon más korpuszok nyelvi anyagának elemzésére (például Ómagyar korpusz (Simon és Sass, 2012), Magyar Történeti Szövegtár (Csengery, 2006)). A módszer korlátozottsága abban rejlik, hogy az eljárás kiszolgáltatott a bemeneti minták előzetes feldolgozásának, illetve a szövegtípusok diszkurzív mintázatainak.

A későbbiekben a bemeneti mintákon érdemes a mozaik szózsákok létrehozását is tesztelni, hogy a sémákkal asszociálódó komponensek szórendtől függetlenül is csoportosíthatóvá váljanak, ezzel általánosabb képet kapva a komponensek mintázatairól.

Bibliográfia

Bajzát T. B., Indig B.: Személyjelölési mintázatok feltárása mozaik n-gram alapon három segédige + főnévi igenév típus konstrukcióiban. In: Tolcsvai Nagy G., Tátrai Sz., Laczkó K. (szerk.) A magyar nyelv igei konstrukciói. Használatalapú konstrukciós nyelvtani kutatás. (Megjelenés előtt) (2023)

Brdar-Szabó R.: Rezultatív mikrokonstrukciók és morfológiai vetélytársaik elmélet és empíria egységében. In: Simon G., Tolcsvai Nagy G., (szerk.) Nyelvtan, diskurzus, megismerés. pp. 141–166. ELTE Eötvös Kiadó. Budapest (2020)

Bybee, J. L.: Language, Usage and Cognition. Cambridge University Press. Cambridge (2010)

Croft, W.: Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press. Oxford (2001)

Csengery K. Az elektronikus korpusz. In: Ittzés N. (szerk.) A magyar nyelv nagyszótára I. Segédletek. MTA Nyelvtudományi Intézet, Budapest (2006)

Dér Cs.: Grammatikalizáció. Nyelvtudományi Értekezések 158. Akadémiai Kiadó, Budapest (2008)

Dömötör A., Gugán K., Novák A., Varga M.: Kiútkeresés a morfológiai labirintusból – korpuszépítés ó- és középmagyar kori magánéleti szövegekből. Nyelvtudományi Közlemények 113. pp. 85–110 (2017)

Diessel, H.: Usage-based construction grammar. In: Dabrowska, E., Dagmar, D. (szerk) Handbook of Cognitive Linguistics. pp. 295–321. Mouton de Gruyger, Berlin (2015)

Glynn, D., Robinson, J. A. (szerk) Corpus methods for semantics: Quantitative studies in polysemy and synonymy. John Benjamins, Amsterdam, Philadelphia (2014)

Goldberg, A. E.: Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago (1995)

Gries, S. T., Stefanowitsch, A. (szerk) Corpora in Cognitive Linguistics. Mouton de Gruyter, Berlin (2007)

Hilpert, M.: Construction Grammar and its Application to English. Edinburgh University Press, Edinburgh (2014)

Hilpert M., Flach, S. (szerk.) Broadening the Spectrum of Corpus Linguistics: New perspectives on variability and change. John Benjamins, Amsterdam (2022)

Hilpert M., Correia Saavedra, D.: Why are grammatical elements more evenly dispersed than lexical elements? Assessing the roles of frequency and semantic generality. Corpora 12/3, pp. 369–392. (2017)

Horváth L.: Az állítmány. In: Kiss J., Pusztai F. (szerk.) Magyar nyelvtörténet. pp. 664–665. Osiris, Budapest (2003)

Hunston S., Gill F.: A corpus-driven approach to the lexical grammar of English. Studies in Corpus Linguistics 4. John Benjamins, Amsterdam (2000)

Imrényi A.: A magyar mondat viszonyhálózati modellje. Doktori (PhD) értekezés (2011)

Imrényi A.: Az elemi mondat viszonyhálózata. In: Tolcsvai Nagy G. (szerk.) Nyelvtan. A magyar nyelv kézikönyvtára 4. pp. 663–761. Osiris. Budapest (2017)

Indig B., Laki L., Prószéky G.: Mozaik nyelvmodell az ANAGRAMMA elemzőhöz. In: Tanács A., Varga V., Vincze V. (szerk.) XII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 260–272. Szegedi Tudományegyetem, TTIK, Informatikai Intézet, Szeged (2016)

Indig B., Sass B., Simon E., Mittelholcz I., Vadász N., Makrai M.: One format to rule them all – the emtsv pipeline for Hungarian. In: Friedrich, A., Zeyrek, D., Hoek, J. (szerk.) Proceedings of the 13th Linguistic Annotation Workshop. pp. 156–166. Association for Computational Linguistics, Florence (2019)

Indig B., Bajzát T. B. Bags and Mosaics: Semi-automatic Identification of Auxiliary Verbal Constructions for Agglutinative Languages. In: Vetulani, Z., Paroubek, P. (szerk.) Human Language Technologies as a Challenge for Computer Science and Linguistics. Adam Mickiewicz University, Poznań (2023)

Kalivoda, Á.: Igekötős szerkezetek a magyarban. Doktori (PhD) értekezés (2021)
Kugler N.: Kontextualizálás elemi mondattal. In: Laczkó K., Tátrai Sz. (szerk.) Kontextualizáció és metapragmatikai tudatosság. pp. 47–67. ELTE Eötvös Kiadó, Budapest (2019)
Langacker, R. W.: Foundations of Cognitive Grammar. Volume I Theoretical Prerequisites. Stanford University Press, Stanford, California (1987)
Langacker, Ronald W.: Construction Grammars: cognitive, radical, and less so. In: Ruiz de Mendoza Ibáñez, F. J., Peña-Cervel, M. S. (szerk.): Cognitive Linguistics Research 32. pp. 101–159. Mouton de Gruyter, Berlin, New York (2005)
Langacker, R. W. Cognitive Grammar: A basic introduction. Oxford University Press, Oxford (2008)
Langacker, R. W. Cognitive (Construction) Grammar. Cognitive Linguistics (20):1. pp. 167 – 176. (2009)
Luodonpää-Manni M., Penttilä, E., Viimaranta, J. (szerk) Empirical Approaches to Cognitive Linguistics. Cambridge Scholars Publishing, Cambridge (2017)
Narrog, H., Heine, B.: Grammaticalization. Oxford University Press, Oxford (2021)
Nemeskey D. M.: Natural Language Processing methods for Language Modeling. PhD thesis. (2020)
Novák, A.: Milyen a Jó Humor? In: Alexin Z., Csendes D. (szerk.) I. Magyar Számítógépes Nyelvészeti Konferencia, pp. 138–144. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2003)
Novák A., Gugán K., Varga M., Dömötör A.: Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. Language Resources and Evaluation 52, pp. 1–28. (2018)
Novák A., Siklósi B., Oravecz Cs.: A new integrated open-source morphological analyzer for Hungarian. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris (2016)
Oravecz Cs., Váradi, T., Sass B. The Hungarian Gigaword Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.): *Proceedings of the ninth international conference on language resources and evaluation (LREC-2014)*. pp. 1719–1723. European Languages Resources Association (ELRA), Reykjavik (2014)
Orosz Gy., Szabó G., Berkecz P., Szántó Z., Farkas R.: Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In: Ekštein, K., Pártl, F., Konopík, M. (szerk.) Text, Speech, and Dialogue. pp. 58–69. Springer (2023)
Sanders, J., Spooren, W.: Perspective, subjectivity, and modality from a cognitive point of view. In: Liebert, W., Redeker, G., Waugh, L. (szerk.): Discourse and perspective in cognitive linguistics. pp. 85–112. John Benjamins, Amsterdam, Philadelphia (1997)
Simon E., Sass B., Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. In: Általános Nyelvészeti Tanulmányok (XXIV). pp. 243–264 (2012)
Simon G.: Az igei jelentés metaforizációjának mintázatai. Nyelvtan- és korpuszvezérelt esettanulmányok. Jelentés és nyelvhasználat 5, pp. 1–36. (2018)
Tátrai Sz.: Pragmatika. In: Tolcsvai Nagy G. (szerk.) Nyelvtan. A magyar nyelv kézikönyvtára 4. pp. 899–1057. Osiris, Budapest (2017)
Tolcsvai Nagy G.: Bevezetés. In: Tolcsvai Nagy G. (szerk.) Nyelvtan. A magyar nyelv kézikönyvtára 4. pp. 23–71. Osiris, Budapest (2017)
Tolcsvai Nagy G., Kugler N.: Jelentéstan. In: Tolcsvai Nagy G. (szerk.) Nyelvtan. A magyar nyelv kézikönyvtára 4. pp. 207–499. Osiris, Budapest (2017)
Tolcsvai Nagy G. (szerk.) Nyelvtan. A magyar nyelv kézikönyvtára 4. Osiris, Budapest (2017)

Traugott, E. C.: ‘All that he endeavoured to prove was ...’: on the emergence of grammatical constructions in dialogual contexts. In: Kempson, R., Cooper, R. (szerk.) Language change and evolution. pp. 143–177. Kings College Publications, London (2008)
Varga M.: Középmagyar kori világi szövegek nyelvtörténeti vizsgálata, különös tekintettel a szövegtani és pragmatikai sajátosságokra. Doktori (PhD) értekezés (2019)