

A deep learning szoftveres vizsgálat lehetőségei a
mentális lexikonelemek szintaktikai szerepeinek
meghatározásában

Jelige: Ginkgo

Tartalomjegyzék

1. Általános bevezetés.....	2
2. A mentális lexikon megközelítésének lehetőségei.....	2
2.1. A mentális lexikon megközelítése a kognitív nyelvelmélet fogalomrendszerével.....	2
2.1.1. A kognitív nyelvészet elméleti alapállásai.....	2
2.1.2. A mentális lexikon általános szerepe a kognitív nyelvészetben	3
2.1.3. A mentális lexikon hálózatos szerveződése a kognitív nyelvészetben	3
2.2. A mentális lexikon megközelítése a hálózatelmélet fogalomrendszerével	4
2.3. A mentális lexikon szintaktikai szerepe a kognitív nyelvelmélet és a hálózatelmélet keresztmetszetében	5
3. A dolgozat témájául szolgáló kutatás ismertetése	5
3.1. A dolgozat tárgya.....	5
3.2. A dolgozat célja	6
3.3. A dolgozat módszerei	7
4. A kutatáshoz használt szoftveres bázis ismertetése	8
4.1. Az NLP deep learning szoftverek általános ismertetése	8
4.2. A kutatáshoz használt deep learning szoftver elemei	9
4.2.1 Futtatási környezet és az alapkészségek moduljai	9
4.2.2. A korpusz előkészítése az elemzéshez.....	10
4.2.3. A szavak beágyazását jellemző vektorok tulajdonságai	11
4.2.4. A szoftveres öntanulás működése és a vektorok kinyerése	11
4.3. A deep learning szoftveres vizsgálat által felismert kapcsolódási mintázatok vizualizálása.....	12
5. A deep learning szoftver által szolgáltatott eredmények.....	12
5.1. A deep learning szoftver által megalkotott lexikon elemei a hipotézis fényében	12
5.2. Konklúziók	18
6. Kitekintés, további lépések	19
Felhasznált irodalom.....	20

1. Általános bevezetés

Dolgozatom tárgya annak megvizsgálása, hogy egy deep learning szoftver (mélytanuló mesterséges intelligencia) milyen keretek között alkalmazható egy, a kognitív nyelvészet talaján álló, illetve a lexikonelemek kapcsolódási valószínűsége mentén szerveződő új mondattani szemlélet elméleti alapjainak lerakásához. Vizsgálatom során egy szakirodalmi források segítségével, általam összeállított deep learning szoftverrel végeztem analízist egy nagyjából 12.500.000 szavas magyar, írott nyelvi korpuszon, olyan lexikonbeli összefüggéseket keresve, melyek új szempontokat adhatnak a kognitív szintaxishoz. A szoftveres vizsgálat eredményei több ilyen tényezőre is felhívták a figyelmet, többek között a szemantikai keretek és forogatókönyvek, illetve az ige-főnév szemantikai viszonyrendszerbe nem illeszkedő elemek feltételezhetően jelentős, lexikonbeli viszonyrendszerükkel magyarázható mondatszerkezeti szerepére. Jelen dolgozat tárgya tehát a kiinduló elméleti és módszertani tájékozódás elvégzése egy hosszú távú kutatási elképzeléshez, a fentebb említett szintaktikai szemlélet kidolgozásához.

2. A mentális lexikon megközelítésének lehetőségei

2.1. A mentális lexikon megközelítése a kognitív nyelvelmélet fogalomrendszerével

2.1.1. A kognitív nyelvészet elméleti alapállásai

A nyelvvel való foglalkozás kognitív paradigmája a nyelvet nem önálló, kontextusfüggetlen rendszerként, hanem a teljes emberi kogníció integráns részeként kívánja vizsgálni, ebből következően a nyelvi jelenségek leírása során is az általános megismerési mechanizmusokra kidolgozott fogalomrendszert használja. Leírási keretrendszerének szemlélete holisztikus, méghozzá a nyelv minden vonatkozásában, azaz a produkciót és a percepciót egyetlen rendkívül komplex, hálózatosan szerveződő rendszer által végzett folyamatnak tekinti, amely a nyelv elsajátítása és használata (illetve az összes többi kognitív funkció működése) során folyamatosan magát a rendszert is módosítja a hálózat új elemekkel való bővítésével, és elemek közti új kapcsolatok létrehozásával. Ebből is látható, hogy a kognitív szemlélet szerint a nyelvelsajátítás és a nyelvhasználat egymással állandó, dinamikus kölcsönviszonyban állnak, hiszen egyetlen rendszernek, a tágran értelmezett megismerésnek a részfolyamatai. Az embernek a világról és a nyelvről való ismerete tehát a nyelvhasználat során

folyamatosan bővül és átstrukturálódik. (Az egész bekezdéshez lásd TOLCSVAI NAGY 2013: 27–83., 2017: 30–33.)

2.1.2. A mentális lexikon általános szerepe a kognitív nyelvészetben

A kognitívista felfogás szerint a nyelvhasználat során működtetett, folyamatosan változó nyelvi tudáshalmaz hálózatosan szerveződő szótárt épít fel. Itt tárolódnak a nyelvi jelek fonológiai és szemantikai szerkezetét felépítő alszerkezetek (mikrojegyek¹), melyek a kategorizáció, az absztrakció és a sematizáció hármas folyamatában nyerik el helyüket a lexikonban. A kategorizáció során az elme a világról vizuális és auditív csatornán keresztül érkező információkat hasonlóság alapján szétválogatja, és csoportokba rendezi, majd absztrahálás útján elhatárolja a felismerhetőség alapjául szolgáló tulajdonságokat, végül pedig ezen absztrakt tulajdonságok között sémák alapján kapcsolatokat teremt.

Az így rögzült mikrojegyek a gondolkodás és a kommunikáció aktuális helyzeteiben a potenciálisan lehetséges kapcsolódási minták közül azt az egyet valósítják meg, amely a rendszer ítélete szerint a leginkább megfelel a pillanatnyi szituációnak. A számos mikrojegy közötti kapcsolatok aktuális aktivizálódása által így profilálódik az adott kommunikációs helyzetben alkalmazott nyelvi jelentés. A profilálódás során gyakran együtt aktiválódó mikrojegyekből felépülő jelentés a fogalom, amely egy jelenség tipikus, sematizált tudati reprezentációja. (Az egész ponthoz lásd TOLCSVAI NAGY 2013: 114–168., 2017: 33–41, 207–225., 2021: 18.)

2.1.3. A mentális lexikon hálózatos szerveződése a kognitív nyelvészetben

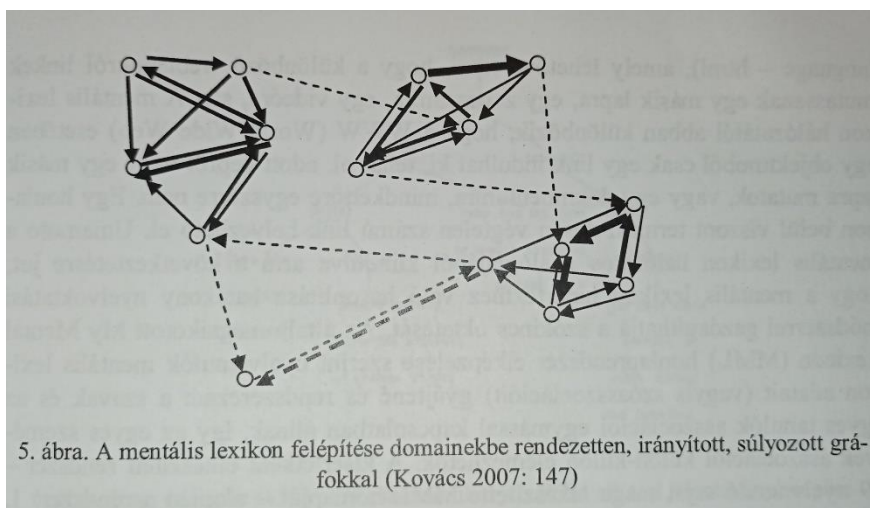
A kognitív szemlélet elméletalkotói szerint a nyelv egészének és ebből következően a lexikonnak a szerveződése is hálózat formájában adható meg: a fogalmakat azon csomópontok jelölik, amelyeket erősebben súlyozott kapcsolatok kötnek össze egy jól körülhatárolható mikrojegy-együttes pontjaival, mint más pontokkal. A fogalmak között hasonlóképpen különböző súlyozottságú kapcsolatok állnak fenn, melyek erőssége a nyelv használata és a lexikon folyamatos átstrukturálódása következtében állandó változásban van. Ezen kapcsolatok rendszere hordozza magában az egyes kommunikációs szituációk közös figyelmi jeleneteiben potenciálisan létrehozható megnyilatkozások szerveződésének valószínűségi mintázatait. Tehát

¹ A kognitív nyelvészeti szakirodalmak nem használják a *mikrojegy* megnevezést, a pszicholingvisztika azonban olyan tudati–nyelvi jelenséget jelöl ezzel a terminussal, amit (bár nem teljesen azonosítható az *alszerkezet* által jelölt jelenséggel) a könnyebb hivatkozás és érthetőség érdekében (a különbözőségek mindenkor szem előtt tartásával) mégis az alszerkezet szinonimájaként alkalmazok.

az elmélet szerint nem létezik a szótártól elszeparáltan működő, külön grammatikai–szintaktikai modul, hanem a lexikonelemek rendszeralkotó elveiből adódnak a mondatok létrejöttében szerepet játszó mintázatok, azaz a szótár belső rendezőelveinek rendszere maga a szintaxis. (Az egész bekezdéshez lásd TOLCSVAI NAGY 2013: 54–58.)

2.2. A mentális lexikon megközelítése a hálózatemélet fogalomrendszerével

A hálózatemélet számos olyan szemponttal egészíti ki a mentális lexikon értelmezését, melyek összeegyeztethetők a kognitív nyelvészet fentebb említett megállapításaival, így bővítve és finomítva a hálózatos struktúrára és annak szintaktikai következményeire vonatkozó elméleti bázist. A számos, egymással sokszor részleges átfedésben levő megközelítés összegzéséként közli KOVÁCS (2011: 26–27) a „domainekben rendezetten, irányított, súlyozott gráfokkal” ábrázolható mentális hálózatról alkotott elképzelését.



1. ábra. Az egyéni mentális lexikon gráfszerű ábrázolásának egyszerűsített szemléltetése.

KOVÁCS 2011: 26–27.

A szóban forgó elmélet szerint a csúcsokban elhelyezkedő lexikonelemek közötti kapcsolódási valószínűségek a gráf éleinek súlyozásával, vagyis a kapcsolódási gyakoriságokból számított értékekkel (valószínűségi változókkal) adhatók meg. Mindez grafikus ábrán az élek vastagságával szemléltethető. Ha a gráffal szemléltetett hálózatban bizonyos elemek között úgy magas a kapcsolódás valószínűsége, hogy eközben más elemekkel szignifikánsan alacsonyabb, akkor ezek egy domaint alkotnak, amely a kognitív nyelvészet nézőpontjából erős szemantikai (és ebből következően szintaktikai) viszonyban álló

lexikonelemek halmazának tekinthető. Ilyen lehet például egy ige és a figuratív szerkezetének üres helyeit nagy valószínűséggel kidolgozó főnévi elemek közös domainje, vagy egy közös prototípus körül szerveződő fogalmak halmaza, vagy akár a névelők és a főnevek gyakori kapcsolatából adódó domain. Ezen példák alapján is sejthető, hogy a domainek egymással nagyon jelentős átfedésben állhatnak. Lehetséges, hogy egyes lexikonelemek között a kapcsolat meglétéén túl szintaktikai szempontból nem hanyagolható el a hierarchikus viszony sem, ezt a gráf éleinek irányítása képes rögzíteni.

A mentális hálózatban elképzelhetőek multiplex viszonyok is, ilyenkor két elem között nem csupán egyfajta minőségben jön létre a kapcsolat; ez a gráfban az elemek között húzódó több éllel képezhető le. (Az egész ponthoz lásd KOVÁCS 2011: 20–90.)

2.3. A mentális lexikon szintaktikai szerepe a kognitív nyelvelmélet és a hálózatelmélet keresztmetszetében

A kognitív nyelvészet tehát a mentális lexikont fonológiai és szemantikai mikrojegyek, illetve belőlük az aktuális kommunikációs helyzetben profilálódó fogalmak olyan hálózataként definiálja, amely egy állandó dinamikus alakulásban levő, súlyozott, irányított gráfként képzelhető el. Az ilyen módon gráffal leképezhető, hálózatszerű szótárban a fogalmak (csúcspontok) között húzódó élek súlyozottsága és ezek esetleges irányítottsága mutatja meg egy adott pillanatban, egy adott kommunikációs helyzet figyelmi jelenete által előhívott szemantikai keretben az egyes elemek kapcsolódásának gyakorisági valószínűségét és a kapcsolódások hierarchikus viszonyát. Ezekkel a kapcsolódási valószínűségekkel és alá-fölé rendeltségi mintázatokkal jellemezhető az adott lexikont birtokló nyelvhasználó által produkált megnyilatkozások szintaxisa.

Az előbbiekből következik, hogy a kognitív szemlélet szerint nem léteznek deklaratív mondattani szabályok, csakis az egyes emberek mentális hálózatainak aktuális állapotait leíró és ezáltal a mondatformálási tendenciákat jellemző szintaktikai valószínűségi változók. (Az egész bekezdéshez lásd Tolcsvai Nagy 2013: 54–57.)

3. A dolgozat témájául szolgáló kutatás ismertetése

3.1. A dolgozat tárgya

Dolgozatomban azzal foglalkozom, hogy a deep learning szoftveres vizsgálat milyen utakat nyit meg egy, a lexikonelemek kapcsolódásának valószínűségi mintázatai alapján szerveződő, a kognitív nyelvészet elméleti megfontolásait szem előtt tartó mondatleírási

keretrendszer kidolgozására. Tekintettel a téma eddigi kezdetleges nyelvtudományos feldolgozottságára, illetve az ide vonatkozó nyelvészeti alapkutatási eredmények és szakirodalmak hiányára², dolgozatom hipotézisét csak kevésbé tudom specifikálni. Feltevésemet ebben a formában fogalmazhatom meg: a deep learning szoftveres vizsgálat új szempontokat ad a kognitív szintaxis elméleti alapjaihoz. Ezen hipotézissel kapcsolatban azzal az előzetes sejtéssel élek, hogy a vizsgálat alapján feltételezhető lesz egy nem teljesen ige-központú, nem elsődlegesen ige-főnév viszonyrendszerre építő, hanem a humán kogníció emergens jellegét specifikusabban figyelembe vevő mondattani szemlélet. A kognitív nyelvtanokban kidolgozott tételes mondattanok ugyanis nem szakadtak el a korábbi nyelvészeti irányzatok (a strukturalizmus, a transzformációs és a nem transzformációs generatív szintaxisok)³ azon elméleti alapállásától, hogy a mondatok szerveződésének elsődleges katalizátora az ige jelentés- (vagy argumentum) szerkezetében immanens módon meglevő üres helyek főnévi típusú elemek általi kitöltődése.⁴ Holott eddig kognitív pszichológiai kutatás nem bizonyította az igei szemantikának a mondatprodukciónban betöltött ilyen domináns szerepét. Óhatatlanul felvetődik tehát ettől különböző elméleti keret létjogosultsága is.

3.2. A dolgozat célja

Dolgozatom célja az előző pontban részletezett mondattani szemlélet kidolgozásához szükséges alapkutatások egyikének elvégzése: annak megvizsgálása, hogy egy deep learning szoftver által egy adott korpuszon végzett elemzés eredményeként kapott kapcsolódási valószínűségi mintázatok alkalmasak-e egy, a kognitív nyelvelméletre alapozó, nem ige-főnév fókuszú szintaktikai megközelítés megalkotásához. Tekintettel a hipotézis meglehetősen laza voltára, az általam összeállított szoftver egyelőre pilot jellegűre és az egyéb hardveres-szoftveres munkakörnyezet korlátozottságára, illetve az eredményekkel kapcsolatos

² A mesterségesintelligencia-kutatás szakirodalma természetesen rendkívül kiterjedt, magyar nyelven is elérhető több kutatócsoport tevékenysége révén (ELKH Nyelvtechnológiai és Alkalmazott Nyelvészeti Intézet, Mesterséges Intelligencia Nemzeti Laboratórium), azonban nem ismert számomra olyan munka, melynek fókuszában kifejezetten a kognitív nyelvelméletre alapozó szoftveres megoldások állnának.

³ Lásd É. KISS-SZABOLCSI 1995: 25–73. (kormányzás és kötés), KOMLÓSY 2001: 10–31. (lexikai-funkcionális nyelvten), TRÓN VIKTOR 2001: 9–73 (fejközpontú frázisstruktúra-nyelvten).

⁴ A kognitív szemléletű grammatikák ugyanis az alábbi alapelvekre építik mondattanukat: az ige jelentésszerkezetnek kettő összetevője (esemény és időszerkezet) a szótárban az igével együtt rögzítve van, míg figuratív szerkezetének kidolgozatlan elemeit (ÁGENS, PÁCIENS, BENEFICIENS, EXPERIENS stb.) főnévi jellegű elemek dolgozzák ki a mondat D1 dimenziójában megnevezendő jelenet tényezőinek megfelelően. Ebben a folyamatban, illetve a mondat további bővülésében szerepet játszó erődinamika-elmélet, agonista-antagonista kölcsönhatás is erőteljesen az ige-főnév szemantikai viszonyrendszerre épül. (A lábjegyzethez és az egész ponthoz lásd TALMY 2000: 406–469., TOLCSVAI NAGY 2015: 27–48., 2017: 391–436.)

ítéletalkotáshoz szükséges egzakt kritériumok szakirodalmi szegénységére, nem tűzhető ki célul a hipotézis és a hozzá kapcsolódó sejtés egyértelmű bizonyítása vagy cáfolata, csupán erős érvek felsorakoztatása a helytállóságuk mellett vagy ellen.

A vizsgálat elvégzésével ily módon szerzett információk várt relevanciája rávilágítani olyan összefüggésekre, melyek a kutatási folyamat következő szakaszainak hipotézisalkotásait, illetve módszertani döntéshozásait katalizálják. További, középtávú kutatási elképzelésem ugyanis a fentebb említett mondatleírási szemlélet megalkotására irányul, jelen dolgozatom tanulságaira alapozva, továbbfejlesztett öntanító és analitikai képességekkel rendelkező szoftveres (illetve lehetőség szerint nagyobb számítási teljesítménnyel bíró hardveres) bázissal, lényegesen kiterjedtebb korpusz elemzésével. Hosszú távon pedig kognitív pszichológiai (vagy akár neuropszichológiai) kísérletekkel tervezem megvizsgálni a szoftveres elemzéssel felismert szintaktikai mintázatoknak a kognitív nyelvelméleten és általában a kognitív tudományon belüli helytállóságát és relevanciáját.

Fontosnak tartom hangsúlyozni, hogy jelen dolgozat tehát a megismeréstudomány részeként felfogott elméleti nyelvészet keretei között, alapkutatás elvégzésének motivációjával jön létre, nem direkt nyelvtechnológiai vagy ipari invenció szándékával, azaz nem valamely mesterséges intelligencia magyar nyelvi képességfejlesztésének céljával. Hosszú távú kutatási elképzelésemtől mindazonáltal olyan eredményeket várok, melyek a kognitív szintaxis elméleti ismereteinek bővítése által a mesterséges intelligenciák nyelvi produkciós és percepciós kompetenciájának fejlesztésére alkalmasak, hozzájárulva az emberi kogníció működési mechanizmusait a jelenleginél még sikerebben adaptáló MI-k kifejlesztéséhez.

3.3. A dolgozat módszerei

A dolgozatom tárgyául szolgáló vizsgálatot egy magyar nyelvű, online forrású, írott nyelvi korpuszon végeztem el, egy szakirodalmi források és online tutorial tartalmak alapján általam összeállított deep learning szoftver segítségével.

Korpuszomat a SZTAKI HLT⁵ kötelékében, magyar nyelvű internetes oldalak szöveges anyagából felépített, 9 milliárd szavas Hungarian Webcorpus 2.0 (NEMESKEY 2020)⁶ első 109 darab textfájljából állítottam össze, így létrehozva egy nagyjából 800.000 soros, 12.500.000 szavas szöveges bázist. Ezt a korpuszt saját Google Drive meghajtómra töltöttem fel, hogy az online futtatókörnyezetben végrehajtott analízis során egyszerűen beolvasható legyen.

⁵ Számítástechnikai és Automatizálási Kutatóintézet, Human Language Technology

⁶ URL: https://nessie.ilab.sztaki.hu/~ndavid/Webcorpus2_text/

Az analízist végző deep learning szoftvert magam állítottam össze RAO–MCMAHAN 2019 programozási elveit végigkövetve, illetve a TensorFlow által közzétett online segédlet (THE TENSORFLOW AUTHORS 2019)⁷ kész moduljait és programozási tanácsait (THE TENSORFLOW AUTHORS (É.N.) ab)⁸ felhasználva. Az így létrejött szoftver a korpusz mondatait vizsgálva képes meghatározni az egyes elemek kapcsolódásának gyakorisági valószínűségét, illetve az egymáshoz képesti viszonyrendszer alapján felépíteni a korpusz szavait (teljes szóalakjait és nem lemmáit) tartalmazó hálózatos mentális szótárat. A szoftver kompetenciái tehát jelenleg addig terjednek, hogy a lexikonelemek kapcsolódási valószínűségi mintázatait képes felismerni, ám a lemmatizáláshoz és az esetleges alá-fölérendeltségi hierarchia kialakításához szükséges matematikai műveletek (konvolúciók) kódjainak beillesztése jelenleg is folyamatban van (így ez természetesen jelen dolgozatnak nem képezi részét).

A vizsgálat elvégzéséhez szándékosan nem támaszkodtam az angol, illetve magyar nyelven már rendelkezésre álló természetesnyelv-feldolgozó mesterséges intelligenciákra, illetve az ezek által analízis útján nyert kapcsolódási valószínűségi adatokra. Ennek motivációja az, hogy a program felépülésének állomásait lépésről-lépésre nyomon követve az ilyen jellegű szoftverek működését lehetővé tevő matematikai háttérrel és programozási elveket, metódusokat legalább alapszinten megismerjem, még akár az időbeli hatékonyság rovására is.⁹

4. A kutatáshoz használt szoftveres bázis ismertetése

4.1. Az NLP deep learning szoftverek általános ismertetése

A természetesnyelv-feldolgozást (*Natural Language Processing*) végző szoftverek olyan mélytanuló (*deep learning*) mesterséges intelligenciák, melyek a beérkező korpuszadatokat először megnyilatkozásokra, majd tokenekre osztják megadott szabályok vagy előre készen álló modulok segítségével, utána pedig ezeknek a kollókációit (egymáshoz viszonyított távolságát, *n*-gramjeit) és a kollokációk gyakoriságát számítják ki. Ezen információk alapján minden egyes tokenhez egy-egy vektort rendelnek hozzá, amely a tokennek a folyamatosan alakuló mentális hálózatban pillanatnyilag elfoglalt helyét jellemzi.

⁷https://colab.research.google.com/github/tensorflow/text/blob/master/docs/guide/word_embeddings.ipynb#scrollTo=Q6mJg1g3apaz

⁸ URL: https://www.tensorflow.org/api_docs/python/tf/keras/losses
https://www.tensorflow.org/api_docs/python/tf/keras/optimizers

⁹ Kutatásom későbbi szakaszaiban viszont tervezek majd támaszkodni a MILAB által fejlesztett HuSpaCy nevű, magyar nyelvű NLP szoftver támogatására.

Újabb adatok bevitelével természetesen számos elem kollokációs mintázatában változás áll be, így a hozzájuk tartozó vektorok értéke is módosul. Egy ilyen mesterséges intelligencia tehát nem előre megadott szabályokra támaszkodva, hanem a szövegtörzs feldolgozása során felismert mintázatok alapján alakítja ki a nyelvi kompetenciáját biztosító neurális hálózatot (lexikont); ezen öntanító képessége miatt nevezik *deep learning* (mélytanuló) szoftvernek. (Az egész bekezdéshez lásd RAO–MCMAHAN 2019: 29–37.)

Ha a hálózat egy adott állapotában a rendszert arra utasítjuk, hogy közölje a saját mesterséges mentális lexikonjának struktúráját jellemző vektorokat, akkor lényegében azokat a valószínűségi mintázatokat kapjuk, amelyek a dolgot tárgyaként részletezett mondattani szemlélet kidolgozásának alapjait adják.

4.2. A kutatáshoz használt deep learning szoftver elemei

4.2.1 Futtatási környezet és az alapkészségek moduljai

A kutatást elvégző szoftver összeállításának elméleti alapjait – ahogyan fentebb említettem – RAO–MCMAHAN 2019 alapján ismertem meg, míg a gyakorlati kivitelezéshez segédletként a TensorFlow fentebb hivatkozott anyagait alkalmaztam. Programom Python nyelven készült a Google Colab prémium GPU-val¹⁰ támogatott online futtatókörnyezetében¹¹.

A természetesnyelv-feldolgozó (NLP) mesterséges intelligenciák működésének alapvető feladata, hogy matematikailag kezelhetővé tegyék a tokenizált szöveges adatokat, amihez már előre készen álló matematikai bázist és szoftverbe építhető valószínűség-számítási blokkokat kínál a TensorFlow által rendelkezésre bocsátott keras modul. Ezért a szoftver alapjait ezen modul és a működését támogató egyéb kiegészítő modulok futtatásával szükséges lerakni, ezt követően a Google Drive-on tárolt textfájlok mappáját kell behívni a program futtatókörnyezetébe.

¹⁰ Az általam előfizetett verzió csak 100 futtatási egységre biztosított lehetőséget, ezért a kísérletezés jellemzően alacsonyabb teljesítményű, CPU-s futtatókörnyezetben történt, csak a végleges programverziót hajtotta végre GPU.

¹¹ Az alábbi linken elérhető a programot tartalmazó teljes Google Colab jegyzetfüzet, amit egy fantom email-címen keresztül tettem közzé, így a személyazonosságom kiderülésének veszélye nélkül megtekinthető:

https://colab.research.google.com/drive/1sJUjwdbvZnBVrZZYV_bmTfVjUoMnAp-4?usp=sharing

4.2.2. A korpusz előkészítése az elemzéshez

Az így beolvasott mappa tartalmának kicsomagolására és a benne található 109 fájl közvetlen elérésére a keras `get_file` blokkját használtam a megfelelő paraméterek beállításával, azaz jelezve, hogy semmilyen almappa nem található a tárhelyen fájlrendszerében, így a behívott tartalom egészében tekintendő a teljes feldolgozandó adathalmaznak, *dataset*nek. Az első részfolyamatok végén ennek köszönhetően a szoftver már felismerte az öntanulás alapjául szolgáló szöveges korpuszt. (Lásd THE TENSORFLOW AUTHORS 2019.)

A behívott szöveg mennyisége rendkívül nagy számítási kapacitást igényelne, ha a számítási folyamatoknak mind a 800.000 sort egyszerre kellene áttekintenie, ezért a programozás kezdeti szakaszának egy rendkívül lényeges lépése a korpusz felosztása, illetve a felosztást meghatározó számarányok optimális megválasztása. Egyáltalán nem indifferens ugyanis, hogy a program milyen ütemben, mekkora adathalmazon futtatja végig az elemző algoritmust. A helyes hiperparaméterek beállításához közelítő támpontot nyújtott a szakirodalom, ám ezek szinte kivétel nélkül korrigálásra szorultak a kísérleti futtatások során felszínre kerülő hibák és a rendkívül valószínűtlen eredmények miatt¹². (Lásd RAO–MCMAHAN 2019: 92–94.)

Számos kísérlet után az tűnt a leghatékonyabbnak, hogy a program egyszerre mindig csak egy 32 elemből álló kishalmazon (*batchen*) futtassa végig az elemző algoritmust, így tehát egy teljes tanítási „korszak” (*epoch*) annyi iterációból (*iteration*) fog állni, ahányszor a 32 megvan a teljes elemszámban. (Lásd THE TENSORFLOW AUTHORS 2019.)

Egy másik rendkívül kritikus pont a korpusz felosztása *training* és *validation* adathalmazokra: a szoftver az előbbin figyeli meg az elemek kapcsolódásának mintázatait és rögzíti az ezeket jellemző számokat vektorokban, majd az utóbbin teszteli, hogy az így számított vektorértékek mennyire pontosan jósolják meg az egyes elemek szintaktikai helyzetét egy másik, nem vektorizált szöveg mondataiban. Annak tekintetében pedig, hogy ezek a jóslatok mennyire voltak pontosak vagy pontatlanok, az szoftver az úgynevezett *loss* értékek segítségével pontosítja a *training* halmazon keletkezett vektorokat. Ezt a folyamatot minden *epoch* során végigjárja a rendszer, ez biztosítja a lexikon vektorainak lehető legpontosabb adaptálhatóságát más szövegekre, más kontextusokra. Jelen kutatásban ideálisnak tűnt a legszélesebb körben alkalmazott 0,8–0,2 arányú felosztás, ami egy 81 fájlból álló *training dataset*et és egy 21 fájlos *validation dataset*et eredményezett. Az egyes batchek kialakításának

¹² A szakirodalom ugyanis elsősorban professzionális felhasználók számára készült, akik nagyobb számítási teljesítményű hardverrel, egy-két (vagy akár több) nagyságrenddel nagyobb elemszámú korpuszon végeznek műveleteket.

paramétereit állandó értékkel biztosító seed esetén szintén éltem a tutorial által javasolt random, 123-as számmal. (Lásd THE TENSORFLOW AUTHORS 2019., RAO–MCMAHAN 2019: 131–133.)

A következő lépés annak biztosítása, hogy a számítások során a teljes dataset és a folyamatosan keletkező új adatok a megfelelő tárhelykihasználással és -struktúrával tárolódjanak a memóriában. Az erre kidolgozott, AUTOTUNE komponenst tartalmazó modulok kész megoldást adnak a problémakörre, így ezek megváltoztatása nem szükséges. (Lásd THE TENSORFLOW AUTHORS 2019.)

4.2.3. A szavak beágyazását jellemző vektorok tulajdonságai

A korpusz minden egyes szavához tartozik egy úgynevezett beágyazás (*embedding*), mely az adott elemnek a környezetében levő más elemekhez viszonyított pozícióját jellemzi egy vektorban/tenzorban rögzített tizedes törtek (*floating point numbers*) sorával. Az öntanulás megkezdésekor minden ilyen vektor/tenzor véletlenszerű számokat tartalmaz, ezek cserélődnek ki egyre pontosabb értékekre a tanulási folyamat során. Ezeknek a jellemzőknek a kialakításához is hiperparaméterek megadása szükséges, amelyeket a kísérletezésben tapasztalt sikeres alkalmazhatóságuk miatt meghagytam az online segédlet által javasolt értékeknek. Azaz a szoftver 1000 szavas különálló szótárakra osztja a teljes adathalmazt, majd az elemek kollokációs mintázatait 5x1-es vektorokban rögzíti (Lásd THE TENSORFLOW AUTHORS 2019.)

Mindezek után a kimeneti szótár méretének megadásával és a TextVectorization nevű, készen rendelkezésre álló modul alkalmazásával véghez kell vinni a szöveg első feldolgozását (*preprocessing*).

4.2.4. A szoftveres öntanulás működése és a vektorok kinyerése

Az előző pontban részletezett módon előkészített, egyszeresen már vektorizált adathalmaz készen áll az öntanulási folyamat megkezdésére, ami *training* hiperparamétereinek megadásával indítható el. Mivel a korpusz adattudományi mércével viszonylag kis terjedelműnek tekinthető, a beágyazási információkat hordozó vektorok alacsony dimenzionalitással működnek hatékonyan, jelen szoftverben ez az érték 5-re lett állítva. (Lásd THE TENSORFLOW AUTHORS 2019.)

Ezt követte az öntanulási folyamatot koordináló matematikai függvény (*optimizer*) és az egyes epochokban a korábbi adatok alapján visszacsatolást végző, korrigáló, finomhangoló függvény (*loss*) kiválasztása. Az *optimizer*ként beállított stochastic gradient descent (*SGD*) az alacsony tanulási rátája (*learning rate*-je) miatt jól alkalmazható kisebb adathalmazok trainingeléséhez, illetve a *loss*ként alkalmazott Poisson-eloszlás ideális mértékben szűri ki a

szélsőségesen alacsony relevanciájú, marginális értékeket. (Lásd RAO–MCMAHAN 2019: 39–48., THE TENSORFLOW AUTHORS (É.N.) a, b.)

Egy ekkora méretű adathalmaz feldolgozásához a szakirodalom közepes *epoch*-értéket javasol (a túlságosan feldolgozatlan, illetve a túláltalánosított eredményeket elkerülendő), így ehhez 30-at adtam meg. A 30 öntanulási periódus végeztével a deep learning szoftver megalkotta a korpusz szavainak kapcsolódási mintázatait jellemző valószínűségi vektorokat, amelyeket egy záró utasítással .tsv kiterjesztéssel ki lehet menteni a programból. (Lásd THE TENSORFLOW AUTHORS 2019., RAO–MCMAHAN 2019: 95–99.)

4.3. A deep learning szoftveres vizsgálat által felismert kapcsolódási mintázatok vizualizálása

Minthogy a programból kinyert vektorok és a hozzájuk tartozó metaadatok (a kimeneti mentális lexikon elemeinek nevei, címkéi) önmagukban nem alkalmasak az utólagos értelmezésre, nyelvészeti elemzésre (hiszen lényegében egy hosszú szám- és szósorról van szó), vizualizálni szükséges az általuk hordozott kapcsolódási valószínűségi mintázatok alapján szerveződő mentális lexikont. Ez az Embedding Projector nevű online felületen történt meg, ahol a teljes, 10.000 szavas kimeneti mentális hálózat megjelenítése lehetséges. Az ilyen módon 10.000 ponttal, háromdimenziós formában megjelenített rendszerben a kiírandó szomszédok számát minimumra (5-re) állítva, az egyes pontokra kattintva egyszerűen megállapítható, hogy egy bizonyos pont a rendszer ítélete szerint mely másik pontokkal kerül egy mondaton belül szomszédságba a legnagyobb valószínűséggel.¹³

5. A deep learning szoftver által szolgáltatott eredmények

5.1. A deep learning szoftver által megalkotott lexikon elemei a hipotézis fényében

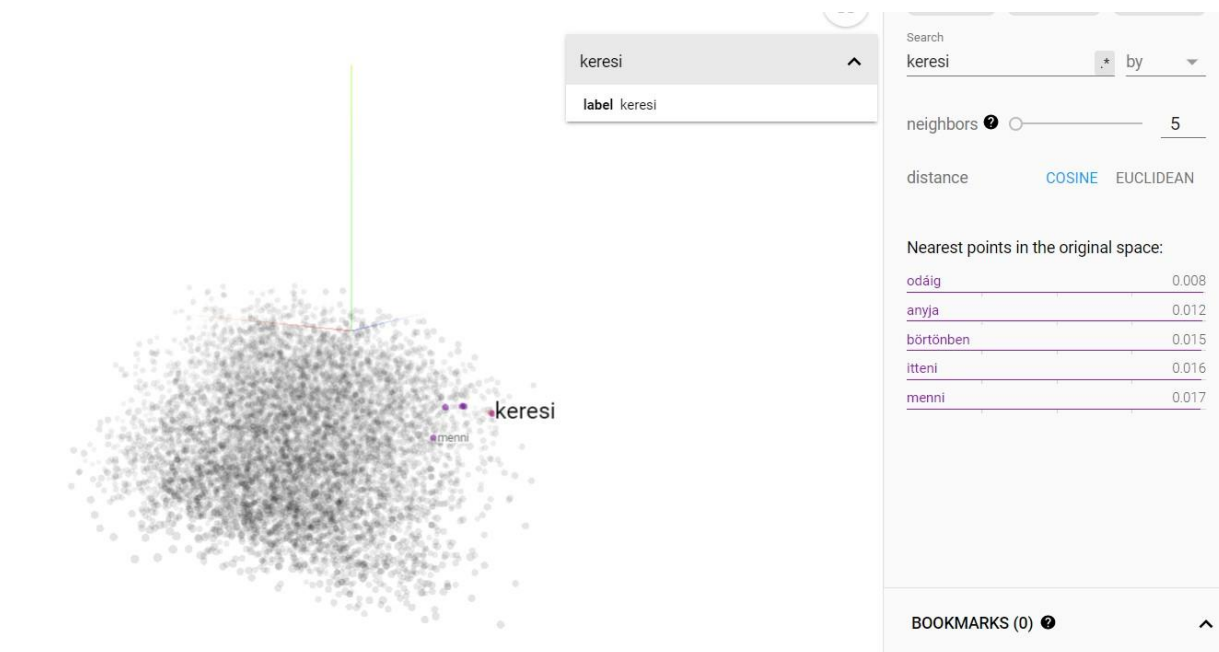
Az Embedding Projector segítségével vizualizált mesterséges mentális lexikon eszközt szolgáltató a dolgozat tárgyaként megadott hipotézis és sejtés helytállóságának megvizsgálásához. Ennek kapcsán természetesen figyelembe kell venni az eredmények korlátozott valenciáját, azaz az egyelőre még pilot, kísérleti jellegű szoftver kimeneti adatait nem lehet szigorúan vett bizonyításnak tekinteni, csupán érvként kezelni a bennük felismerhető

¹³ A vizualizált hálózat reprodukálhatóságának érdekében az alábbi Google Drive linken elérhetővé tettem a szoftver által nyert vektorokat:
<https://drive.google.com/drive/folders/1ZBfHLx-bJGo0KgaxQ0JbwlJBkx2j4cBV?usp=sharing>
A linken található két fájlban a <https://projector.tensorflow.org/> oldalon a Load-ra kattintást követő beolvasásával megjeleníthetővé válik a dolgozatban hivatkozott mentális lexikon.

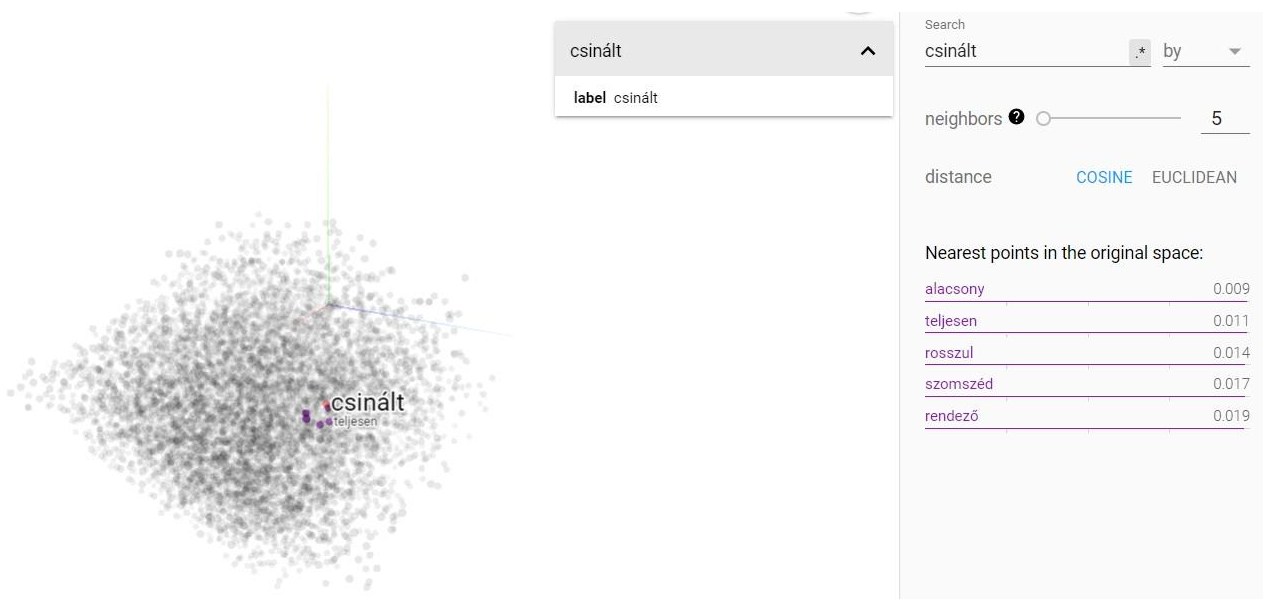
tendenciákat. Az sem hanyagolható el, hogy a szoftver a nagyjából 12.500.000-es bemeneti elemszámból hozott létre egy 10.000 szavas szótárt, vagyis a kapcsolódási valószínűségek az esetek legnagyobb részében nem a konkrét elemekre, hanem az elemek nagyobb szemantikai kategóriájára vonatkoznak – például arra, hogy egy adott ige mellett leggyakrabban milyen főnév áll, de nem feltétlenül az aktuálisan megnevezett főnévre. Illetve azzal is számolni kell, hogy jelen dolgozat keretei nem teszik lehetővé mind a 10.000 pont teljeskörű elemzésének ismertetését, hanem csupán néhány jellegzetes adat mutatóváltó közlését és a hipotézis fényében történő elemzését. A kutatási folyamatban természetesen lényegesen több elem kapcsolódási mintázatát megvizsgáltam, azonban a részletes közlésben csupán a legjellemzőbb csoportok egy-egy tagjára szorítkozik dolgozatom. Konklúzióimat viszont természetesen az összes értékelt elem tulajdonságaira alapozom, így ezek helytállósága a mentális hálózat 13. lábjegyzetben közölt módon történő újbóli vizualizálásával bármikor ellenőrizhető.

Elsőként igeik esetében vizsgálom meg, hogy a szoftver által számított kapcsolódási valószínűségek mennyiben tükrözik a kognitív szintaxis ismeretében velük kapcsolatban felmerülő preszuppozíciókat.

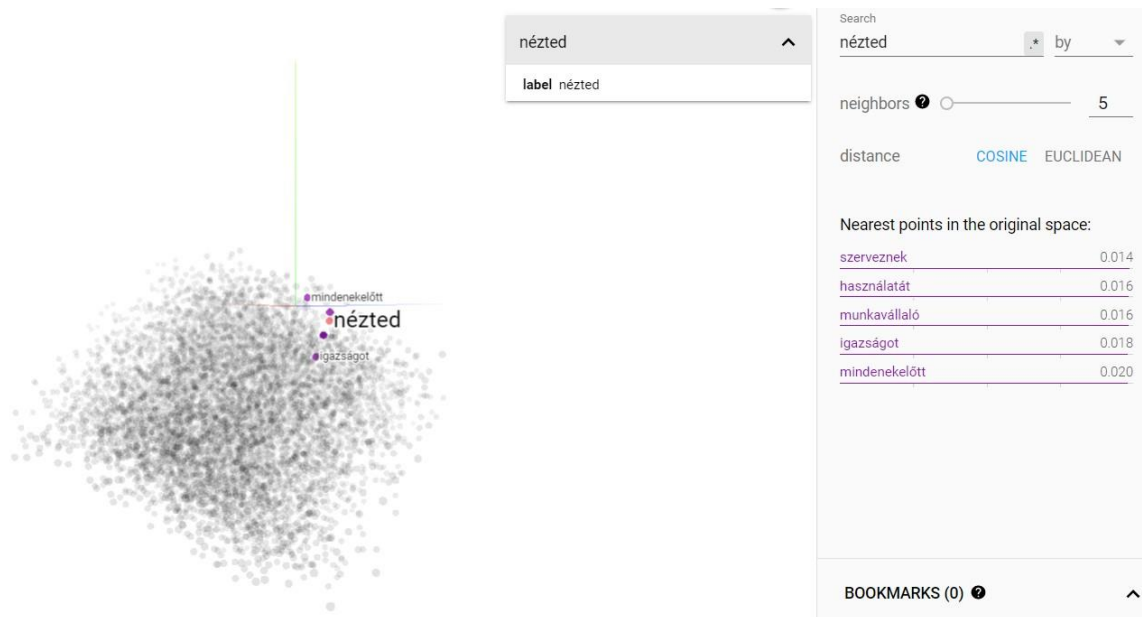
A keresi lehorgonyzott igealak estében a figuratív szerkezet kidolgozásához elsősorban egy ÁGENST és egy PÁCIENST várnánk leggyakrabban kapcsolódó elemekként, a program analízise szerint azonban PÁCIENS nincs az öt legvalószínűbb szomszédos szó mellett, viszont található az ÁGENS mellett két elem, ami valamilyen helyen levéssel kapcsolatos (*börtönben, itteni*). Ezekben az esetekben feltehetőleg az ige inherens jelentésszerkezetéhez prototipikusan nem kapcsolódó, térbeli viszonyokkal kapcsolatos tényezők kifejtése történik meg (például: *keresi a börtönben, keresi az itteni piacon*). A kognitív szemléletű szintaxis ezt a jelenséget térbeli referenciakeretként kezeli (TOLCSVAI NAGY 2021: 135–138.), ugyanakkor ezzel kapcsolatban fontos, tisztázandó kérdés, hogy az ilyen (rendkívül gyakori) esetekben az ige pontosan hogyan, milyen szemantikai szerkezettel (a térbeli referenciakerettel milyen viszonyban) tárolódik a mentális lexikonban.



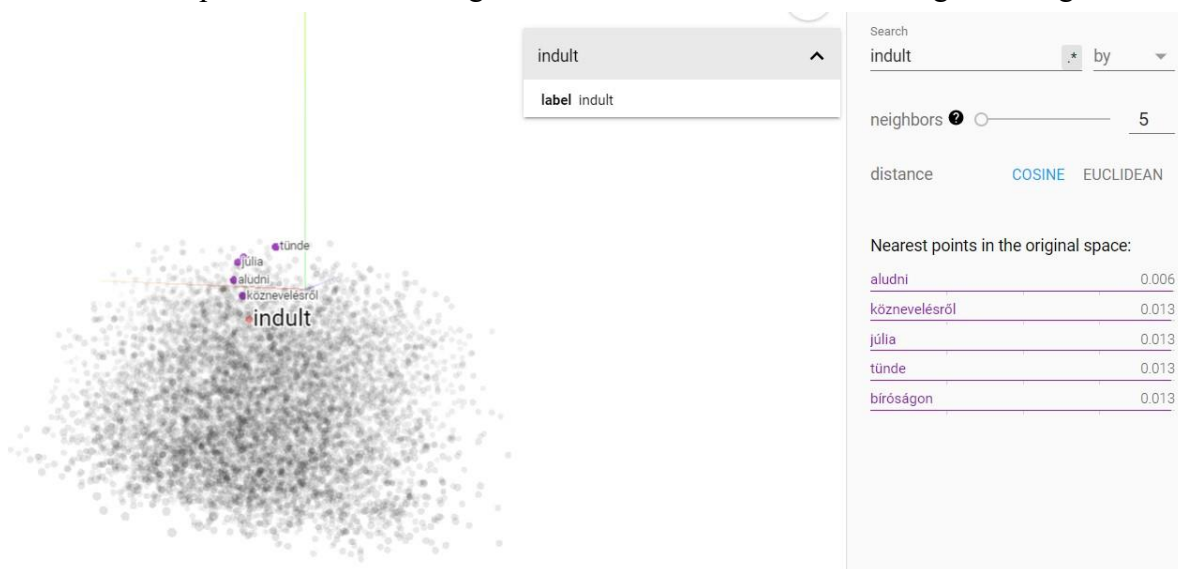
Egy másik, elméletben hasonló figuratív szerkezetet implikáló lehorgonyzott igealak, a *csinált* esetében is hiányzik a legvalószínűbb öt kapcsolatból a PÁCIENS szerepét betölteni képes elem, miközben (két ÁGENSI szerepű szó mellett) olyan szavak is nagy valószínűséggel alkotnak vele egy jelenetet, amik nem illeszthetők be az *csinál* ige jelentésszerkezetébe: a *teljesen* és a *rosszul*. Ez utóbbi kettő sokkal inkább az ige által hordozott folyamat, illetve mondatban kidolgozott jelenet más, immanens figuratív szerepkörrel nem jellemezhető tényezőjét dolgozza ki (például: *valamit rosszul csinált*). Szükséges annak a feltérképezése, hogy az ilyen jellegű, leginkább módbelinek nevezhető tényezők milyen szerepet töltenek be az igékhez kapcsolódva a szótárban.



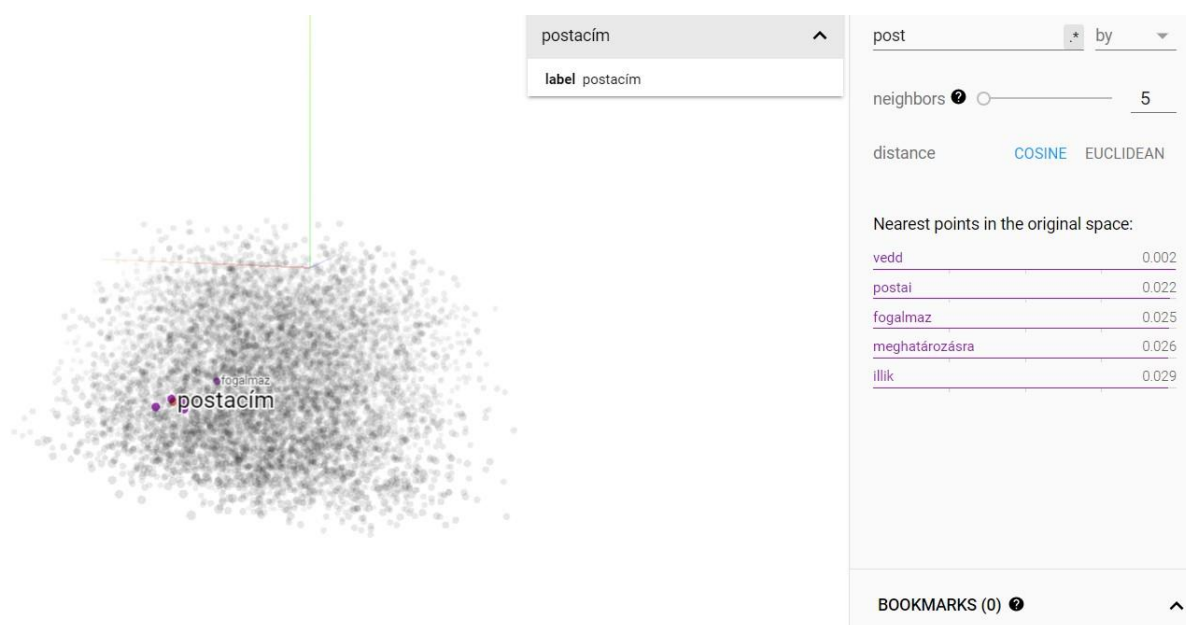
Különösen érdekes a *nézted* igealakhoz kapcsolódó elemek között első helyen felbukkanó másik igealak is (*szerveznek*), ez talán összefüggésben állhat azzal, hogy a *néz* ige gyakran dolgozza ki alárendelő összetett mondat főmondatának jelenetét (például: *Nézted, ahogy valamit szerveznek.*), így akár az is elképzelhető, hogy a mentális lexikonban is rögzül ez, mint szemantikai tulajdonság. Ebben az esetben egyébként prototipikusnak tekinthető módon PÁCIENSEk (*használatát, igazságot*) is helyet foglalnak az első öt leggyakoribb kapcsolat között.



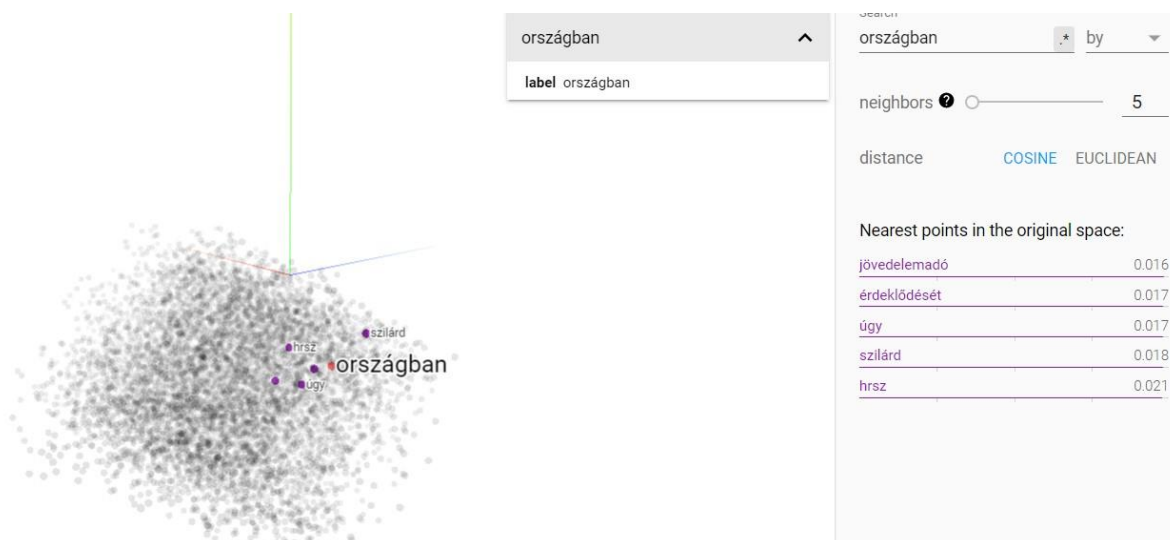
Arra is akad példa, hogy egy ige jelentésszerkezete több különböző folyamat tényezőinek megfelelően többféleképpen is szerveződik. Ilyen többek között az *indult* esete, ami a szoftveres vizsgálat során kapott gyakorisági adatok alapján a leggyakoribb eseteknek is csak egy részében jeleníti meg a prototipikusan hozzá társított folyamatot (például: *Tünde indult valahova, Júlia indult aludni*), míg képes más lefolyású jelenetek kidolgozására is egy mondatban, ráadásul az értékek szerint nagyon gyakran (például: *vita indult a köznevelésről, eljárás indult a bíróságon*). Az ilyen poliszémikus igék (TOLCSVAI NAGY 2021: 113–121.) szótárbeli és szintaktikai szerepének közelebbi vizsgálatára is alkalmas lehet a mesterséges intelligencia.



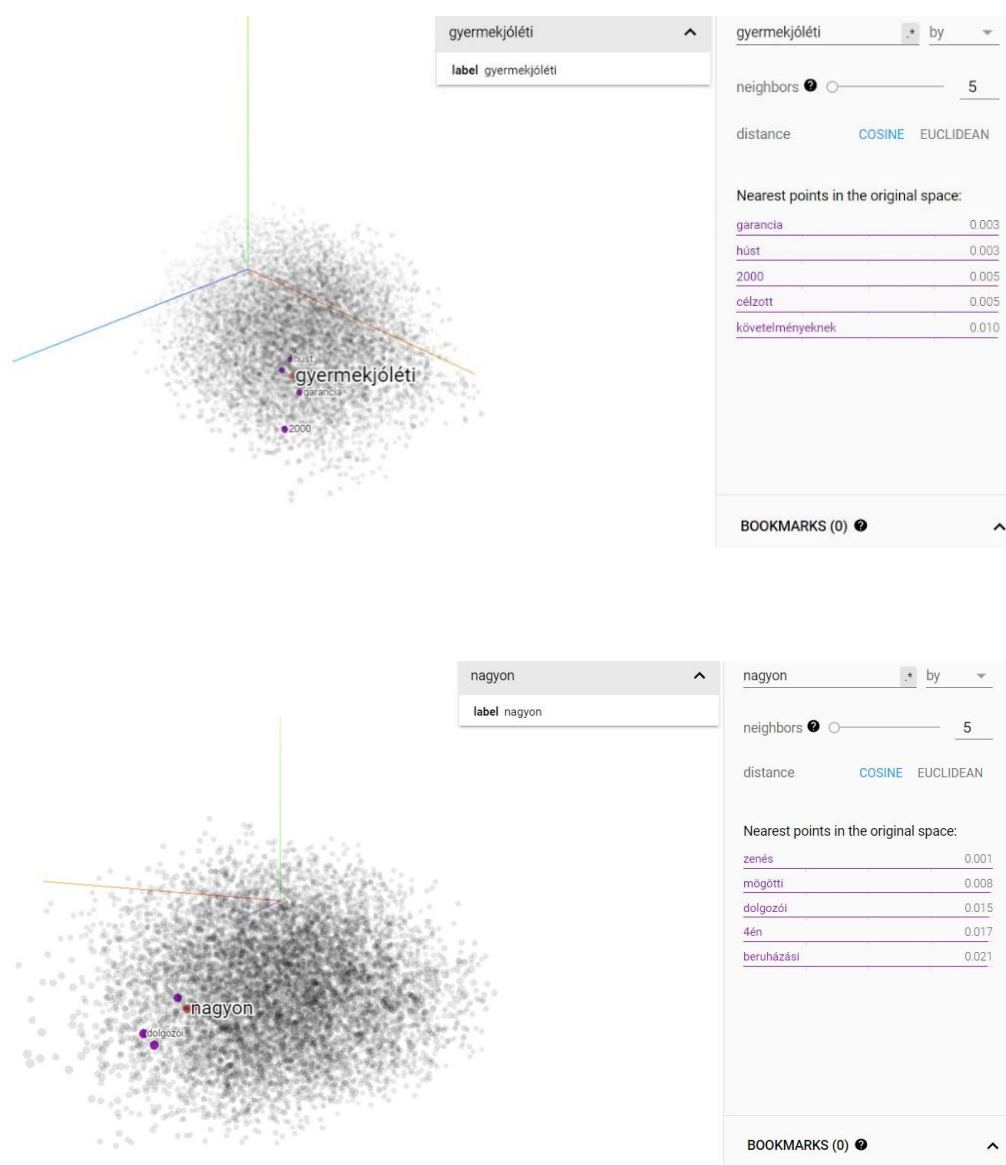
A főnevek esetében is megfigyelhető olyan eset, amikor az adott szó nem csupán a tőle prototipikusan várt szerepekben kapcsolódik más elemekhez. A *postacím* szóalaktól ugyanis leginkább egy ÁGENS vagy TÉMA szerepének kidolgozását várnánk, ámde ez csak a negyedik és ötödik leggyakoribb esetben fordul elő (*meghatározásra, illik*; például: *a postacím meghatározásra kerül, a postacím illik a valósághoz*), a gyakoriság szerinti második és harmadik helyen álló esetekben ugyanis olyan együtt előfordulásokról van szó, amit a különböző nyelvtanok hagyományosan nem is a mondattan tárgykörében értelmeznek. A *postai* és *fogalmaz* szavak ugyanis a POSTA szemantikai keretében, illetve a LEVÉLÍRÁS forгатókönyvében szerepelnek a *postacímmel* együtt, miközben egymásnak semmilyen szemantikai alszerkezetét nem dolgozzák ki.

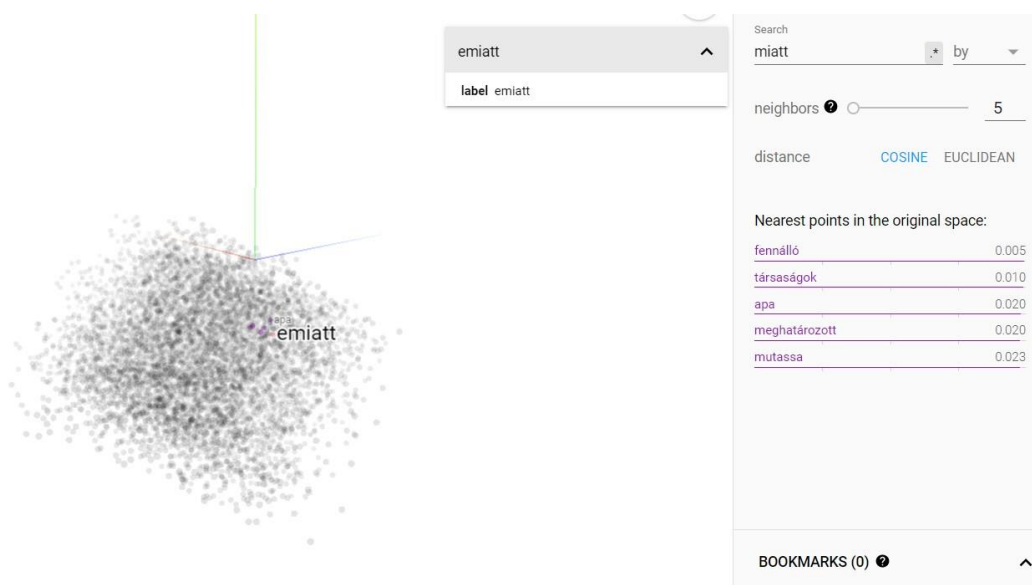


Az előzőhöz hasonló módon az *országban* szóalakhoz sem kapcsolódik a leggyakoribb öt helyen ige, hogy főnévként az ige által kifejtett jelenet LOKÁLIS összetevőjét dolgozza ki, hanem az ORSZÁG, ÁLLAM szemantikai keretéhez nagy valószínűséggel tartozó kifejezésekkel áll együtt a legtöbbször (*jövedelemadó, szilárd, hrsz*).



Ha pedig távolabb lépünk az ige-főnév kompozitmuszerkezet szemantikája által meghatározott mondatszerveződési tendenciáktól, és más szemantikai tartalommal bíró lexikonelemek kapcsolódási mintáit vizsgáljuk, akkor az látszik, hogy a legvalószínűbb kapcsolatok (*gyermekjóléti: garancia, húst, 2000, célzott, követelményeknek; nagyon: zenés, mögötti, dolgozói, 4-én, beruházási; emiatt: fennálló, társaságok, apa, meghatározott, mutassa*) létrejöttében is a hétköznapi nyelvi viselkedésünk számára értelmezhető és elfogadható mintázatok játszanak szerepet. Ezek a mintázatok ugyanakkor nem írhatók le a kognitív szintaxis ige-főnév fókuszú, erődinamika-elvű mondatszerveződési elveivel, így jellemzésük megkívánja egy másfajta szintaktikai szemlélet kidolgozását. A kutatási folyamat későbbi szakaszaiban tehát elsősorban ezeknek az kognitív szintaxisba még be nem épített elemeknek a lexikonbeli viszonyait és azok alapján a mondatszervezésben betöltött (sejtésem szerint jelentős) szerepét szükséges megvizsgálni.





5.2. Konklúziók

A programból kinyert vektorok vizualizációt követő elemzése, vizsgálata a mentális lexikon szerveződésének olyan mintázataira mutatott rá, melyek erős érvnek tekinthetők dolgozatomban hipotézisének helytállósága mellett. Azaz a szemléltetett példák alapján extrapolálva erősen valószínűsíthető, hogy a deep learning szoftveres vizsgálat valóban új szempontokat ad a kognitív szintaxis elméleti alapjaihoz. Ezen új szempontokat az alábbi területeken valószínűsítem: az igékhez kapcsolódó, nem immanens jelentésszerkezeti tényezőket kidolgozó elemek kapcsolódási mintázatai; a több lehetséges kompozitumszerkezettel rendelkező igék lexikonbeli–szintaktikai működése; a nagy alárendelési potenciált vélhetően a szemantikájukban hordozó igék szintaktikai szerepe; a szemantikai keretek és forgatókönyvek szerepe a lexikonban és a szintaxisban; illetve az ige–főnév szemantikai viszonyrendszerébe nem illeszkedő, rendkívül nagyszámú lexikonelem szerepe a szintaxisban.

Mindezek fényében nagyon meggyőzőnek tűnik, hogy a deep learning szoftverrel végzett analízis hatékony módszer lehet egy, a lexikonelemek kapcsolódásának valószínűségi mintázatai alapján szerveződő, a kognitív nyelvészet elméleti megfontolásait szem előtt, a nem igei–főnévi elemek szerepét jelentősebben inkorporáló mondatleírási keretrendszer kidolgozására.

6. Kitekintés, további lépések

Az előző bekezdésben leírtak alapján dolgozat célkitűzéseiben megfogalmazott kutatási elképzelések a továbbiakban folytathatóknak bizonyulnak. A középtávra meghatározott célok felé a következő lépés a jelen dolgozat eredményeképp létrejött elméleti sejtések megfogalmazása hipotézisként, majd azok bizonyításának megkísérlése a deep learning szoftver továbbfejlesztett verziójával. A szoftveres bázis produktívabb működéséhez először is nagyon szükséges a szóalakok belső struktúrájába is belátni tudó, lemmatizációs kompetenciával rendelkező és a lexikonelemek kapcsolódásának immár nem csupán valószínűségi tendenciáit, hanem hierarchikus viszonyait is feltérképezni képes modulok beépítéséhez. Ezekről már nem csupán erős érveket, hanem bizonyítást is várok majd leendő hipotéziseimhez. A kognitív szemléletnek egy további rendkívül fontos eleme a kontextus, azaz a beszédhelyzet aktuális figyelmi jelenetének a kimagasló szemantikai–szintaktikai szerepe. Így terveim között szerepel kísérletet tenni olyan programrészlet összeállítására is, mely képes a kontextus szerepéről is információkat szolgáltatni.

Az így létrejött eredményeket pedig olyan formában lesz szükséges rendszerezni és interpretálni, hogy az a kognitív nyelvészet számára szintaktikai szemléletként hasznosítható legyen, illetve a kognitív pszichológiai vizsgálatok és a mesterségesintelligencia-kutatás számára is gyakorlati relevanciával bírjon. Ehhez a továbbiakban még több tudományterület között szükséges tájékozódni és párbeszédet teremteni.

Felhasznált irodalom

- É. KISS KATALIN–SZABOLCSI ANNA 1995. Grammatikaelméleti bevezető. In: KIEFER FERENC (szerk.): *Strukturális magyar nyelvtan 1. Mondattan*. Budapest, Akadémiai Kiadó. 25–73.
- KOMLÓSY ANDRÁS 2001. *A lexikai-funkcionális grammatika mondattanának alapfogalmai. Segédkönyvek a nyelvészet tanulmányozásához 7*. Budapest, Tinta Könyvkiadó. 10–37.
- LANGACKER, RONALD W. 2008. *Cognitive Grammar. A basic introduction*. Oxford, Oxford University Press. 183–214.
- KOVÁCS LÁSZLÓ 2011. *Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban. Segédkönyvek a nyelvészet tanulmányozásához 119*. Budapest, Tinta Könyvkiadó. 7–90.
- RAO, DELIP–MCMAHAN, BRIAN 2019. *Natural Language Processing with PyTorch*. Sebastopol CA, O'Reilly.
- TALMY, LEONARD 2000. *Toward a Cognitive Semantics. Volume I. Concept Structuring Systems*. Cambridge, The MIT Press. 406–469.
- TOLCSVAI NAGY GÁBOR 2013. *Bevezetés a kognitív nyelvészetbe*. Budapest, Osiris Kiadó. 13–168.
- TOLCSVAI NAGY GÁBOR 2015. *Az ige a magyar nyelvben. Funkcionális elemzés. Segédkönyvek a nyelvészet tanulmányozásához 181*. Budapest, Tinta Könyvkiadó. 27–68., 112–123.
- TOLCSVAI NAGY GÁBOR 2017. *Nyelvtan*. Budapest, Osiris Kiadó. 23–67., 207–234., 310–352., 391–462.
- TOLCSVAI NAGY GÁBOR 2021. *Kognitív szemantika. Diagram Könyvek 4*. Budapest, ELTE Eötvös Kiadó. 26–93., 126–153.
- TRÓN VIKTOR 2001. *Fejközpontú frázisstruktúra-nyelvtan. Segédkönyvek a nyelvészet tanulmányozásához 9*. Budapest, Tinta Könyvkiadó. 9–73.

Online források a programozáshoz

- NEMESKEY, DÁVID MÁRK 2020. *Natural Language Processing methods for Language Modeling. PhD thesis*. Eötvös Loránd University. URL: https://nessie.ilab.sztaki.hu/~ndavid/Webcorpus2_text/ (utolsó elérés: 2023. 01. 11.)
- THE TENSORFLOW AUTHORS (É.N.) a. *Module: tf.keras.losses*. URL: https://www.tensorflow.org/api_docs/python/tf/keras/losses (utolsó elérés: 2023. 01. 11.)
- THE TENSORFLOW AUTHORS (É.N.) b. *Module: tf.keras.optimizers*. URL: https://www.tensorflow.org/api_docs/python/tf/keras/optimizers (utolsó elérés: 2023. 01. 11.)
- THE TENSORFLOW AUTHORS 2019. *Word embeddings*. URL: https://colab.research.google.com/github/tensorflow/text/blob/master/docs/guide/word_embeddings.ipynb (utolsó elérés: 2023. 01. 11.)