Kornai András

Vektorszemantika

Kornai András

Vektorszemantika

TYPOTEX

A kiadvány megjelenését az Innovációs és Technológiai Minisztérium Nemzeti Kutatási, Fejlesztési és Innovációs Alapból nyújtott támogatásával a Mecenatúra 2021 pályázati program finanszírozásában megvalósuló MEC_K 141539 számú projekt tette lehetővé.

A fordítás a következő kiadás alapján készült:
András Kornai: *Vector Semantics.* Springer International Publishing, 2023

ISBN 978 963 493 309 0

Kedves Olvasó!
Köszönjük, hogy kínálatunkból választott olvasnivalót!
Újabb kiadványainkról és akcióinkról a `www.typotex.hu`
és a `facebook.com/typotexkiado` oldalakon értesülhet.

Typotex Kiadó
Alapította Votisky Zsuzsa, 1989
A kiadó az 1795-ben alapított Magyar Könyvkiadók
és Könyvterjesztők Egyesülésének tagja.

Felelős kiadó: Németh Kinga
Felelős szerkesztő: Balázs Péter
Szerkesztette: Erő Zsuzsa
Tördelés: Kornai András
A borítót készítette: Szalay Éva

Áginak

Semmi sincs olyan gyakorlatias, mint egy jó elmélet (Lewin, 1943)

A matematika annak a művészete, hogy minden problémát a lineáris algebrára vezessünk vissza (William Stein, idézve Kapitula (2015)-ben)

Az algebra az ördög ajánlata a matematikusnak. Az ördög így szól: Neked adom ezt a csodálatos gépezetet, minden kérdésedet meg fogja válaszolni. Csak annyit kell tenned, hogy nekem adod a lelkedet. Add fel a geometriát, és máris a tiéd ez a csodamasina (Atiyah, 2001)

Előszó

Ez a könyv (Kornai, 2019) közvetlen folytatása, de elődjétől eltérően már nem tankönyv. A korábbi kötet, a továbbiakban S19, többnyire olyan anyagot fedett, ami jól ismert a területen, a jelenlegi kötet ezzel szemben kutatási monográfia, amelyben a szerző saját, a `4lang` rendszerre összpontosító kutatása dominál. S19 négy tudományág diákjait próbálta kiszolgálni: nyelvészet, számítástechnika, kognitív tudomány és filozófia. Heinrich Schütze akkor ezt írta: „Ez a tankönyv interdiszciplinaritása révén különbözik a többi szemantikakönyvtől: bemutatja a nyelvészet, a számítástechnika, a filozófia és a kognitív tudomány nézőpontjait. Az elkövetkező években nagy változások várhatók ezen a területen, ezért az alapok széles körének lefedése a helyes megközelítése annak, hogy a diákokat ellássuk a szükséges ismeretekkel ahhoz, hogy most és a jövőben szemantikával foglalkozzanak."

A nagy változások valójában már folyamatban voltak, nem kis részben Schütze, 1993-nak köszönhetően, aki megtette az alapvető lépést a szójelentéseknek a közönséges euklideszi tér vektoraival való modellezésében. S19:2.7. részben tárgyalja ennek matematikai alapjait. Ez az anyag ma már sztenderd, olyannyira, hogy a természetes nyelvfeldolgozás (natural language processing, NLP) alapvető tankönyve, Jurafsky és Martin (2022) már beépíti új kiadásába (erre az új változatra fogunk hivatkozni). A vektoros szemantikának azonban egyelőre viszonylag kevés érintkezési pontja van a mainstream nyelvi szemantikával, olyan kevés, hogy a legátfogóbb (ötkötetes) kortárs összefoglaló, Gutzmann és tsai. (2021), egyetlen fejezetet sem szentelt a témának. Hatvan évvel ezelőtt McCarthy (1963) így írt:

> A matematikai nyelvészek komoly hibát követnek el, amikor a szintaxisra és még szűkebben a természetes nyelvek nyelvtanára összpontosítanak. Ennél még fontosabb lenne a matematikai megértésnek és a természetes nyelven közvetített különféle információk formalizálásának a fejlesztése.

Mi pedig folytatjuk az eredeti tervet olymódon, hogy nemcsak szóvektorokat – statikusokat és kontextuálisokat egyaránt – próbálunk használni, hanem a lineáris és a multilineáris algebra bővebb gépezetét is olyan jelentésreprezentációk leírására, amelyeknek értelme van mind a nyelvészek, mind az informatikusok számára. Ennek során újra

fogjuk értékelni magukat a szóvektorokat is, amellett érvelve, hogy a szavak a legtöbb esetben nem vektoroknak felelnek meg, hanem az n dimenziós tér politópjainak, és új modelleket fogunk nyújtani a nyelvi szemantika számos hagyományos problémájához a preszuppozícióktól az indexikusokig, a merev jelölőktől a változók kötéséig. Kornai, 2007-ben ezt írtuk:

> A matematikai nyelvészet talán leginkább lenyűgöző aspektusa nem csupán az, hogy léteznek diszkrét mezoszkópikus struktúrák, hanem az a tény, hogy ezek – olyan módokon, amiket nem teljesen értünk – beágyazódnak folytonos jelekbe.

A vektorszemantika pedig erényt csinál a szükségből: akár teljesen értjük a dolgot, akár nem, amikor beágyazunk nyilvánvalóan diszkrét szavakat a folytonos euklidészi térbe, belső szerveződésük egy lényeges jellemzőjéről adunk számot. Mindeközben a beszédfelismerésben is hasonló változások mennek végbe, lásd pl. Bohnstingl és tsai., 2021. Itt nyilvánvalóan nem tudjuk részletekbe menően tárgyalni a beszédet, de egyértelműnek tűnik, hogy a neurális modellezés korai céljai, amelyeket akkoriban nem tett elérhetővé a gépek számítási teljesítménye, végre látótérbe kerülnek. A közelmúltbeli áttérés a *dinamikus* vagy a *kontextuális* beágyazásokra, amik mára a számítógépes nyelvészetben (computational linguistics, CL) és az NLP-ben beépült sztenderdek, megválaszolatlanul hagynak egy kulcskérdést, a kompozicionalitást (S19:1.1.): hogyan reprezentáljuk a nagyobb kifejezések jelentését. A kérdés fontosságát korán felismerte Allauzen és tsai., 2013, de mindeddig egyetlen javasolt megoldás (mint például Purver és tsai., 2021) sem nyert széleskörű elfogadást. Valójában a CL/NLP közösség a kérdést nagyrészt szem elől vesztette, annak a hatására, amit Noah Smith „az e2e vallásra való áttérésnek és a differenciálhatóság kultuszának” nevezett (lásd LeCun, Bengio és Hinton, 2015 és Goldberg, 2017 az elejétől végéig differenciálhatóság paradigmájának világos összefoglalásához).

Egyelőre áthidalhatatlannak tűnik a szakadék a nyelvészek és a számítógépes nyelvészek között: az előbbiek nagy súlyt helyeznek a köztes struktúrákra a morfémától a bekezdésig és azon túl, az utóbbiak pedig egyre inkább az „elejétől a végéig” (end-to-end, e2e) rendszereket részesítik előnyben, amelyek határozottan nem támaszkodnak közbülső egységekre vagy struktúrákra, még a lexikon alapvető hasonlósági struktúrájára sem, amit a statikus szóvektorok tettek láthatóvá. Mégis egyértelműnek tűnik, hogy mindkét fél ugyanazt akarja, a nyelvi viselkedés tanulásra képes modelljeit, és a különbség stratégia kérdése: a nyelvészek olyan magyarázható, moduláris rendszereket keresnek, amelyeknek tanulási képessége menet közben tanulmányozható, míg a számítógépes nyelvészek ragaszkodnak az azonnal tanulásra képes modellekhez, akár olyan problémák árán is, mint az egy példa alapján és nulla példa alapján tanulás, amelyek inkább központi helyet foglalnak el az elméleti nyelvészetben, ahol a jelenség produktivitás néven ismert. Ezenkívül a CL/NLP teljesen elégedett a sokmilliárd szavas tanító korpuszok használatával, míg a nyelvészek olyan algoritmust szeretnének, amely érzékeny az elsődleges nyelvi adatokra; ezek valószínűleg összesen sem haladják meg a néhány millió szót.

Ebben a könyvben megpróbálunk mindkét oldalnak megfelelni, olymódon, hogy (i) használunk köztes reprezentációkat, és (ii) nyújtunk ezekhez tanulási algoritmusokat. Az

1. fejezetben annak a formális rendszernek a meghatározásával kezdjük, amellyel a szavakhoz szimbolikus technikák segítségével jelentést fogunk rendelni. Mint Gérard Huet annak idején megjegyezte, S19:4.,5.8.-ban „az Eilenberg-gépek elegáns formalizmusát” használtuk „a szalagokkal és olvasófejekkel működő, vacak tákolmányok” helyett. De ezzel valójában csak későbbre halasztottuk a tanulásra való képességet, főleg ha azt nézzük (Angluin, 1981; Angluin, 1987), hogy a véges állapotú (FS) eszközök megtanulása egyáltalán nem triviális. Az FS tanulhatósággal kapcsolatos munka frontvonala most a fonológia (Rogers és tsai., 2013; Yli-Jyrä, 2015; Chandlee és Jardine, 2019; Rawski és Dolatian, 2020), ahol az adatok lényeges időbeli szerkezettel bírnak. Még nem tudjuk, ebből mennyit lehet átvinni a szemantikába, ahol a memória jellemzően véletlen hozzáférésű (lásd 7.4.) és az időbeli szerkezet, a szavak sorrendje jórészt irreleváns lehet (a szabad szórendű nyelvekben). Ezért a jelenlegi kötet főiránya az, hogy a nyelvi szemantika elméletét folytonos vektorterekkel, ne pedig Eilenberg-gépekkel kapcsoljuk össze, de eközben igyekszünk minél többet megőrizni a relációs gondolkodás eleganciájából.

Megközelítésünk formális, és kifejezett célja a számítógépes nyelvészek számára hasznos formalizmus kialakítása. Mégis nagyon sokat köszönhet egy határozottan informális elméletnek, a kognitív nyelvészetnek. Kötetünket valójában *Formális lexikai szemantiká*nak is lehetne nevezni, ha a megrögzült terminológia nem tekintené a *formális*t és a *lexikai*t egyenesen ellentétnek. A „kognitív” munkák (Jackendoff, 1972; Jackendoff, 1983; Jackendoff, 1990; Lakoff, 1987; Langacker, 1987; Talmy, 2000) hatása végig látható lesz. Ezen kognitív elméletek jó részének prezentációja informális (valójában a kognitív nyelvtan legtöbb képviselője, Jackendoff figyelemre méltó kivételével, kifejezetten antiformalista), mások pedig, mind a mesterséges intelligenciában (artificial intelligence, AI, MI), mind a tulajdonképpeni kognitív tudományban hallgatnak a szójelentésről; Fodor (1998) egészen világosan kimondja, hogy a szavak atomiak. A 2. fejezetben bemutatunk egy nem kompozicionális formális elméletet, amely alkalmazható a morfológiára, azaz a klitikumok és a kötött toldalékok szemantikájának leírására is, továbbá természetes módon kiterjeszthető a kompozicionális tartományra. Az, hogy valami ilyesmire valóban szükség van, nyilvánvaló, ha több nyelvet tekintünk, mivel ugyanaz a jelentés, amit az egyik nyelvben a morfológia fejez ki, egy másikban gyakran szintaktikai eszközökben fejeződik ki.

Kurt Lewin nevezetes mondása szerint „semmi sincs olyan gyakorlatias, mint egy jó elmélet”. Ezt a tézist azzal illusztráljuk, hogy bemutatjuk a kognitív nyelvtan nagy részének egy teljesen formális rekonstrukcióját, habár a Jackendoff által előnyben részesített generatív masinéria helyett algebrai köntösben. Felvállaltunk egy sor kényes kérdést, így például az időbeliség és térbeliség szemantikáját a 3. fejezetben; a tagadást a 4. fejezetben; a valószínűségi okfejtést az 5. fejezetben; a modalitást és tényellentétességet a 6. fejezetben; az implikatúrát és a skaláris mellékneveket a 7. fejezetben; a tulajdonneveket és a a világról való tudás integrálását a 8. fejezetben; továbbá bemutatunk néhány alkalmazást a 9. fejezetben.

Talán ennél is fontosabb, hogy felvállaljuk az *egész* lexikont, mind szélesség, mind mélység tekintetében. A `4lang` számítógépes projekt célja a teljes szókincs visszave-

zetése egy központi definiáló halmazra. A szélesség kedvéért megvizsgálunk minden sztenderd (Buck, 1949) szemantikai mezőből néhány reprezentáns elemet, a „fizikai világ”-tól a „vallás és hiedelmek”-ig (lásd: S19:6.4. és 5.3.). A teljességet az osztályok és nem az egyedek szintjén célozzuk be; például nem vállalkozunk a Buck által figyelembe vett ötvennél több „testrész és funkció” szisztematikus katalogizálására. A `4lang` csak néhány tucatot használ ezekből, és úgy látjuk, nem szükséges túllépni a reprezentatív példákon: ha az olvasó látja, hogyan kezeljük ezeket, világos lesz az általános elképzelés. Ami a többit illeti (pl. a *köldök* kívül van a `4lang` -on) az általános célú szótárakra hagyatkozunk, különösen az LDOCE-re (Procter, 1978), és kielégítőnek tekintjük azt, hogy ‘a kis üreges vagy emelt hely a hasad közepén’ mindaddig, amíg a definícióban szereplő szavak és összetételük módja definiálva van. Az olvasót a 275. o.-n kezdődő függelék segíti, ahol a `4lang` definiáló szavainak mindegyike fel van sorolva egy mutatóval a szöveg fő részének arra a pontjára, ahol a szócikkről szó esik, és számos kereszthivatkozással azok számára, akik szívesebben merülnek bele a részletekbe, mintsem követik a szélességi kifejtésnek azt a sorrendjét, amit a tárgy megkíván.

Mélység tekintetében gyakran a szószint alá megyünk, lexikai tételként vizsgálva kötött morfémákat, mind gyököket, mind toldalékokat (lásd: 2.2.). Sok elődjével ellentétben a `4lang` nem áll meg a primitívek halmazánál, hanem ezeket is definiálja a többi primitív segítségével, ahol csak ez lehetséges. Marad egy maroknyi valóban redukálhatatlan elem, mint például a *wh* kérdésmorféma, de érdekesebb az esetek 99%-a, mondjuk a *bíró* meghatározása: `ember, bíróság része/3124, dönt, alkot hivatalos(vélemény)` (a definíciókban használt formális nyelv szintaxisát lásd: 1.3.), ahol tetszőleges mélységig nyomon követhetjük az alkotórészeket.

A legfontosabb megfigyelés itt az, hogy a valódi definiálhatatlanság inkább anomália, mint a normális eset. Egyszerűen nem tudjuk a kisszámú definiálhatatlan elemre akasztani a szókincs többi részét, mert a definíciókban kiküszöbölhetetlen körkörösséggel találkozunk már jóval azelőtt, hogy minden mást ezekre vezettünk volna vissza. Ebben a könyvben végig elfogadjuk ezt a körkörösséget, sőt a heurisztikus módszer szintjére emeljük: ha elegendő masinéria áll rendelkezésre (főleg a 6. fejezetben és utána), jelentős időt fogunk fordítani arra, hogy definíciók különféle láncait alkossuk meg ismételt behelyettesítések segítségével.

Tekintsük a hét napjait. A Longman Defining Vocabulary felveszi a kesztyűt, és primitívként felsorolja az összeset: vasárnap, hétfő, . . . , szombat. Viszont világos, hogy mihelyt ezek közül az egyiket primitívnek vesszük, a többi már definiálható. Ahelyett, hogy önkényesen kijelölnénk valamelyiket alapnak, `4lang` minden definíciót egyenletként kezel, a teljes lexikont pedig olyan egyenletrendszerként, amely minden jelentésre kölcsönös kényszerfeltételeket ír elő. Hogy ez hogyan történik, az a könyv tárgya. A türelmetlen olvasó előreugorhat 9.5.-ra, ahol a kötet egészében alulról felfelé felépített algoritmust összefoglaljuk felülről lefelé stílusban.

Kinek érdemes elolvasnia ezt a könyvet?

A szemantika azt vizsgálja, hogyan közvetítődik a jelentés egyik személytől a másikhoz. Ez nagy kérdés, és több tudományág is szeretne egy darabot kihasítani. A listán szerepel a nyelvészet, a logika, a számítástechnika, a mesterséges intelligencia, a filozófia, a pszichológia, a kognitív tudomány és a szemiotika. Ezeknek a tudományágaknak sok művelője azt mondaná a hallgatóinak, hogy a szemantikának *csak* akkor van értelme, ha az ő tudományának nézőpontjából vizsgáljuk. Mi itt szinkretikus nézetet vallunk, és üdvözlünk minden olyan fejleményt, ami úgy tűnik, hogy hozzájárul a nagy kérdéshez.

Akárcsak S19 esetében, az ideális olvasó egy hekker, „olyan személy, aki örömét leli egy rendszer belső működésének mély megértésében". De ezúttal a doktori hallgatót célozzuk meg, és nemcsak S19 ismeretét feltételezzük, hanem azt is, hogy hajlandó szakcikkeket olvasni. A Zeitgeist egyik központi eleme a mesterséges általános intelligencia (AGI) világra hozása. Ez bonyolult ügy, különösen annak biztosítása tekintetében, hogy az AGI-k nem veszélyeztetik az emberiséget (a szerző álláspontjához lásd Kornai, 2014, de ezt mostanra már túlhaladja Fuenmayor és Benzmüller, 2019, és S19:9.). Nyilvánvaló, hogy az AGI egyik legfontosabb szempontja az emberekkel való kommunikáció képessége, és a könyv célja, hogy segítsen megteremteni ennek módját (szemben az érzékelőrendszer, a motoros képességek stb. megsegítésével). Ez egy számos embert foglalkoztató vállalkozás, akik többnyire nemcsak hogy központi irányítás nélkül, de gyakran egymást sem ismerve tevékenykednek. Bár csak a 9.4. foglalkozik közvetlenül a kérdéssel, a könyv ajánlott mindazoknak, akiket érdekelnek az AGI nyelvi vonatkozásai.

Ugyanakkor kifejezett célunk, hogy a nyelvészeket és kognitív tudósokat, akár szkeptikusak az AGI céljával kapcsolatban, akár nem, ismét bevonjuk a játékba. A mély modellek, különösen a transzformátorok hatalmas sikere az előrejelzésben, különösen a nyelvtanilag kifogástalan, folyékony szövegek előállításában egyértelművé teszi, hogy a szintaxis tanulás tekintetében könnyebb, mint a szemantika. Jelenleg ennek a munkának a frontvonala az AlphaCode (Li és tsai., 2022), amely figyelemre méltó szemantikai megértéssel rendelkező szoftvert generál angolul megfogalmazott programozási problémákból, hasonlóan ahhoz, ahogy számítási modellek korábbi generációi egyenletrendszereket állítottak elő MCAS szintű szöveges feladatokból (Kushman és tsai., 2014). Véleményünk szerint az ilyen rendszerek kihagyják a „gyors gondolkodás" kognitív kompetenciát, ami az emberi nyelvi megértést jellemzi, és a „lassú gondolkodást", Kahneman, 2011 2. típusú folyamatait modellezik. A mi érdeklődésünk itt az előbbire irányul, különösen a mai tudományos világképet mind ontogenetikus, mind filogenetikus értelemben megelőző, *naiv* világnézetre.

Hogyan olvassuk?

A könyvet elsősorban számítógépen való olvasásra terveztük. Nagymértékben használjuk a szövegen belüli hivatkozásokat, kékkel szedve, főleg a Wikipédiára (WP) és a Stanford Encyclopedia of Philosophyra (SEP), különösen olyan fogalmakat és gondola-

tokat illetően, amelyeket úgy érezzük, hogy az olvasó már ismer, de esetleg fel akarja frissíteni ismereteit. Mivel e linkek követése nagyban javítja az olvasási élményt, a papíralapú változat olvasóinak ajánlott, hogy legyen kéznél a mobiltelefonjuk, hogy beolvassák a hiperhivatkozásokat, amelyek QR-kódként is megjelennek a margón. A jelenlegi kötet külső indexet is tartalmaz, amely a 271 oldalon kezdődik, továbbá hozzáférhető a http://hlt.bme.hu/semantics/external2, amely a külső hivatkozások befagyasztott másolatainak gyűjteménye avégett, hogy megvédjük a olvasót a halott linkektől. Rendelkezésre áll egy hagyományos, több száz terminust tartalmazó index is, de az olvasót arra biztatjuk, hogy mindkettőben keressen, vagy ha egy kifejezés hiányzik ezekből, akár az egész fájlban is. A függelékben (275 o.) azok a definíciók, amelyeket a szövegben kifejtünk, szintén indexelve vannak. Ezeket a szavakat kiemeljük a margón ott, ahol a definíció található.

A nyelvi példák általában dőlt betűsek, és ha egy jelentést (parafrázist) adunk meg, az szimpla idézőjelben szerepel. A dőlt betűket a szakkifejezések első megjelenésénél és hangsúlyozás végett is használjuk. A `4lang` számítógépes rendszer tartalmaz egy fogalomtárat, amely eredetileg négy nyelven, az Európában beszélt főbb nyelvcsaládok reprezentatív mintáiból készült: germán (angol), szláv (lengyel), román (latin) és finnugor (magyar). Ma már több mint 40 nyelven léteznek kötések (Ács, Pajkossy és Kornai, 2013; Hamerlik, 2022), de az itt nyomtatásban megjelenő változat (lásd 275 o.) csak az angol kötéseket tartalmazza. A szövegben a szócikkek, definíciók és más, számítással kapcsolatos dolgok `írógépes betűtípussal` szerepelnek.

Ahogy az nagyszabású, sokkomponensű kutatási programoknál lenni szokott, a `4lang` számos sajátossága megváltozott a kezdeti tanulmányok közzététele óta. S19-ben ezt a kérdést nagyrészt elfedte az a tudatos törekvés, hogy mások munkáját állítsuk előtérbe, ahogy az egy tankönyvhöz illik, és minimalizáljuk `4lang` közvetlen tárgyalását. A lassú elmozdulás problémáját (nem voltak nagyobb koncepcionális fordulatok, még az Eilenberg-gépekről a vektorszemantikára való áttérés is megvalósulhatott mindössze egy ág elévülésével és egy másik hozzáadásával) most verziószámozással kezeljük: S19 megfelel `4lang` V1 kiadásának, a jelenlegi kötet pedig a V2-nek. Hacsak nincs külön feltüntetve, minden itt tárgyalt definíció, formula és statisztika a V2-es verzióból származik, lásd a 9.5. kiadási megjegyzéseket. Még sok munka marad a további kiadásokra. Ezt a szövegben időnként megemlítjük, mindig felhívásként, hogy lehet csatlakozni a nagy szabad szoftveres közösséghez, és részletesebben is tárgyaljuk a 9. fejezetben.

Köszönetnyilvánítás

Az itt bemutatott anyag egy része először tanulmányokban jelent meg, amelyek sok esetben másokkal közös munkák voltak, akiknek a hozzájárulása rendkívül jelentős. Hadd emeljem itt ki Marcus Kracht Bielefeld, (Kornai és Kracht, 2015; Borbély és tsai., 2016), és Gyenis Zalán Jagellonian University, Krakow (Gyenis és Kornai, 2019) nagyvonalúságát, akik megengedték munkánk egy részének újbóli felhasználását.

A nehéz munka nagy részét, különösen számítástechnikai oldalon, de a koncepcionális tisztázások és új ötletek tekintetében is a jelenlegi és korábbi HLT diákok, többek

között Ács Judit (SZTAKI), Borbély Gábor (BME), Makrai Márton (Kognitív Idegtudományi és Pszichológiai Intézet), Kovács Ádám (TU Wien), Lévai Dániel (Upright Oy), Nemeskey Dávid Márk (Digital Heritage Lab), Recski Gábor (TU Wien) és Zséder Attila (Lensa) végezték.

Az anyag egy részét tanítottam a 2020/21-es tanév őszi félévében a BME-n és a North American Summer School in Logic Language and Information (NASSLLI 2022)-n. Külön köszönettel tartozom ezen kurzusok hallgatóinak és a könyv korai változatai minden olvasójának, akik számos helyesírási hibát és stilisztikai botlást észrevettek, nagyon jó hivatkozásokat javasoltak és segítették az előadás menetét: Ács Judit (BME AUT), Eper Miklós (BME TTK), Gémes Kinga (BME AUT), Havas Tamás (BME TTK), Juhász Máté (ELTE), Koncz Máté (BME TTK), Kovács Ádám (BME AUT) és Tauber Boglárka (BME TTK).

Hálásan köszönöm a segítséget azoknak, akik a kézirat egyes részeihez hozzászóltak, fontos tanácsokat adtak sok kérdésben, amelyek csak az ő szakértelmükkel bírók számára kézenfekvőek; különösen Avery Andrewsnak (Australian National University), Cleo Condoravdinak (Stanford), Cser Andrásnak (PPKE), Hans-Martin Gärtnernek (Nyelvtudományi Intézet), Máté Andrásnak (ELTE Logika), Richard Rhodesnak (Berkeley), Simonyi Andrásnak (PPKE), Szabolcsi Annának (NYU) és Madeleine Thompsonnak (OpenAI). Az újonnan (V2) létrehozott japán és kínai kötések Cseresnyési László (Shikoku Gakuin Egyetem) és Bartos Huba (Nyelvtudományi Intézet) szakértelmét és nagylelkűségét tükrözik. Az eredetileg Anna Cieśliktől (Cambridge) származó lengyel kötések frissítése sokat köszönhet Malgorzata Suszczynska (Szegedi Tudományegyetem) segítségének. Különösen hálás vagyok kollégámnak, Wettl Ferencnek (BME), és Doktorvateremnek, Paul Kiparskynak (Stanford), akik mindketten részletes észrevételekkel segítettek. Mondanom sem kell, hogy nem mindenben értenek egyet a könyvben foglaltakkal, az itt kifejtett nézetek nem a finanszírozó szervezetek álláspontját tükrözik és minden hiba és hiányosság kizárólag a sajátom.

A munkát részben támogatta a 2018-1.2.1-NKP-00008 program: A mesterséges intelligencia matematikai alapjainak feltárása; az OTKA, 120145-ös szerződésszám; az Innovációs és Technológiai Minisztérium NKFI Hivatala a Mesterséges Intelligencia Nemzeti Laboratórium program keretében, valamint a MILAB, a Magyar Mesterséges Intelligencia Laboratórium. A megírásra a Budapesti Műszaki és Gazdaságtudományi Egyetem (BME) Algebra tanszékén, valamint a Számítástechnikai és Automatizálási Kutatóintézetben került sor.

Hálás vagyok szerkesztőimnek a Springernél, Celine Changnak és Alexandru Ciolannak a folyamatos professzionális, kitartó támogatásért. Az angol eredeti nyílt hozzáférését az NKFIH (Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal) MEC_K 141539 számú ösztöndíja tette lehetővé.

Előszó a magyar kiadáshoz

A magyar kiadás három dologban tér el az angoltól. Egyrészt itt van ez az előszó, ami az angolból teljesen hiányzik. Másrészt az éles szemű fordító, Máté András (ELTE) számos apróbb-nagyobb hibát talált az eredetiben, amiket ott is javítottunk, a hibajegyzéket ld. https://www.kornai.com/Books/VectorSemantics alatt. Ilyen értelemben a magyar kiadás jobb az eredetinél, és természetesen 'magyarabb' is annyiban, hogy a nyelvtani példák egy részét a magyarból vett példákkal pótoltuk, és az irodalomjegyzék is bővült magyar nyelvű munkákkal, illetve ahol a szakirodalom magyar fordításban is elérhető, ott ezt igyekeztünk feltüntetni. Bár a szövegben az angol nyelvű S19-re hivatkozunk, megemlitjük, hogy ennek van magyar fordítása is, Kornai, 2018. A magyar tipográfiai konvenciók gépi betartását ismét Wettl Ferenc hathatós LaTeX segítsége tette lehetővé.

Harmadszor, az implementáció alapjait képző `4lang` számítógépes projekt magszókincsét tartalmazó 700.tsv fájl nyomtatott változatában az angol kiadás függeléke a névadó négy nyelv (angol, magyar, latin, lengyel) közül csupán az angolt tüntette fel, míg a magyar kiadásban már szerepelnek ezek mellett az újonnan (V2) létrehozott japán és kínai kötések is, melyek Cseresnyési László (Shikoku Gakuin Egyetem) és Bartos Huba (Nyelvtudományi Intézet) szakértelmét és nagylelkűségét tükrözik. Az eredetileg Anna Cieśliktől (Cambridge) származó lengyel kötések frissítése sokat köszönhet Malgorzata Suszczynska (Szegedi Tudományegyetem) segítségének.

Különösen hálás vagyok a fordítónak, Máté Andrásnak, nemcsak a hibák kiszúrásáért, hanem azért a számtalan apró és nagyobb tanácsért is, amik a magyar kiadást jobban érthetővé, logikusabbá tették; Balázs Péternek, a kiadó szerkesztőjének; és a sasszemű Erő Zsuzsának, akik nélkül ez a kiadás csúnya lenne és hemzsegne a hibáktól. A fennmaradó hibákért természetesen a szerző a felelős. A magyar kiadás előkészítését az NKFIH (Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal) MEC_K 141539 számú ösztöndíja tette lehetővé.

Tartalomjegyzék

1.

A nemkompozicionalitás alapjai

Tartalom

Az elmúlt fél évszázadban a nyelvi szemantika főleg a kompozicionalitás kérdéseivel foglalkozott, olyannyira, hogy az atomi egységeknek (általában a szavakat vagy azok töveit tekintik annak) a jelentése alig kapott figyelmet. Itt a szó jelentését helyezzük előtérbe, és az egész könyvet úgy építjük fel, hogy a jelentéssel bíró legalsó egységekkel, a morfémákkal kezdjük, majd felfelé építkezünk. A 1.1. szakaszban berendezzük a színt, végiggondolva a három fő megközelítést a szemantikában, amelyeket formális apparátusuk alapján megkülönböztethetünk: logikai, geometriai és algebrai. A 1.2. szakaszban összefoglalunk néhány lexikográfiai elvet, amelyeket az egész könyv során alkalmazni fogunk: egyetemesség, reduktivitás és a lexikon mentesítése az enciklopédikus ismeretektől. A 1.3. szakaszban leírjuk a lexikális jelentés logikai elméletét. Ezt kapcsoljuk a 1.4. szakaszban a geometriai és a 1.5. szakaszban az algebrai elmélethez. Az algebrai és a geometriai elmélet közötti kapcsolatokról a 1.6. szakaszban beszélünk, ahol megvizsgáljuk egy meta-formalizmus lehetőségét, amely összekapcsolhatná mindhárom megközelítést.

1.1. Háttér

A **logikai** (formula alapú) szemantika elmélete (S19:3.7.), a Montague-grammatika (MG) és egyenesvonalú leszármazottai, mint például a diskurzusreprezentáció-elmélet (DRT) és a dinamikus szemantika a 21. századig uralkodó szerepet játszottak a nyelvi szemantikában, jól ismert hiányosságaik ellenére, mert nem volt és bizonyos tekintetben

ma sincs más megközelítés: az alternatív „kognitív” elmélet igazából nem lett formalizálva, a logikai alapú iskola „markeréznek” (Lewis, 1970) titulálta. Itt megkíséreljük formalizálni a kognitív elmélet számos – de távolról sem minden – felismerését. Ezt a vállalkozást még szükségesebbé teszi az, hogy az MG úgyszólván semmit sem mond az atomi egységek természetéről (Zimmermann, 1999).

Talán a Schütze, 1993; Schütze, 1998 munkákkal kezdve, majd Collobert és Weston, 2008; Collobert és tsai., 2011 általános sikerének hatására sztenderddé vált a számítógépes nyelvészetben egy teljesen új, **geometriai** elmélet, amely a jelentéseket alacsony dimenziós euklideszi tér vektoraihoz rendeli (S19:2.7. 2.3. példa skk.). A szemantika központi témái, mint például a kompozicionalitás vagy a szintaktikai és szemantikai reprezentációk viszonya, amiket eddig kizárólag logika alapú keretrendszerben tárgyaltak, a geometriai elmélet figyelmének középpontjába kerültek (Allauzen és tsai., 2013), azonban még mindig nincs széles körben elfogadott megoldás ezekre a problémákra. Az egyik váratlan fejlemény a geometriai elméletben az volt, hogy a morfológia, a szintaxis és a szemantika bizonyos mértékig különböző rétegekben találhatók a többrétegű modellekben, amelyeknek szóvektorok a bemenetei (Belinkov és tsai., 2017b; Belinkov és tsai., 2017a), de a modellek „kikérdezése” még mindig művészet, lásd néhány korai munkát ebben az irányban (Karpathy, Johnson és Fei-Fei, 2015; Greff és tsai., 2015), és a kontextuális beágyazásokkal kapcsolatos frissebb munkákat (Clark és tsai., 2019; Hewitt és Manning, 2019).

Ugyanakkor a szemantika **algebrai** elmélete (S19:Def 4.5. skk.), amit az 1960-as évek óta kutatnak a mesterséges intelligencia területén (Quillian, 1969; Minsky, 1975; Sondheimer, Weischedel és Bobrow, 1984), és ami a mondatok és nagyobb egységek jelentésének ábrázolásához (hiper)gráfokat használ, új lendületet kapott a Google erőfeszítéseinek hatására. Ezek a világról való tudás egy nagy adatbázisban való összegyűjtésére irányultak, azzal a módszerrel, hogy azonosították a tulajdonneveket a szövegben, és ezeket egy nagy külső tudásbázishoz, a KnowledgeGraphhoz kapcsolták. Ez jelenleg több mint 500 millió entitást tartalmaz, amelyeket 170 milliárd reláció vagy „tény” köt össze (Pereira, 2012). A nyelvészeti szempontból motivált algebrai elméletek (Kornai, 2010a; Abend és Rappoport, 2013; Banarescu és tsai., 2013), párosulva a függőségi elemzés iránti újraéledt érdeklődéssel (Nivre és tsai., 2016), hozzájárulnak a háttértudás szerepének újjáértékeléséhez és a hipergráfok használatához a szemantikában (Koller és Kuhlmann, 2011).

Ebben a könyvben megpróbáljuk összekapcsolni ezt a három megközelítést, matematikai formába öntve azt a meggyőződést, hogy nem mások, mint ugyanannak az elefántnak a törzse, lába és a farka. Ez nem azt jelenti, hogy csupán „jelölésbeli változatok” lennének (Johnson, 2015), ellenkezőleg, mindegyikük tesz olyan előrejelzéseket, amelyek a többinél hiányoznak. Jobb analógia a lineáris algebra algebrai (mátrix) és geometriai (transzformációs) megközelítése: mindkettő egyformán érvényes, de nem minden helyzetben egyformán hasznos.

Nem árt óvatosságra inteni: az 1.3. szakaszban tanulmányozott formulák nem annak a magasabb rendű intenzionális logikának a formulái, amelyet az MG hallgatói

jól ismernek, hanem egy sokkal egyszerűbb, komplexitásban jóval az elsőrendű logika (FOL) alatti proto-logika alapvető építőkövei. A gráfok, amelyeket a 1.5. szakaszban kezdünk tanulmányozni, hipergráfok, nagyon hasonlítanak a kognitív nyelvtan, a függőségi nyelvtan (DG, Dependency Grammar), a lexikai funkcionális nyelvtan és az HPSG (Head-driven Phrase Structure Grammar), valamint az MI (mesterséges intelligencia) jelölési eszközeire, de nem azonosak betű szerint a korábbi javaslatok széles választékának egyik elemével sem. Az egyetlen dolog, ami ugyanolyan, a geometria, az n dimenziós euklideszi geometria, amit mindenki más is használ, de még itt is lesz néhány furcsaság, lásd az 1.4. szakaszt.

1.2. Lexikográfiai alapelvek

Univerzalitás A `4lang` egy fogalomtár, amelyet az alábbiakban részletezett értelemben univerzálisnak szánunk. Hogy megtegyük az első óvatos lépéseket a nyelvfüggetlenség irányába, a rendszert négy nyelv kötéseivel hoztuk létre. Ezek reprezentatív mintát adnak a fő európai nyelvcsaládokból: germán (angol), szláv (lengyel), román (latin) és finnugor (magyar). Az 1. verzió több mint 40 nyelv automatikusan létrehozott kötéseit tartalmazza (Ács, Pajkossy és Kornai, 2013), de érdemes figyelembe venni, hogy ezek a kötések csak durva szemantikai megfelelésben állnak a megcélzott fogalommal. A jelenlegi 2. verzióhoz (lásd 9.5.) Cseresnyési László és Bartos Huba kézi úton adott hozzá két keleti nyelvet, a japánt, illetve a kínait, továbbá újabb automatikus kötések lettek létrehozva (Hamerlik, 2022).

A négy nyelvben végrehajtott párhuzamos fejlesztés tapasztalata megerősített egy egyszerű tényt, amit a lexikográfusok mindig is magától értetődőnek tartottak: a szavak, illetve szójelentések nem illeszkednek össze a különböző nyelveken, még ennek a négy nyelvnek az esetében sem, amelyek pedig osztoznak egy közös európai kulturális / civilizációs háttérben.Nem csak az a probléma, hogy egyes fogalmak egyszerűen hiányoznak bizonyos nyelveken (ez a kölcsönzés gyakori oka), hanem az is, hogy az egész fogalmi tér (lásd 1.4.) másképpen osztható fel.

Például az angol nyelv általában nem különbözteti meg azokat a cselekvést leíró igéket, amelyek az alanyra és amelyek a tárgyra hatnak azonos módon: hasonlítsuk össze a *John turns, John bends* kifejezéseket a *John turns the lever, John bends the pipe* kifejezésekkel. A lengyel nyelvben ahhoz, hogy kifejezzük, hogy az első esetben John az, aki fordul/hajlik, szükség van egy tárgyesetű visszaható névmásra, a *się* 'önmagát' kifejezésre. A szemantika ugyanaz, de az angolban a *??John turns/bends himself* kifejezés furcsán hangzik. A magyarban különböző igéket kell használnunk, amelyek ugyanabból a tőből származnak: az 'önmagát fordítja' *ford-ul*, míg a 'valamit megfordít' *ford-ít*, és hasonlóan az 'önmagát hajlítja' *haj-ol*, míg a 'valamit meghajlít' *hajl-ít*, hasonlóan a latin *versor/verso, flector/flecto* igékhez, de a latin nyelv lehetőséget nyújt a *me flecto/verso* névmás használatára is.

Hová jutunk így az egyetemesség magasztos célkitűzésével kapcsolatban? Az egyik végleten találjuk az erős Sapir–Whorf-hipotézist, miszerint a nyelv meghatározza a gondolkodást. Ez azt jelentené, hogy az angol beszélő nem tudná megosztani a hajlítás fogalmát a magyar beszélővel, mivel az angol nyelv egy szót használ két különböző fajta helyzetre, amire a magyarnak két különböző szava van. A másik végleten az itt követett módszert találjuk: nagyon absztrakt egységeket (maglexémákat) alkalmazunk, amelyekről feltételezzük, hogy a nyelvek között közösek, viszont lehetővé tesszük, hogy a nagyobb egységek ezekből nyelvek szerint különböző módokon épüljenek fel. Jelen esetself ben a szükséges kulcsfogalmak közé tartozik a *visszahatás*, amelyet a `=pat[=agt]`, `=agt[=pat]` alakban definiálunk (lásd még 3.3.), valamint a *hajlít* fogalma, amelyet az intranzitív alakban tekintünk alapvetőnek (lásd 2.4.). A 3.1. szakaszban térünk rá arra, hogy általában hogyan határozhatók meg a tranzitívek a tárgyatlan megfelelőjük segítségével.

Azt, hogy hogyan kell létrehozni, kezelni és értelmezni ilyen formulákat, az 1.3. szakaszban tárgyaljuk; itt a magas szintű formázással kezdjük. A fő `4lang` fájl 11 tabulátorral elválasztott mezőre van felosztva, amelyek közül az utolsót a megjegyzések számára tartjuk fenn (ezek százalékjellel kezdődnek). Egy tipikus szócikk, amely egy sorban van megírva a fájlban, de itt a szövegben általában két sorba van tördelve a jobb olvashatóság érdekében, így nézne ki:

```
water víz aqua woda mizu ? shui3 ? 2622 u N
      liquid, lack colour, lack taste, lack smell, life need
```

Ahogy látható, az első négy oszlop az eredeti négy nyelvi kötés EN HU LA PL sorrendben. Az 1. verzióban az összes kiterjesztett latin karaktert az alapkarakterrel és egy számmal helyettesítettük, például o3 az ő betűhöz, o2 az ö-höz és o1 az ó-hoz. Ennek az volt a célja, hogy az alapvető Unix segédprogramok, mint például a `grep` viselkedése állandó maradjon különböző platformokon (elérhetők voltak konverziós szkriptek az utf8-ra és -ről való átalakításhoz). A 2. verzióban két új oszlopot adtunk hozzá a negyedik után a JA ZH nyelvek számára (lásd 9.5.), és mindenütt az utf8-kódolt ékezetes karaktereket használjuk. A hetedik (a V1-ben ötödik) oszlop minden fogalomhoz egy egyedi szám, amely különösen fontos, amikor az angol kötések megegyeznek:

```
cook főz     coquo  gotowac 825  V
     =agt make <food>, ins_ heat
cook szakács coquus kucharz 2152 N
     person, <profession>, make food
```

A nyolcadik (a V1-ben hatodik) oszlop a reducibilitási státusz becslése. Csak négy értéket vehet fel: `p` azt jelenti, hogy primitív, azaz olyan szócikk, amelyet más szócikkekre lehetetlen visszavezetni. Egy példa erre a *wh* kérdésmorféma, amit itt `wh ki/mi/hogy quo kto/co/jak 3636 p G wh`-ként adunk meg. Vegyük észre, hogy a definiendum (1. oszlop) megjelenik a definiensben (10. oszlop), ami világosan mutatja ennek a szócikknek az irreducibilitását. A másik végletbe `e`-vel megjelölt szócikkek tartoznak; ez azt jelenti, hogy kiküszöbölhető. Egy példa erre a *three*

wh

three

`three három tres trzy 2970 e A number, follow two`. Közöttük helyezkednek el a `c`-vel jelölt szócikkek, ezek a jelöltek a magszókincsre: például a *see* `see lát video widziec 1476 c V perceive, ins_ eye`; valamint az `see u` megjelölés, ami ismeretlen egyszerűsíthetőségi státuszt jelent.

A kilencedik (a V1-ben hetedik) oszlop egy vázlatos lexikai kategória-szimbólum, további tárgyalását lásd a 2.1. szakaszban. Fő témánk itt a tizedik (a V1-ben a nyolcadik) oszlop, amely a `4lang` definíciót adja. A definíciók formális szintaxisát a 1.3. szakaszra hagyjuk, miután megtárgyaltunk néhány további lexikográfiai elvet, és – először informálisan – bevezettünk néhány jelölést. Sok technikai eszköz, mint például `=agt, =pat, wh, gen, ...` itt jelenik meg először, de csak a későbbi fejezetekben lesz teljesen megmagyarázva. Nagyon gyakran lesz okunk arra, hogy csak rövidített formában mutassuk be a szócikkeket, csak a címszót és a definíciót mutatva (az indexet, reducibilitást, a fogalom számát és a lexikai kategóriát szükség szerint megjelenítve vagy elhallgatva):

```
bend 975 e V has form[change], after(lack straight/563)
```

Amikor ilyen rövidített szócikkek megjelennek a folyó szövegben, például itt a *drunk*, `drunk ittas potus pijany 1165 c A quality, person has cause_, lack control`, a címszót kiemeljük a margón. Az olvashatóság érdekében a fogalom számát elhagyjuk abban az esetben, ha az angol kötés egyedi, tehát az előző definícióban `person` szerepel, nem pedig `person/2185`, de a `man` szót kiírjuk `man/659` 'homo' formában, hogy megkülönböztessük a `man férfi vir mezczyzna 744 e N person, male` fogalomtól. A folyó szövegben általában elhagyjuk a japán és kínai megfelelőket a könnyebb nyomtathatóság érdekében.

drunk

man

A példákat általában a V2/700.tsv fájlból vesszük, de néha szükségünk lesz arra, hogy az adott kérdés szemléltetése végett túlmenjünk a `700.tsv` készleten, és (nagyon ritkán) még a V1 fájlon is.

Reduktivitás `4lang` számos szempontból logikus folyománya annak a modern, számítástechnikára orientált lexikográfiai munkának, amely a Collins-COBUILD (Sinclair, 1987), Longman Dictionary of Contemporary English (LDOCE) (Boguraev és Briscoe, 1989), WordNet (Miller, 1995), FrameNet (Fillmore és Atkins, 1998) és VerbNet (Kipper, Dang és Palmer, 2000) művekkel vette kezdetét. A rendszeres reduktivitás fő motivációját a (Kornai, 2010a) cikkben a következőképpen fejtettük ki:

„A lexikon formális modelljének létrehozásában a döntő nehézség a hagyományos szótárdefiníciók körkörössége – már az első angol szótár, a Cawdrey, 1604 a *heathen*-t mint `gentile`-t és a *gentile*-t mint `heathen`-t definiálja. A problémát már Leibniz is felismerte (idézi Wierzbicka, 1985):

> Tegyük fel, hogy nagy pénzösszeget ajándékozok neked, mondván, hogy megkaphatod Titiustól; Titius elküld téged Caiushoz, Caius pedig Maeviushoz; ha ilyen módon folyton egyik személytől a másikig küldenek, sosem fogsz semmit kapni.