Véges erőforrás végtelen sok igekötős igére

Kalivoda Ágnes

MTA Nyelvtudományi Intézet
MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport
Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
kalivoda.agnes@nytud.mta.hu

Kivonat A PrevLex egy szabadon elérhető, manuálisan ellenőrzött erőforrás, amely 54 955 igekötős igét tartalmaz gyakorisági adatokkal együtt. Bár lefedi az MNSZ 2.0.4 korpusz összes igekötős finit igéjét, soha nem lehet teljes: bizonyos igekötők rendkívül produktívak, és tetszőleges számú új szó képezhető velük. A cikk központi témája az, hogyan mérhető az igekötőknek ez a tulajdonsága, és hogyan használhatók fel a kvantitatív eredmények a lexikai erőforrások teljesebbé tételére. Az ismeretlen szavak mintázatainak számítógépes vizsgálata rámutat azokra a szabályokra, amelyekkel az ilyen szavak nagy része automatikusan felvehető a lexikonba. Nyelvészeti szempontból szintén lényegesek ezek a szabályok, mivel az anyanyelvi beszélő is ezek mentén képes korábban ismeretlen szavakat alkotni és érteni.
Kulcsszavak: igekötős igék, produktivitás, korpusznyelvészet

1. Bevezetés

A morfológiai produktivitás szerves része a természetes nyelvek működésének. Ennek segítségével folyamatosan, tudatos erőfeszítés nélkül hozunk létre új szavakat [1]. Nyelvtechnológiai szempontból ez felvet egy fontos kérdést: Hogyan dolgozzunk fel olyan szavakat, amelyek nem szerepelnek a lexikonban? A neurális hálón alapuló módszerek számára ez kevésbé problémás, a lexikalista megközelítésben viszont nehézséget jelent. A cikk az utóbbi szellemében járja körül a problémát, az igekötős igék morfológiai produktivitását vizsgálva.

A cikk első felében bemutatom a Magyar Nemzeti Szövegtár 2.0.4 [2] felhasználásával készült PrevLex-et[1], amely a magyar igekötős igék jelenleg legbővebb, manuálisan ellenőrzött táblázata. Szerepelnek benne a korpuszban `UNKNOWN`-nak címkézett szavak és a hapaxok (egyszer előforduló szavak) is. Az utóbbiak alkalmassá teszik a PrevLex-et arra, hogy meg lehessen vele határozni az egyes igekötők produktivitásának mértékét. Erre teszek kísérletet a cikk második felében. A mérések alapján sorra veszem a legproduktívabb igeképzési szabályokat, valamint a produktivitás kapcsolatát a stílusregiszterrel és a gyakorisággal. Végül szó lesz arról, hogy a cikkben ismertetett módszert lehet-e használni az igekötő-állomány meghatározására.

[1] https://github.com/kagnes/prevlex

2. A PrevLex

2.1. Az adatfeldolgozás menete

A PrevLex előállításához közvetlenül az MNSZ 2.0.4 forrásfájlt használtam. Három szűrést végeztem az eredeti korpuszon annak érdekében, hogy a lehető legjobb minőségű szöveganyagot kapjam. Egyrészt kiszűrtem a verseket, mivel sokuk nem természetes nyelvhasználatot tükröz. Másrészt – amennyire csak lehetett – eltávolítottam az idegen nyelvű, valamint a magyar, de ékezet nélkül írt mondatokat, mert torzíthatták volna a keresések eredményét. Például az ékezetet eleve nem tartalmazó igekötős igék sokkal gyakoribbnak tűntek volna, mint az ékezetet tartalmazók. Ehhez azt a heurisztikát alkalmaztam, hogy töröltem minden olyan mondatot, amelyben a tokenek 80%-a `UNKNOWN` vagy `SKIP` elemzést kapott. Ez a módszer inkább a pontosságnak, mintsem a fedésnek kedvezett. Végül igyekeztem kiszűrni a korpuszban található duplumokat. Itt is a pontosságot tartottam szem előtt. Csak a nyolc tokennél hosszabb mondatokat vettem figyelembe a szűrésnél, feltételezve, hogy ennél rövidebb mondatoknál (pl. köszöneséknél) természetes lehet a többszörös jelenlét. Még ezzel az óvatos módszerrel is rendkívül magasnak bizonyult a duplumok aránya (20,12%), a személyes alkorpuszon belül akadt olyan – meglehetősen hosszú – mondat, amely száznál többször ismétlődött. A szűrések eredményét az 1. táblázat foglalja össze.

korpusz	**token**	**százalék**
eredeti MNSZ2	**1 348 000 000**	**100**
versek	5 661 000	0,42
`UNKNOWN`/`SKIP`	26 825 200	1,99
duplumok	271 217 600	20,12
módosított MNSZ2	**1 044 296 200**	**77,47**

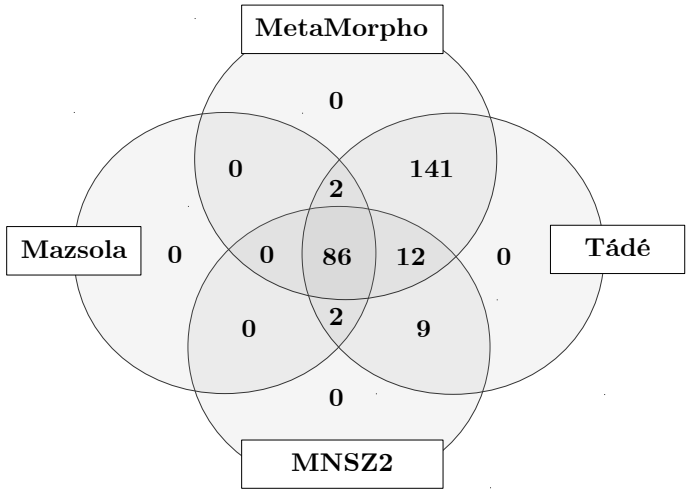
1. táblázat: Az MNSZ 2.0.4 mérete a versek, értelmes elemzés nélküli mondatok, valamint a duplumok szűrése előtt és után. A tokenszám írásjelekkel együtt értendő.

A lehetséges igekötők listája a Manócska[2] integrált igei vonzatkeret adatbázisból származik [3]. Olyan szavak szerepelnek benne, amelyeket a magyar igei vonzatkerettárak közül legalább egy igekötőnek jelöl. Az adatbázis készítői átnézték ezeket a szavakat, és javították az egyértelműnek tűnő hibákat (pl. a *vissz*, *nyug* igekötőnek jelölését a *visszhangoz*, *nyugdíjaz* szavak esetében). Ennek ellenére a végső lista hosszú – összesen 252 tagot számlál –, és több mint kétharmada esetében (pl. *szénné, pofon, zsebre*) az igekötői státusz erősen vitatható. Ez egyúttal jól tükrözi azt is, mennyire nincs egyetértés abban, hogy mely szavakat tekintjük igekötőnek (erről áttekintést ad Komlósy [4]). A 1. ábra azt

[2] http://github.com/ppke-nlpg/manocska

szemlélteti, hogy melyik erőforrás hány igekötőt nevez meg. A MetaMorpho [5] és a Tádé [6] kezelik a legtágabban ezt a kategóriát. Az MNSZ2 alapján készült listák [7] feleannyi jelöltet sem tartalmaznak, a legszigorúbb pedig a Mazsola ([8] és [9]), 90 igekötővel.

A MetaMorpho adatbázis esetében elsősorban a szubjektív annotátori osztályozás határozza meg, hogy mi igekötő és mi nem. A többi erőforrásnál az tűnik vízválasztónak, hogy a kérdéses szó elég gyakori-e, és elég gyakran van-e egybeírva az igével. Például az *utol* és a *zokon* is csak egy-egy igével állnak, de az *utolérte* lényegesen gyakoribb, mint a *zokonvette*. Csak az előbbi annotált igekötős igeként. Megjegyzendő, hogy az egybeírás nem szükséges feltétele az igekötős igévé válásnak. Az egybeírásra való hajlandóságban számíthat a szavak hossza és az is, hogy a főnév ragos vagy ragtalan-e.



1. ábra: A Manócskában szereplő erőforrások összesen 252 szót minősítenek igekötőnek. A halmazok azt mutatják, hogy az egyes erőforrások hány másikkal és hány darab szót illetően értenek egyet.

A kiinduló, 252 szavas listában 13 hibát találtam (ilyen pl. a *vízi*, amely egyszerűen bizonyos szóösszetételek első tagja). Így végül 239 szó maradt, amely az ismertetett források valamelyike szerint igekötő, és én is fenntartom ennek a lehetőségét – hangsúlyozva, hogy az igekötő-állomány összetétele bizonytalan.

A következő lépésben lekértem a módosított MNSZ2-ből minden olyan finit igeként vagy `UNKNOWN`-ként annotált szót, amely egy adott igekötővel kezdődik. Ennek a döntésemnek két része is magyarázatra szorul. Először is az, hogy miért csak a finit igéket vettem figyelembe, amikor az igekötők például igenevekhez is kapcsolódhatnak. A korpuszvizsgálat során azt tapasztaltam, hogy az igenevek esetében erős a tendencia az igemódosító és az igenév egybeírására (pl. *jóltáplált vendég, földreszállt angyal*), míg ugyanezeket az igemódosítókat a fi-

nit igével már kevésbé írják egybe. Valószínű, hogy az igenevek figyelembevétele nem változtatott volna jelentősen a PrevLex összetételén, viszont sokkal több ellenőrizendő adathoz vezetett volna. Másodszor, az `UNKNOWN` szavakra azért volt szükség, mert sok jó találat csak így jelenik meg (pl. *visszacuccol, felstócol, benyammog*). Ugyanakkor az `UNKNOWN` szavak legnagyobb része hibás találat (elírt vagy idegen nyelvű szó), és a finit igék között is akadnak álpozitív találatok (pl. a *túlélősködik* mint igekötős ige). Emiatt az eredményül kapott, közel 178 000 szavas listát át kellett nézni.

Ez a munka körülbelül huszonegy órát vett igénybe. Először eltávolítottam a lehetséges igekötőket a szavak elejéről, és a megmaradó szórészeket néztem át aszerint, hogy egyáltalán igék-e vagy sem. Ezután a már jóval rövidebb listát átnéztem úgy, hogy az ige az adott igekötővel is megfelelő-e (ezen a szinten szűrtem ki pl. a *túlélősködik* és *feltűnősködik* igéket). Néhány olyan esetben, ahol az igekötő+ige kombináció nem volt értelmetlen, viszont nagyon valószínűtlennek tűnt, csak a konkrét szövegbeli előfordulások segítségével tudtam dönteni (pl. a *túltejesít*-ről így derült ki, hogy mindig a *túlteljesít* hibásan írt változata). Ezután lokálisan újraelemeztem a forrásfájlt a javított adatokkal (pl. a korábban `UNKNOWN` *hype-olok, hype-ol* szavakat összevontam egy lemmává). A javított korpuszból állt elő a PrevLex végső változata.

2.2. Nehézségek

Az adatok átnézése során többször felmerült a kérdés, hogy bizonyos szóalakokat nem kellene-e valahogyan normalizálni. Három esetben az igető okozott bizonytalanságot, mert (1) teljesen azonos jelentésű igék történetileg eltérő tőváltozattal rendelkeznek (pl. *verekedik – verekszik*), (2) két igének minimálisan eltérő töve van (pl. *gyömöszköl – gyömöcköl*), (3) egy-egy neologizmus többféle írásváltozatban létezik (pl. *dizájnol – design-ol – designol*). Egyedül az utóbbi csoport kapcsán voltam biztos abban, hogy a különbség csak ortográfiai jellegű. Ezeket a szavakat normalizáltam – rendszerint a magyar kiejtés szerint írt változatra –, mindenhol megőrizve az eredeti szóalakot is.

Elkülöníthető továbbá három olyan probléma, amely a képzőt érinti: amikor (1) két vagy több ige képzőjében csak a kötőhang tér el (pl. *feccel – feccöl – feccol*, (2) opcionálisan -ikes végződésű az ige (pl. *szörföz – szörfözik*), (3) ugyanaz az ige -(O)z és -(O)l képzővel is előfordul (pl. *offtopicol – offtopicoz*). Bár itt is szólhatnak érvek a normalizálás mellett, annyi biztos, hogy nem egyszerű ortográfiai különbségekről van szó. A (3)-asban látható példák egyelőre még ugyanazt jelentik, de elképzelhető, hogy idővel kis jelentésbeli eltérés kapcsolódik hozzájuk (ahogy azt pl. a *házal – házaz* párnál látjuk). A normalizálást ezekben az esetekben önkényesnek találtam, és nem vállalkoztam rá.

2.3. A PrevLex felépítése

Az erőforrás egy `TSV` fájlként érhető el, amely öt oszlopból áll. Az első oszlopban szerepel az ige (igekötő+igelemma formában). Ezt követi az MNSZ2-ben

mért tokengyakoriság. A harmadik oszlopban kétféle érték szerepelhet attól függően, hogy az ige kapott-e megfelelő annotációt az MNSZ2-ben (`FIN`, ha igen és `UNKNOWN`, ha nem). A negyedik oszlop azt jelzi, hogy az ige hány dokumentumban fordult elő. Ez fontos információ lehet akkor, ha a tokengyakoriság és a tartalmazó dokumentumok száma nincs arányban (pl. az ige százszor fordul elő, de mindössze egy dokumentumban). Utolsóként szerepel a normalizált alak, amely csak a neologizmusoknál térhet el az első oszlop tartalmától.

Bár az igekötős igék listája manuálisan ellenőrzött, a gyakorisági adatok fenntartással kezelendők. Néhány igealak ugyanis több lehetséges elemzéssel rendelkezik (pl. a *leszel* egyik lehetséges elemzése a *lenni* E/2. alakja, a másik a *leszel* igekötős ige). Ezek az elemzések sokszor eleve rosszak a forrásfájlban, így kissé torzíthatják a gyűjtött statisztikát.

kategória	**típus**	**token**
összes igekötős ige	54 955	11 959 379
hapaxok	22 043	22 043
`UNKNOWN` szavak	5 156	26 542
`UNKNOWN` hapaxok	3 335	3 335

2. táblázat: A PrevLex számokban. Az értékek az eredeti igealakokra vonatkoznak, nem a normalizáltakra.

A 2. táblázat számszerű áttekintést ad a PrevLex-ről. A várakozásnak megfelelően az igekötős igék Zipf-eloszlást mutatnak: néhány ige rendkívül nagy tokengyakorisággal bír, míg a hapaxok ritkák, de sokfélék. Az utóbbi tulajdonságuk miatt bizonyulnak hasznosnak a morfológiai produktivitás kvantitatív meghatározásában.

3. Az igekötők produktivitásának vizsgálata

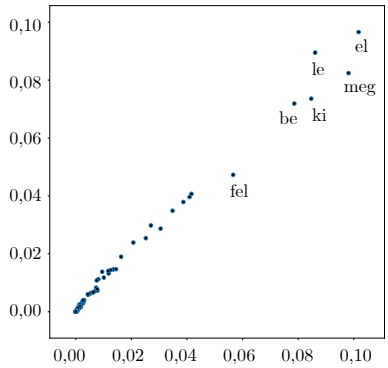
Kiefer és Ladányi (2000) [10] szerint egy szóalkotási mintát akkor tekinthetünk produktívnak, ha a minta alapján tetszőleges számú, szemantikailag transzparens szó hozható létre egy adott szemantikai tartományban. A morfológiai produktivitás esetében is – mint minden nyelvi jelenségnél – célszerű rávilágítani a kompetencia és a performancia közti különbségre. A nyelvi rendszer szintjén a produktivitás egy lehetőség, a minden további nélkül létrehozható szóalakok nem biztos, hogy ténylegesen létrejönnek, és még kevésbé valószínű, hogy benne lesznek egy korpuszban. Emiatt több elméleti morfológus (pl. Dressler [11]) nem tartja helyesnek a kompetenciaszintű lehetőség performanciaszintű valószínűségen alapuló vizsgálatát. Ez a cikk mégis az utóbbit célozza meg, mivel általában véve ezt a módszert tartom a leginkább objektívnek és reprodukálhatónak, az eredmény pedig tendenciák szintjén érdekes lehet a kompetenciát kutatóknak is.

A morfológiai produktivitás kvantitatív meghatározása Baayen nevéhez köthető ([12] és [13]). Három típust különböztet meg: a megvalósult (*realized*), a terjeszkedő (*expanding*) és a lehetséges (*potential*) produktivitást. A következőkben arról lesz szó, hogy pontosan mik ezek a típusok, és hogyan jellemeznek egy-egy igekötőt.[3]

3.1. Megvalósult és terjeszkedő produktivitás

A megvalósult produktivitás annak a mértéke, hogy egy adott affixum a mérés időpontjáig mennyire vett részt a szóalkotásban, tehát a múltbeli és a jelenlegi szerepe jellemezhető ezáltal. Úgy kapjuk meg, hogy az affixumot (itt: igekötőt) tartalmazó lemmák darabszámát elosztjuk a korpuszban (itt: a PrevLex-ben) található összes lemma darabszámával.

A terjeszkedő produktivitás arról ad jóslatot, hogy az affixumnak várhatóan mekkora szerepe lesz a szóalkotásban a közeljövőben. Ehhez az affixumot tartalmazó hapaxok darabszámát osztjuk el a korpuszban található összes hapax darabszámával. Ez azért is jó mérték, mert a hapaxok jelentése szinte minden esetben kompozicionális, ezért kevesebb „hamis produktív" találat adódik hozzá az eredményhez, mint a megvalósult produktivitás esetében. A 2. ábra áttekintést ad a PrevLex igekötőinek kétféle produktivitásáról.[4]



igekötő	P_m	P_t
el	0,1018	0,0970
meg	0,0984	0,0825
le	0,0863	0,0899
ki	0,0848	0,0737
be	0,0785	0,0718
fel	0,0563	0,0471
át	0,0418	0,0408
bele	0,0411	0,0398
vissza	0,0389	0,0380
össze	0,0348	0,0349

2. ábra: Az igekötők megvalósult (P_m) és terjeszkedő (P_t) produktivitása. Bal oldalt a két mérték összefüggése síkban ábrázolva látható, az X-tengelyen a P_m, az Y-tengelyen a P_t értékeivel. Jobb oldalt a tíz legmagasabb értéket kapott igekötő szerepel.

[3] A következő igekötőket alakvariánsokként kezeltem, és összevontam minden mérés előtt (a párok első tagját meghagyva): *bele – belé, be – bé, fel – föl, odább – odébb, rá – reá, tele – teli.*

[4] Az ábrát Makrai Márton készítette. Kiegészítő információkkal együtt elérhető az alábbi címen is: https://github.com/makrai/misc/blob/master/tade/kalivoda19-mszny.ipynb

Az ősi igekötők (*meg, el, le, ki, be, fel*) mindkét mérték szerint kiugróan produktívak, bár a *fel* elmarad a többitől. Látható az is, hogy a kétféle produktivitás nagyjából egyenesen arányos. A tendenciától csak a *le* és a *meg* térnek el. A *le* azért is figyelemre méltó, mert a terjeszkedő produktivitása nagyobb, mint a *meg*-é. Ez azt jelenti, hogy várhatóan az igekötős igék alkotásában is egyre nagyobb szerepe lesz.

Mielőtt áttérnénk a harmadik produktivitás-típusra, érdemes alaposabban megvizsgálni az eddig látott eredmények okait azok fontossági sorrendjében. Három tényezőről lesz szó, amelyek közül leglényegesebbek a produktív szóképzési szabályok.

3.2. Produktív szóképzési szabályok

Minden olyan igekötőnél, amelynek a P_m-je és P_t-je nagyobb 0-nál, megfigyelhető az igéből igét képző produktív szabályok megléte (ilyen például a *-gat/get* gyakorító képző használata, ha ennek nincs szemantikai korlátja). Lényegesen kevesebb igekötőre igaz viszont az, hogy névszóból képzett igéhez kapcsolódhat. Épp ezért a névszóból igét képző szabályok azok, amelyek produktivitás szempontjából látványosan kiemelnek bizonyos igekötőket a többi közül.

A korábban nem hallott, de „mintaszerűen" képzett szavaknak azért tudunk jelentést tulajdonítani, mert egy ismert, alapjelentéssel bíró sémába illeszkednek (ld. 3. táblázat). Az igekötők ezt teszik specifikusabbá, illetve gyakran további jelentéssel gazdagítják a létrejövő szót.

alapjelentés	**sematikus szerkezet**	**néhány példa**
N-nel kapcsolatosat csinál	N+(O)z(ik) N+(O)l	*kisfiamozik, testékszerez* *rajzszögel, bokroscsomagol*
N-né változik	N+Ul/sUl N+Odik/sOdik	*mémesül, szinglisül* *vékonyodik, tahósodik*
N-né változtat	N+ít/(O)sít	*részegít, szálkásít*
N-ként viselkedik	N+kOdik/skOdik	*vandálkodik, jópofáskodik*

3. táblázat: A hat legproduktívabb szabály, amellyel névszóból ige képezhető. A táblázatban az N tetszőleges névszót jelöl, az /O/ archifonéma az /o/, /ö/ és /e/, az /U/ pedig az /u/ és /ü/ fonémák helyett áll.

Példaként vizsgáljuk meg – a teljesség igénye nélkül – azokat a többletjelentéseket, amelyeket a *le* igekötő ad a névszóból képzett igének. Kifejezhet (1) lefelé történő mozgást (pl. *leszánkózik, leteherautózik*), (2) egy felület lefedését valamivel (pl. *leszőnyegpadlóz, leszemfedelez*), (3) támadást vagy rombolást valamilyen eszközzel (pl. *lemacsetéz, levízágyúz*), és azt, hogy (4) valakit vagy valamit nevezünk valahogy (pl. *lelófogúzik, lebölcsészlányoz, legyíkarcoz*). Ez a

sokféle többletjelentés az oka annak, hogy terjeszkedő produktivitás szempontjából a *le* felülmúlja a *meg* igekötőt.[5]

Említést érdemelnek még azok a produktív szabályok is, amelyek egy szótagszerkezeti séma alapján tetszőleges számú hangutánzó igét hoznak létre. Ezek aztán tovább kombinálódhatnak bizonyos – főként irányjelölő – igekötőkkel. A leggyakoribbak a `CVC:+Og` (pl. *cimmog, nyammog, kaffog, hümmög*) és a `CVC:+An`, `CVC:+En` (pl. *nyekken, csisszen, toccsan, suppan*) sémákra illeszkedő igék.

3.3. Produktivitás és stílusregiszter

Az MNSZ 2.0.4 összesen 2952 dokumentumból áll, amelyek mindegyike egy alkorpuszhoz van rendelve. Az alkorpuszok a következők: személyes, beszéltnyelvi, szépirodalom, sajtó, tudományos, hivatalos. Ezeknek a metaadatoknak a segítségével könnyen ki lehetett mérni, hogy van-e összefüggés az igekötők produktivitása és a vizsgált szövegek stílusregisztere között. A 4. táblázatban látható, hogy tíz igekötő hapaxai milyen arányban szerepelnek az egyes alkorpuszokban.

	személyes	**beszélt**	**szépirod.**	**sajtó**	**tud.**	**hivatalos**
MNSZ2	**28,96**	**7,33**	**7,75**	**35,09**	**11,34**	**9,52**
el	↑ 34,62	6,10	↑ 28,56	↓ 18,21	9,57	↓ 2,93
meg	31,02	7,44	↑ 31,25	↓ 14,94	12,59	↓ 2,75
le	↑ 45,18	6,90	↑ 17,64	↓ 20,33	7,00	↓ 2,95
ki	33,06	6,53	↑ 29,66	↓ 17,36	10,38	↓ 3,01
be	↑ 41,70	8,17	↑ 22,09	↓ 16,20	8,77	↓ 3,08
fel	↑ 34,55	6,59	↑ 24,54	↓ 19,05	13,68	↓ 1,59
át	27,46	6,84	↑ 28,16	↓ 23,06	11,24	↓ 3,24
bele	↑ 33,99	6,23	↑ 33,13	↓ 17,97	↓ 5,99	↓ 2,69
vissza	27,74	7,21	↑ 33,21	↓ 20,52	7,21	↓ 4,10
össze	32,70	6,89	↑ 31,62	↓ 17,03	9,86	↓ 1,89

4. táblázat: A tíz legmagasabb P_m és P_t értékű igekötő hapaxainak százalékos eloszlása az MNSZ 2.0.4 alkorpuszaiban. A második sorban vastagon kiemelve az látható, hogy az adott alkorpusz tokenjei az MNSZ2-nek mekkora részét képezik. A ↑ azt jelzi, ha az adott igekötőnél egy alkorpusz legalább 5%-kal nagyobb arányban van jelen az eredetinél, a ↓ azt jelzi, ha legalább 5%-kal kisebben.

Ahogy várható volt, a formális regiszter (a hivatalos, tudományos és sajtószövegek nyelvezete) kevés teret enged a produktivitásnak. Érdekes viszont, hogy a sajtónyelvre a tudományoshoz képest kevésbé jellemző a produktivitás, pedig

[5] A többletjelentések is eltérő produktivitással bírnak, például a *le* esetében a (4)-es jelentés jóval produktívabbnak tűnik, mint a (3)-as. A jelentéscsoportok automatikus elkülönítése nem lehetetlen ugyan – például szóbeágyazást alkalmazó módszerekkel –, de komoly utómunkálatot igényel, ezért ebben a cikkben nem vállalkozok rá.

a beszélt nyelv jobban hat rá. Az új szóalakok alkotása az informális regiszterhez (főképpen a személyes szövegekhez) és a szépirodalomhoz köthető. Néhány igekötő (pl. *tova, által*) produktivitása szinte csak a szépirodalmi alkorpuszban mutatkozik meg.

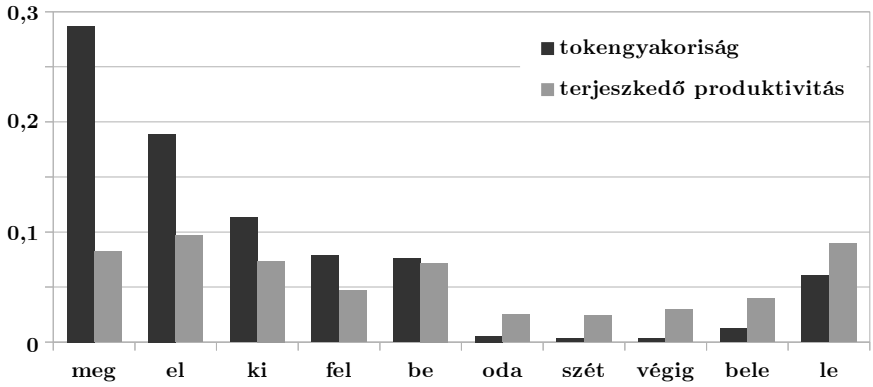
3.4. Produktivitás és gyakoriság

A kapott eredmények alapján feltételezhetnénk, hogy a produktivitás szorosan összefügg a gyakorisággal. Ez nem így van: jó ellenpélda az *agyon*, amely a tokengyakoriság szerinti rangsorban a 46., a P_t szerintiben a 20. helyet foglalja el. A produktív affixumok nem feltétlenül gyakoriak. A gyakoriság és a produktivitás kapcsolata egy 2x2-es mátrixszal írható le, amelyet az 5. táblázat szemléltet.

	produktív	**nem produktív**
gyakori	• sok típus, sok token • pl. *el, be*	• kevés típus, sok token • pl. *létre, egyet*
nem gyakori	• sok típus, kevés token • pl. *pofon, szénné*	• kevés típus, kevés token • pl. *hajba, síkra*

5. táblázat: A produktivitás és a gyakoriság kapcsolata, Pakerys [14] alapján.

A szoros összefüggés hiányát igazolja a 3. ábra is. Az egyik szélsőséges csoportot a *meg*, *el*, *ki*, *fel* és *be* igekötők alkotják, amelyek sokkal kevésbé produktívak, mint ahogy a gyakoriságuk alapján várnánk. A másik szélsőséges csoport az *oda*, *szét*, *végig*, *bele* és *le*, amelyek esetében épp az ellenkező tendencia látható.



3. ábra: Az a tíz igekötő, amelynél a legnagyobb az eltérés a tokengyakoriság és a terjeszkedő produktivitás mértéke között.

3.5. Lehetséges produktivitás

A Baayen által definiált produktivitás-típusok harmadik tagja a lehetséges produktivitás. Ez az egészen távoli jövőről ad jóslatot: mik azok a most még viszonylag ritkán előforduló affixumok, amelyeknek jó esélye van arra, hogy később sok szó képzésében vegyenek részt? A terjeszkedő produktivitáshoz hasonlóan ez is hapax-alapú mérték, de az eddig látottaktól élesen eltérő eredményt hoz.

Úgy kapjuk meg, hogy egy adott affixumhoz tartozó hapaxok tokengyakoriságát elosztjuk az affixumhoz tartozó összes szó tokengyakoriságával. A méréshez célszerű gyakorisági küszöböt választani. Minél kisebb tokengyakorisággal osztunk, annál magasabb – és annál kevésbé informatív – lesz a lehetséges produktivitás. A mérést 5-ös és 5000-es küszöbbel (ld. 6. táblázat) végeztem el.

igekötő	token	hapax	P_1
mennybe	6	4	0,6667
oldalba	18	9	0,5000
égbe	10	5	0,5000
szarrá	18	8	0,4444
szénné	7	3	0,4286
fejen	16	6	0,3750
torkon	11	4	0,3636
seggre	9	3	0,3333
tűzbe	13	4	0,3077
porba	10	3	0,3000

igekötő	token	hapax	P_1
tele	5 824	225	0,0386
agyon	7 852	293	0,0373
körbe	11 304	298	0,0264
ide	15 499	305	0,0197
körül	12 940	238	0,0184
hátra	8 381	142	0,0169
végig	41 772	629	0,0151
utána	8 407	125	0,0149
előre	12 633	166	0,0131
egybe	10 163	132	0,0130

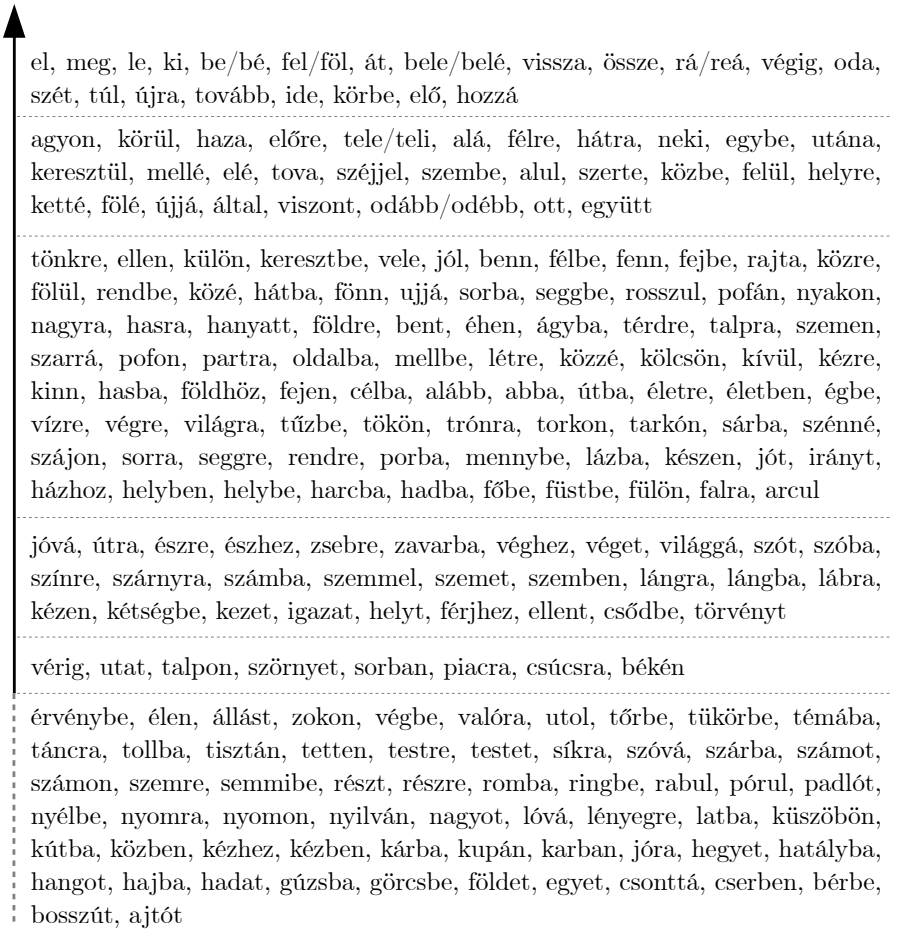
6. táblázat: Lehetséges produktivitás (P_1), a 10 legmagasabb értéket kapott szó 5-ös küszöb (bal oldalon) és 5000-es küszöb (jobb oldalon) mellett.

A 6. táblázat bal oldalán szereplő szavakat egyetlen nyelvészeti szakirodalom sem sorolná az igekötők közé. Tény viszont, hogy „igekötőszerűbben" viselkednek sok más igemódosítónál, és emiatt nem ritka, hogy egybe is íródnak az igével. Az előfordulásaik több dokumentumra oszlanak el, tehát nem csak egy ember szóhasználatát látjuk. Többségük egy jól meghatározható szemantikai tartományban mutatja a produktivitás jeleit, például a *mennybe* és az *égbe* mozgásigékkel (*megy, száll*), az *oldalba* támadást kifejező igékkel (*szúr, rúg*) áll. A *szarrá* (*ázik, fagy, bombáz*) és a *szénné* (*ég, vakuz, tetovál*) már absztraktan értendők – a *szét* stilisztikailag jelölt változatai.

A 6. táblázat jobb oldalán rangsorolt szavak státusza kevésbé megosztó: a szakirodalomban is felbukkannak igekötőként, bár az egyetértés ezeket illetően sem általános. Könnyen elképzelhető, hogy idővel lényegesen több új szó alkotásában vesznek majd részt. Az itt látható igekötők mindegyike kapcsolódhat névszóból képzett igéhez (pl. *végigszambáz, telekommentel, agyonpárnáz, körbekordonoz, idekontárkodik*), ami táptalajt nyújt a kreatív szóalkotásnak.

3.6. Következtetések

A kapott eredmények amellett szólnak, hogy az igekötők állományát ne intuitív módon megszabott szükséges és elégséges feltételek mentén határozzuk meg. Célravezetőbb lehet az a felfogás, amely szerint az igekötőségnek különböző fokozatai vannak (ld. még [15] és [16]). Az általam vizsgált 239 szó a háromféle produktivitása alapján jól elhelyezhető egy kontinuumban, és ezen belül hat nagyobb csoportra osztható (ld. 4. ábra).



4. ábra: Kontinuum, amely a vizsgált 239 szó háromféle produktivitása alapján rajzolódik ki. A nyíl mentén felfelé haladva a produktivitás mértéke egyre nő. A vízszintes, szaggatott vonalak az egyes csoportok közötti, semmiképp sem éles határokat jelzik.

A 7. táblázat azokat a szempontokat mutatja be, amelyek szerint az igekötőjelölteket csoportosítottam. Ezek sorrendje lényeges: ha egy szó a produktivitásértékei alapján az 1. csoportba tartozás feltételét nem teljesítette, akkor vizsgáltam a 2. csoportba tartozást, és így tovább. A csoportra bontás nem áll ellentétben az igekötő-kategória kontinuum-jellegével. Egyszerűen az eredmények áttekintését és tárgyalását hivatott segíteni.

csoport	**feltétel**
1.	$P_m > 0{,}01$ és $P_t > 0{,}01$
2.	$P_m > 0{,}001$ és $P_t > 0{,}001$
3.	$P_m > 0{,}0001$ és $P_t > 0{,}0001$
4.	$P_m > 0$ és $P_t > 0$
5.	$P_m = 0$ és $P_t = 0$ és $P_l > 0$
6.	$P_m = 0$ és $P_t = 0$ és $P_l = 0$

7. táblázat: Feltételek, amelyek mentén a hat csoport kialakult. A P_m a megvalósult, a P_t a terjeszkedő, a P_l a lehetséges produktivitás mértékét jelöli.

Az 1. csoportot, egyúttal a kontinuum egyik végpontját a kimagaslóan produktív igekötők alkotják (pl. *be, össze*). A 2. csoport igekötői (pl. *agyon, félre*) közepesen produktívak. A legnépesebb, 84 tagú 3. csoport most még nem túl produktív, de várhatóan azzá váló szavakat foglal magába (pl. *tönkre, szénné*). Ezek között már nagy számban találunk olyan igemódosítókat, amelyeket a nyelvészeti szakirodalmak többsége nem sorolna az igekötők közé.

A 4. csoport tagjai (pl. *jóvá, világgá*) jellemzően csak egy-két igével állnak (pl. *zsebrevág*, ritkábban *-dug, -tesz, -rak*). Az 5. csoportról az mondható el, hogy a tagjai (pl. *csúcsra, békén*) nem produktívak, de minimális esély van arra, hogy idővel produktívabbak lesznek. A kontinuum másik végpontját alkotó 6. csoport 61 olyan igemódosítót tartalmaz, amely kizárólag egy igével áll (pl. *pórul* → *póruljár*, *lóvá* → *lóvátesz*).

4. Összefoglalás

A cikkben bemutattam a nyilvánosan elérhető, 54 955 igekötős igét tartalmazó PrevLex táblázatot, amelyet arra használtam, hogy kvantitatív módon meghatározzam az egyes igekötők morfológiai produktivitását. Az így kapott eredmények azt az elképzelést támasztják alá, miszerint az igekötő-kategória kontinuumként értelmezendő.

A PrevLex anyagával bővíthetők a morfológiai elemzők (például az `emMorph` [17]) lexikonjai, ezáltal csökkenthető az `UNKNOWN`-ként elemzett szavak száma. A produktív szóképzési szabályoknak a 3.2. alfejezetben felvázolt, ám ennél szisztematikusabb és teljesebb leírásával a lexikonírás kevesebb humán erőforrást igényel. Mindez jobb lexikont – ezáltal pontosabb nyelvmodelleket – eredményezhet.

Köszönetnyilvánítás

Köszönöm Olsvay Csabának, Indig Balázsnak és Makrai Mártonnak a cikk többszöri átnézését és a hozzá fűzött értékes megjegyzéseket. Köszönet illeti Prószéky Gábort és mindkét névtelen bírálót a hasznos tanácsokért. Sass Bálintnak köszönöm az MNSZ 2.0.4-hez adott közvetlen hozzáférést.

Jelen kutatás az FK 125217 és a PD 125216 számú projekt keretében az FK17 és a PD17 pályázati program finanszírozásában a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással valósult meg.

Hivatkozások

1. Ladányi, M.: Produktivitás és analógia a szóképzésben: elvek és esetek. Tinta Könyvkiadó, Budapest (2007)
2. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., et al., eds.: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Izland, European Language Resources Association (ELRA) (2014) 1719–1723
3. Kalivoda, Á., Vadász, N., Indig, B.: Manócska: A Unified Verb Frame Database for Hungarian. In Sojka, P., Horák, A., Kopeček, I., Pala, K., eds.: Proceedings of the 21st International Conference on Text, Speech and Dialogue (TSD), szeptember 11–14, 2018, Brno, Csehország, Springer-Verlag (2018) 135–143
4. Komlósy, A.: Régensek és vonzatok. In Kiefer, F., ed.: Strukturális magyar nyelvtan 1., Mondattan, Budapest, Akadémiai Kiadó (1992) 299–527
5. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: a robust high-performance parser and generator. In Hutchins, J., ed.: Proceedings of the 9th EAMT Conference, La Valletta, Málta, Foundation for International Studies (2004) 138–142
6. Kornai, A., Nemeskey, D.M., Recski, G.: Detecting Optional Arguments of Verbs. In Calzolari, N., et al., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Szlovénia, European Language Resources Association (ELRA) (2016) 2815–2818
7. Kalivoda, Á.: A magyar igei komplexumok vizsgálata. (2016) Mesterszakos szakdolgozat. PPKE-BTK. https://github.com/kagnes/hungarian_verbal_complex.
8. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest (2010)
9. Sass, B.: 28 millió szintaktikailag elemzett mondat és 500 000 igei szerkezet. In Tanács, A., Varga, V., Vincze, V., eds.: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015), Szeged, Szegedi Tudományegyetem Informatikai Intézet, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2015) 399–403
10. Kiefer, F., Ladányi, M.: A szóképzés. In Kiefer, F., ed.: Strukturális magyar nyelvtan 3., Morfológia, Budapest, Akadémiai Kiadó (2000) 137–164
11. Dressler, W.U.: Degrees of grammatical productivity in inflectional morphology. Rivista di Linguistica (Italian Journal of Linguistics) **15**(1) (2003) 31–62
12. Baayen, H.: A Corpus-Based Approach to Morphological Productivity (Statistical Analysis and Psycholinguistic Interpretation). (1989) Doktori értekezés. Centrum voor Wiskunde en Informatica, Amszterdam, Hollandia.
13. Baayen, H.: Corpus linguistics in morphology: morphological productivity. In Lüdeling, A., Kytö, M., eds.: Corpus Linguistics. An international handbook, Berlin, Mouton De Gruyter (2009) 900–919

14. Pakerys, J.: Measuring morphological productivity (2017) Graduate School of Linguistics, Philosophy and Semiotics (GSLPS), Tartu, Észtország, március 20, 2017. Handout. http://web.vu.lt/flf/j.pakerys/wp-content/uploads/pakerys-measuring-morphological-productivity-tartu-2017-handout.pdf.
15. Kerekes, J.: Az igekötők meghatározásának problémái. In Gécseg, Zs., ed.: LingDok 10. Nyelvészdoktoranduszok dolgozatai, Szeged, JATEPress (2011) 109–130
16. Forgács, T.: Grammatikalizálódás az igekötők körében. In Oszkó, B., Sipos, M., eds.: Uráli grammatizáló, Budapest, MTA Nyelvtudományi Intézet (2005) 88–116
17. Novák, A., Siklósi, B., Oravecz, Cs.: A New Integrated Open-source Morphological Analyzer for Hungarian. In Calzolari, N., et al., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Szlovénia, European Language Resources Association (ELRA) (2016) 1315–1322