Kornai András

Szemantika

Kornai András

Szemantika

TYPOTEX

A könyv megjelenését támogatta:
a Magyar Tudományos Akadémia és
a Nemzeti Kulturális Alap a kiadói program keretében.

A fordítás a következő kiadás alapján készült:
Semantics, Springer International Publishing, 2018

Hungarian edition © Kornai András, Typotex, Budapest, 2018
Hungarian translation © Varasdi Károly (Előszó, 1. fej.);
Máté András (2.); Kálmán László (3–5.), Makrai Márton (6.),
Gyenis Zalán (7–8.), Takó Ferenc (9.)

Engedély nélkül semmilyen formában nem másolható!

ISBN 978 963 279 970 4

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
Főszerkesztő: Horváth Balázs
A borítót készítette: Szalay Éva
Készült a Kódex Könyvgyártó Kft. nyomdájában
Felelős vezető: Marosi Attila

Áginak

Előszó

A szemantika a nyelvi kifejezések jelentésének (értelmének) tanulmányozása. Az utolsó fejezet kivételével a könyv középpontjában a természetes nyelv kifejezéseinek, jellemzően teljes mondatoknak és hosszabb szövegeknek a jelentése áll, nem pedig a számítógépes programok, matematikai képletek jelentése vagy tágabb szemiotikai megfontolások. A mindennapos használatban a *szemantika* szó leginkább a szavak jelentésére vonatkozik: az Urban Dictionary egyenesen úgy definiálja a szemantikát, mint

> Szavak vagy szavak csoportjai által egy adott kontextusban kifejezett jelentések/értelmezési lehetőségek tanulmányozása; rendszerint annak érdekében használjuk, hogy lezárjunk valamilyen vitát. Példa: *Ugyan, ne ragadjunk már le a szavak jelentésénél.*

A területet ugyanúgy tekinthetjük a nyelvtudomány, a számítástechnika, a filozófia vagy a kognitív tudomány egyik fejezetének, és bizonyos mértékig a könyv szerkesztése tükrözi is ezt a sokértelműséget oly módon, hogy a margó egyes pontjait a **Nyelv** és **Szám**, és alkalmanként a **Fil** vagy **Kog** jelöléssel látja el. Mivel senkitől sem várható el, hogy szakértő legyen mindezeken a területeken, az egyes területekre vonatkozó előfeltételeket külön-külön tárgyaljuk.

Kinek ajánljuk ezt a könyvet

A tágabb célunk az, hogy bemutassuk a szövegek megértésére képes szemantikai rendszerek megépítéséhez szükséges fogalmi és formális eszközöket, mind az olyan konkrét feladatok esetében, mint az információkinyerés és a kérdésmegválaszolás, mind pedig az olyan nagyobb léptékű fejlemények esetében, mint amilyen a szemantikus web. A szűkebb célunk az, hogy bemutassuk a működő rendszerek alapjául szolgáló elképzeléseket, és tankönyvünk elsősorban a a szemantikai rendszerek kifejlesztésében érdekelt számítástechnikus vagy mérnök hallgatót célozza. A könyv ideális olvasója a jó értelemben vett *hekker*, azaz 'olyan személy, aki örömét leli egy adott rendszer belső működésének mély megértésében'. Ez nem csak a dolgok kipróbálását és a velük való kísérletezés hajlandóságát, hanem a kutatás, különösen a matematikai modellezés iránti

pozitív hozzáállást jelenti. A könyv elég sokat követel ebben a tekintetben: a mondatok hosszúak, gyakoriak a több mint három szótagos szavak, a gyorsan-találjuk-meg-a-kulcsszót-és-essünk-túl-rajta jellegű olvasást támogató oldalsáv és kapcsolódó tipográfiai fogások teljes egészében hiányoznak, és számítunk arra, hogy az olvasó elegendő időt szán a gyakorlatok megoldására és nem habozik utánaolvasni az ismeretlen anyagrészeknek, ha szükséges.

Szám

A hangsúly az eszméken van, de a GitHubon elérhető (a könyv írásakor olyan 2500 sornyi) programkód, és a könyv kifejezetten támogat egy nagyon gyakorlatias, magával a kóddal kezdődő tanulási tervet, amelynek során csak akkor nézzünk bele a könyvbe, amikor erre feltétlenül szükség van. Az e tervet követő olvasók figyelmét azonban fel kell hívni arra, hogy a könyvben tárgyaltaknak csak mintegy a harmada van ténylegesen lefedve a kóddal, a többit saját maguknak kell előállítaniuk. Más jellegű, hagyományosabb olvasási terveket az 1.5. szakaszban javasolunk.

Az olvasók, akik azért tanulmányozzák a szemantikát, mert szükségük van valamiféle szemantikai feladatot ellátó programkódra, a kód dokumentációjában megtalálják a hivatkozásokat a könyv megfelelő oldalaihoz. Hiperhivatkozások biztosítják a kétirányú kapcsolatot a szöveg és a fő gondolatokat implementáló Python-kód egyre növekvő törzse között. Ez az anyag a tapasztalt szoftverfejlesztő számára készült — maga a könyv semmilyen bevezetést nem tartalmaz a Pythonba — és nem szabad a kód részletes dokumentációjának gondolni. Az olvasókat erősen arra szeretnénk ösztönözni, hogy saját maguk is járuljanak hozzá a kódhoz a https://github.com/kornai/4lang címen CC attribútum vagy hasonló licenc alatt, amely gyengébb (megengedőbb), mint a GPL, amennyiben lehetővé teszi a kereskedelmi felhasználást és nem korlátozza a kód többi részének a terjesztését (a GNU LPGL, a BSD és a hasonló licencek tökéletesen megfelelnek).

A számítástechnika felé orientálódó olvasó, aki ideális esetben doktori hallgató vagy felsőéves informatika vagy mérnök hallgató, úgy fogja találni, hogy a könyv önmagában is tökéletesen megáll, a matematikai előzményeket összefoglaló 2. fejezet kivételével. A könyv hátralévő fejezeteiben ezeket az eszközöket fogjuk használni és továbbfejleszteni, elsősorban a természetes és nem pedig a programozási nyelveket szem előtt tartva. Emiatt a speciális tárgyválasztás miatt a könyvben feldolgozott anyag meglepően csekély átfedést mutat a progamozási nyelvek szemantikájához matematikai logikából manapság megkövetelt előzetes ismeretekkel, ahol az egyik fő attrakció a bizonyítások programokként történő értelmezése. A kétféle szókincs figyelemre méltó egybeesése — amelyek egyikét logikusok hozták létre a matematikai tételbizonyítás elemzése céljából (erről röviden beszélünk a 2.6. pontban), a másikat a számítógéptudósok a kiszámítási eljárások tanulmányozása céljából — a képzett programozó vagy logikus számára szinte ellenállhatatlan vonzerővel bír, de a természetes nyelv egy sokkal egyszerűbb, *nulladrendű* elmélet felé fog vinni minket, a kijelentéslogika egy modális kiterjesztése felé, ahol a fogalmak gépi tanulhatósága a nehéz kérdés.

Nyelv

A természetes nyelvi inputot kezelni képes rendszerek tervezése jelentős nyelvészeti tudást igényel és a **Nyelv** szimbólummal jelölt bekezdések az olyan olvasókat célozzák,

akik még nem rendelkeznek ilyen tudással. A kötet keretein belül nem volt kivitelezhető a nyelvészetben feltételezett technikai háttér bemutatása, ezért ezek a bekezdések a megfelelő nyelvészeti szakirodalomra utalnak, céljuk elsősorban az önálló tanulás megkönnyítése. Figyelmeztetjük az olvasót, hogy az anyag kiválasztása kizárólag hasznossági megfontolásokon alapult, és még a hivatkozásokat lelkiismeretesen követő olvasó sem fog tudni a nyelvészeti elképzelésekről pusztán ezek alapján koherens képet kialakítani, hogy a modern nyelvészeti szemantikáról már ne is beszéljünk. Jacobson (2014) teljes egészében a kompozicionális szemantikának szentelt egy kiváló kötetet, amelynek jól időzített megjelenése lehetővé tette a jelen szerző számára, hogy több helyet hagyjon a lexikai szemantikának azzal együtt, hogy a kötet kezelhető méretkorlátok között maradjon.

A nyelvészet szakos hallgatók, különösen azok, akik egyben számítástechnikai szemléletűek, is valószínűleg könnyebben elsajátítják a szükséges anyagot, habár ez a klasszikus formális szemantika tananyagának jelentős átcsoportosítását és bizonyos fokú elfelejtését igényli. A gyengébb számítógépes háttérrel rendelkező diákoknak ajánlatos Jurafsky és Martin (2009) vagy Bird, Klein, és Loper (2009) munkáival kezdeni. Nagyon keveset feltételezünk filozófiából vagy kognitív tudományból (lásd alább), de a könyv olvasásához vannak bizonyos előzetes matematikai ismeretek, amelyeket a 2. fejezet tárgyal.

Fil

A **Fil** szimbólummal jelölt szakaszokban tárgyalt elképzelések leggyakrabban az olyan filozófiai ágakra vonatkoznak, mint a nyelvfilozófia és a tudományfilozófia. A filozófiai gondolatokra való hivatkozás különösen akkor válik majd gyakorivá, ha valamilyen kérdést a puszta alapoktól kezdve kell felfejtenünk, mint például a 3. fejezetben. Időnként megragadjuk a lehetőséget, hogy rámutassunk a filozófusok elképzeléseivel való kapcsolatra, de ezek a megjegyzések, kivéve talán a 9.1. szakaszt, nem jelentenek célirányos bevezetést a filozófia vonatkozó fejezeteihez és önmagukban nem is elégségesek az önképzéshez. Ezeket inkább azoknak az olvasóknak szánjuk, akik már eleve érzékenyek a filozófiára, és tesszük ezt azzal a céllal, hogy tájékozódási pontokat nyújtsunk az ilyen olvasók számára arra vonatkozóan, hogy az itt megfogalmazott álláspontok hogyan is illeszkednek az e témákat körülvevő nagyobb filozófiai vitákba.

Új megoldásokat kínálunk néhány jól ismert filozófiai rejtvényre, különösen a kupac paradoxona (szóritész) és a szupererogáció rejtvényeire, de ezek a megoldások a gyakorlatra vannak kihegyezve (mivel e problémák valóban felmerülnek a rendszertervezésben), és nem átfogó filozófiai megoldásnak vannak szánva. Általában is igaz az, hogy a filozófia kutatói is sok új ismeretre tesznek majd szert a nagy filozófiai kérdések némelyikéről, olyanokról mint: Mi a jelentés? Mi a tudás? Mi az igazság? de ezeknek a nagy kérdéseknek semmiféle rendszeres filozófiai kezelésével nem szolgálunk a könyvben. Ehelyett egy felettébb technikai apparátust adunk az olvasó kezébe, amellyel képes *modellezni* a jelentést, a tudást, és kisebb mértékben az igazságot, mind számítási, mind pedig matematikai (inkább algebrai, mint logikai alapú) eszközök használatán keresztül.

A (filozófiai) logika oktatása általában nem fedi le sem a szükséges számítástudományi, sem a matematikai előismereteket, és ennek tudatában összeállítottunk egy, a filozófusoknak legmegfelelőbb olvasási tervet a 1.5. szakaszban. Ennek a könyvnek az olvasásakor az egyik dolog, amit a filozófusnak el kell felejtenie, vagy legalábbis nagyon korlátoznia kell, az az erősen technikai nyelvezet iránti elfogultság. Amikor a *helyes* fogalmát elemezzük, akkor amit adunk, az a mindennapi fogalom elemzése, nem pedig egy kifinomult jogelmélet. Persze meg fogjuk különböztetni $right_1$ 'dextra', $right_2$ 'bonus' és $right_3$ 'ius' értelmeket, de az utolsó felsorolt értelem definíciója nálunk egyszerűen 'törvény', ahol a *törvény*t mint `rule, system, society/2285 HAS, official, 'ACCEPT, ABOUT can/1246[person[=TO]]` értelmezzük (lásd 6.5. a fenti definíció formális elméletét illetően). A törvények első közelítésben olyan szabályok, amelyekkel társadalmak rendelkeznek, hivatalos státuszuk van, az emberek elfogadják őket, és e törvények arra vonatkoznak, hogy mit tehetnek meg az emberek. De itt álljunk csak meg egy kicsit — hát nincsenek talán igazságtalan törvények, vagy olyanok, amiket az emberek nem fogadnak el? Nincsenek olyan jogok, amelyek túllépnek a társadalmon? Úgy véljük, hogy ezek a kérdéseket, legyenek bármennyire tiszteletreméltóak is, nem lehet gyümölcsözően megközelíteni a hétköznapi nyelv tanulmányozása révén. Idézzük Kornai (2008)-at:

> Mivel szinte az összes társadalmi tevékenység végső soron a nyelvi kommunikáción alapul, nagy a kísértés arra, hogy egyéb tudományterületről származó problémákat tisztán nyelvészeti problémákra vezessük vissza. A skizofrénia megértése helyett talán először azon kellene eltöprengenünk, hogy mit is jelent a *többszörös személyiség* kifejezés. A matematika a többszörös fogalmára ad egy jól használható fogalmat, de mi a személyiség, és hogyan lehet egynél több ilyen személyenként? Lehet, hogy az *-i* és *-ség* képzők jelentik a probléma megértésének a kulcsát?

Kog

Eredetileg megértésre képes rendszereket olyan a mesterséges intelligencia területén dolgozó kutatók építettek, mint Allen Newell és Herbert Simon, akik az emberi megismerési képességet próbálták meg modellezni, de legalábbis fogalmakat kölcsönözni abból, amit az adott korban tudtak az elme szerkezetéről. Néhány ilyen rendszer, például a SOAR vagy az ACT-R, még mindig használatban vannak, míg másokat felváltottak az olyan újfajta kognitív architektúrák, mint például a OpenCog. A funkcionális MRI technikák megjelenésével a kutatási terület óriási fejlődésen ment át, de ebben a könyvben nagyon kevés van, ami ehhez a máris hatalmas és még mindig gyorsan növekvő irodalomhoz köthető. Az ok, a szerző nyilvánvaló korlátain túl az, hogy a természettől ellesett ideák használata zsákutcának bizonyult a természetes nyelv feldolgozásának területén.

Az olyan fontosabb szemantikai rendszerek, mint az IBM Watsonja, nem óriási elektronikus agyak; tulajdonképpen rendkívül keveset kölcsönöznek a biológiai rendszerekre vonatkozó tudásunkból. Előfordulhatnak persze különböző rendszerösszetevőkben neurális hálók, ám igen gyakran találkozunk olyan statisztikai tanulórendszerek-

kel, mint például a támaszvektoros gépek, amelyek teljes egészében szakítanak a biológiai metafórával. A kognitív tudományon belül a BICA Society vezetésével megújult erőfeszítéseket tesznek a biológiailag ihletett modellek felé, de eddig ezek nagyon kevés fogást találtak a szemantika központi kérdésein.

Mivel az algoritmusok egyre inkább képesek korábban csak ember által végezhető feladatok ellátására, a természetes nyelvi megértés igényei által követelt alaparchitektúra a filozófus és a kognitív tudós érdeklődésére is számot tarthat, és a 3. és a 9. fejezet sok ide vonatkozó anyagot tartalmaz. Mondani sem kell, hogy ezek a tudományágak sok olyan kérdéssel is foglalkoznak, amely nem tartozik a jelen könyv hatókörébe, ezért a filozófus hallgatónak tanácsos átolvasni legalább a Boden (2006) 16. fejezetét. A kognitív tudomány szakos hallgató természetesen olvassa csak el az egész kétkötetes munkát, és az újabb Gordon és Hobbs (2017) kötetet, nem egyszerűen előzetes olvasmányként a jelen könyvhöz, hanem azért is, hogy mélyebb ismereteket szerezzen. A korszerű szemantika technikai eszközeinek alapoktól kezdődő részletes bemutatása ezt a könyvet nagyrészt a Boden-könyv kiegészítő párjává teszi, és meglehetősen sok olyan fejleményt is bemutat, amit Boden nem tudhatott figyelembe venni azon egyszerű oknál fogva, hogy ezek az ő könyve megjelenését követően kerültek be a köztudatba.

A könyv szemléletéhez nagyon közel állnak a megtestesített megismerés elméletének központi állításai, de ezzel együtt is a formális szimbólum-manipuláció irányából közelíti meg a dolgokat, mivel a hangsúly az olyan szemantikai feladatok elvégzésére alkalmas algoritmusok létrehozásán van, mint például a sematikus következtetések (lásd 7.1.) elvégzésére alkalmas eljárások — még akkor is, ha ez a kognitív realizmus rovására történik. A könyv azok számára íródott, akik repülő gépek építésében érdekeltek, függetlenül attól a módtól, ahogy a madarak ténylegesen repülnek, és az egyetlen vigasz a kognitív tudomány kutatója számára itt az lehet, hogy a technikai eszközök egy része — hogy úgy mondjuk: a megértés aerodinamikája — szükségszerűen jelen van mind a két területen. A megismeréstudománnyal foglalkozó számára a 1.5. szakaszban felvázolt olvasási terv hangsúlyozza ezt a közös nézőpontot.

Betűszedési konvenciók

A könyvet elsősorban a számítógépen történő olvasásra tervezték. Folyamatosan használunk kékkel szedett szövegközi hivatkozásokat, különösképpen a Wikipédia (WP), a PlanetMath és a Stanford Encyclopedia of Philosophy (SEP) weboldalakra történő hivatkozások során, és elsősorban azon fogalmak és elképzelések esetében, amelyekről úgy gondoljuk, hogy az olvasó már valószínűleg ismeri őket ugyan, de esetleg fel kívánja frissíteni a tudását. Mivel ezeknek a linkeknek a követése igen jelentősen javítja az olvasási élményt, a papírverzió olvasóinak azt javasoljuk, hogy vegyenek egy mobiltelefont a kezükbe, hogy be tudják olvasni a hiperhivatkozásokat, amelyeket a margókon lévő QR kódok szintén megjelenítenek.

A könyv egy újdonsága, hogy egy külső indexet is tartalmaz, amely a PDF-ben a 317. oldalról indul, és a nyomtatott változat olvasóinak a http://hlt.bme.hu/szemantika

weblapon érhető el. Ez egyes külső hivatkozások egy befagyasztott másolatát tartalmazza, annak érdekében, hogy megkímélje az olvasót a már érvényüket vesztett hivatkozások lekövetésétől. A hagyományos index több száz indexelt kifejezéssel továbbra is rendelkezésre áll, de az olvasót arra biztatjuk, hogy keressen a könyv szöveges állományában, ha az indexből hiányzik egy adott kifejezés. Bizonyos esetekben egy kifejezést esetleg informálisan használunk (szövegközi hivatkozással vagy anélkül), mielőtt egy formálisabb meghatározást adnánk. Az említett forrásokban használt jelölési konvenciók nem mindig egyeznek meg a könyvben használtakkal: például mi a $\langle a, b \rangle$ jelölést használjuk arra a rendezett párra, amelyet a WP úgy jelöl, hogy (a, b).

A könyvben bemutatott technikai anyagok sokféleségét némileg ellensúlyozni fogja az egységes módszertani szemlélet. A filozófiában és a logikában meglehetősen gyakori, hogy normatív módon közelítenek a dolgokhoz: elítélik azokat a használati formákat, amelyeket a szerző „illogikusnak” tart, és egy olyan ideális nyelvet dolgoznak ki, amely csakis a logikailag következetes használatot támogatja. A normatív szemlélet nagymértékben átszivárog a számítástechnikába is, ahol az ember egy formális nyelvet egy formális nyelvtan megadásán keresztül definiálhat, amelyhez aztán kompozicionális szemantikát rendelhet a szokásos szoftvereszközök segítségével, például a yacc-kal. Minket most azonban egy a természetes nyelvi kifejezésekhez rendelhető működőképes szemantika felépítése érdekel („működőképes” alatt egyszerűen azt értjük, hogy számítógépes programok megírásának alapjául szolgálhat), és az elmélet elsődleges empirikus tesztelési anyagának a tényleges nyelvhasználatot fogjuk tekinteni.

A könyv számos gyakorlatot tartalmaz, amelyek többnyire meglehetősen egyszerűek (a Knuth, 1971 rendszerében a 30. szint alatt vannak), de gyakran meglepően mély következményekkel, mint amilyen például a Schur-lemma. Sok esetben a megoldások meglehetősen könnyen fellelhetők az interneten, vagy akár csak néhány oldallal később, de annak az olvasónak, aki aktív tudást szeretne elsajátítani e kutatási területről, nyomatékosan javasoljuk, hogy saját maga oldja meg a feladatokat. Néhány gyakorlat célja — ezeket egy megemelt $^\circ$ jelöli — az, hogy ellenőrizze, az olvasó megértette-e az éppen tárgyalt anyagot, és a legjobb olvasási terv ezekhez a feladatokhoz az, hogy ezeket a problémákat *nyomban* meg kell oldani, ahogy az olvasó találkozik velük a szövegben, és nem várni velük a szakasz vagy a fejezet végéig. Más feladatok, amelyeket egy megemelt $^\rightarrow$ jelez, olyan anyagra utalnak, amellyel a könyvben nem tudtunk foglalkozni, és gyakran támaszkodnak további olyan ismeretekre vagy előfeltevésekre, amelyek nem tesznek lehetővé egy teljesen egyértelmű választ. Mégis, az olvasót a legjobban az szolgálja, ha nekifut ezeknek a feladatoknak is, egyrészt azelőtt, hogy belepillantana a kötet végén összegyűjtött megoldási útmutatásokba, majd pedig utána még egyszer. A nehezebb gyakorlatokat felemelt * jelöli, és általában a szövegben nem sokkal azután ismertetjük a megoldásukat is. A későbbi fejezetekben mind gyakrabban jelölünk meg gyakorlatokat egy megemelt † szimbólummal, ami azt jelenti, hogy nincs egy egyértelműen jó megoldás, hanem az olvasónak érdemes kísérleteznie a problémával, elsősorban egy formális vagy számítógépes modell építése útján, amely a kívánt tulajdonságokat mutatja. A definíciók és a gyakorlatok számozása a fejezetszámot is

tartalmazza, hogy megkönnyítse a kereszthivatkozásokat és a végén található útmutatások ellenőrzését, de a táblázatok, számok és egyenletek számozása minden fejezetben újraindul.

A nyelvészeti példákat rendszerint a *dőlt betűkkel* adjuk meg, és ha egy jelentés (parafrázis) van megadva, akkor ez félidézőjelek 'jelentésjelek' között jelenik meg. A dőlt betűket a legelső alkalommal bevezetett technikai kifejezésekre és nyomatékosításra is használjuk. A `4lang` számítógépes rendszer tartalmaz egy fogalomszótárat, amely eredetileg négy olyan nyelvet kötött össze, amelyek a nagy európai nyelvcsaládokat reprezentálták: a germánt (angol), a szlávot (lengyel), az újlatint (latin) és a finnugort (magyar). Ma a kötések már több mint 40 nyelven léteznek (Ács, Pajkossy, és Kornai, 2013). Az ebben a szótárban szereplő tételek angol nyelvű nyomtatási képe, valamint más, számítógépes szempontból releváns anyagok `írógép betűtípussal` szerepelnek.

Minden fejezet egy, a további irodalmat részletező szakasszal végződik. Általában azokat a cikkeket és könyveket javasoljuk itt, amelyekben először jelenik meg az adott elképzelés. Mivel az itt tárgyalt témák közül soknak évtizedes, néha pedig évszázados kutatási múltja van, ez a döntés lehet, hogy a könyvet jóval régimódibbnak mutatja, mintha az ellentétes elvet követnénk és csak a legújabb kutatásokra hivatkoznánk. Úgy véljük azonban, hogy az általunk követett irányt nem csak az a szükséglet indokolja, hogy megadjuk a megfelelő elismerést az elképzelések kitalálóinak, hanem az is, hogy a korai munkák gyakran olyan nézőpontokat és meglátásokat szolgáltatnak, amiket a későbbi tárgyalás már magától értetődőnek tekint. (Következetesen kivételt tettünk azonban azokkal a monográfiákkal és tankönyvekkel, amelyek átdolgozott kiadásban később újra megjelentek: ezek mindig a legutolsó kiadásban szerepelnek, mivel azok gyakran jobbak és könnyebben beszerezhetők.) E művek azonban gyakran csak papíron állnak rendelkezésre, és egyre kevesebb diák vagy kutató hajlandó ellátogatni a könyvtárba. Mivel ez a tendencia nyilvánvalóan visszafordíthatatlan, arra törekedtünk, hogy kattintható formában internetes hivatkozásokat adjunk meg, a lehető legnagyobb mértékben elkerülve eközben a jelszóval védett részeit az olyan gyűjteményeknek, mint amilyen a Project MUSE és a JSTOR.

Köszönetnyilvánítás

Az első fejezetekben bemutatott témák közül néhány elsőként a következő három cikkben került publikálásra: Kornai (2010, 2010a, 2012) , és ehelyütt fejezzük ki hálánkat a következő személyeknek, akik még ezeket az első változatokat kommentálták: Beke Tibor (UMass Lowell), Szabó Zoltán (Yale), Terry Langendoen (UArizona), Donca Steriade (MIT), Varasdi Károly (Henrich Heine Universität). A 8. fejezet nagyrészt az először a 2014c valamint a Kornai 2014d cikkekben leírt gondolatokon alapul. Beke Tibornak, Vida Péternek (Mannheim) és Gyenis Zalánnak (MTA Rényi Intézet) tartozunk hálával az e cikkeket illető éles szemű kritikájukért. A munkát részben az OTKA 77476. számú pályázata és a FuturICT.hu (TAMOP-4.2.2.C-11/1/KONV-2012-0013) pályázaton keresztül az Európai Unió és az Európai Szociális Alap támo-

gatta. A könyv megírása részben a Harvard egyetemhez tartozó Kvantitatív Társadalomtudományok Intézetében (IQSS), a Bostoni Egyetem Számítógéptudományi Tanszékén, illetve az egyetem Rafik B. Hariri Intézetében, továbbá a Budapesti Műszaki és Gazdaságtudományi Egyetem (BME) Algebra Tanszékén és a Magyar Tudományos Akadémia Nyelvtudományi Intézetében (MTA NyTI) történt, de a munka legnagyobb része a Számítástechnikai és Automatizálási Kutatóintézetében (MTA SZTAKI) zajlott.

Külön köszönet illeti a könyv korai változatainak alábbi olvasóit, akik sok sajtóhibára és stilisztikai ügyetlenségre mutattak rá, kiváló hivatkozásokat, gyakorlatokat és a gyakorlatokhoz tartozó megoldási útmutatókat javasoltak, és kitűnő tanácsokat adtak számos olyan részletre vonatkozóan, amely csak az ő szakértelmükkel rendelkező ember számára tűnhet fel: Ács Judit (BME AUT), Eric Bach (Wisconsin-Madison), Michael Covington (University of Georgia), Gérard Huet (INRIA), Paul Kay (Berkeley), Marcus Kracht (Bielefeld), Makrai Márton (MTA NyTI), Orthmayr Imre (ELTE Általános Filozófia Tanszék), Pajkossy Katalin (BME), Recski Gábor (BME AUT), Simonyi András (PPKE), Takó Ferenc (ELTE Filozófiatudományi Doktori Iskola), Madeleine Thompson (Empirical Systems), és Zséder Attila (Lensa). Mondanom sem kell, hogy az említett személyek nem értenek egyet mindennel, ami a könyvben van; a könyvben kifejtett nézetek nem feltétlenül tükrözik a támogató szervezetek véleményét, és minden esetleges hiba és mulasztás engem terhel.

A vezető fejlesztők a következők: Borbély Gábor (BME), Makrai Márton (MTA NyTI), Nemeskey Dávid, (MTA SZTAKI), Recski Gábor (MTA AUT), és Zséder Attila. Az itt leírt tananyagot kétszer tanítottuk a BME-n, 2010 tavaszán és 2014 őszén, és itt szeretném kifejezni a hálámat sok résztvevő diáknak, többek között Kara Greenfieldnek (jelenleg az MIT Lincoln Labsnál), Sarah Juddnak (jelenleg a Girls Who Codenál) és Vásárhelyi Dánielnek (MTA NyTI).

Végezetül szeretném kifejezni hálámat a könyv magyar fordítóinak. Varasdi Károly (Előszó és 1. fejezet); Máté András (2. fejezet); Kálmán László (3.-5. fejezet); Makrai Márton (6. fejezet); Gyenis Zalán (7.-8. fejezet); és Takó Ferenc (9. fejezet) hihetetlen türelemmel és precizitással, gyakran az angol eredeti javítani való részeire való rámutatással, hozták létre a magyar szöveget. Ahol ez jól érthető, az az ő érdemük, ahol kevésbé, az az én hibám. A bábeli latex-könyvtár magyar szekciójában Wettl Ferenc (BME) igazított el. Erő Zsuzsa számtalan ponton hozta közelebb a kéziratot a magyar helyesírás bevett szabályaihoz, az itt-ott mégis fennmaradó eltérések a szerző tudatos választásai. A magyar változat nem jöhetett volna létre Votisky Zsuzsa† lelkesedése, bátorítása, és támogatása nélkül. Nem is lehet elhinni, hogy ezt a kötetet már nem vehette kezébe. Non omnis moriar; multaque pars mei vitabit libitinam.

Tartalomjegyzék

1

Bevezetés

Tartalom

Az 1.1. szakaszban azzal kezdjük a munkát, hogy bevezetjük a nyelvi kifejezéseket (szavakat, mondatokat, nagyobb szövegdarabokat) a jelentésükkel összekapcsoló *interpretációs relációt*, majd elhatároljuk egymástól a jelentéstan (szemantika) két fő részterületét, a *lexikai* szemantikát és a *kompozicionális* szemantikát. Mivel a természetes nyelv által közvetített ideák, érzelmek, gondolatok és tények hihetetlenül széles skálája igencsak parttalanná teheti a természetes nyelvi kifejezések szemantikai elemzésének feladatát, sorrendeznünk kell a teendőket. Ezt fogjuk tenni az 1.2. és 1.3. szakaszokban, először az előfordulási gyakoriságokra, majd pedig az információs tartalomra alapozva. Miután tisztáztuk ezeket az alapokat, az 1.4. szakaszban felvázoljuk a könyv általános felépítését, melynek alapján az olvasó kellő megalapozottsággal választhatja ki a neki megfelelő olvasási tervet az 1.5. szakaszban vázoltak közül.

1.1. Kompozicionalitás és kontextualitás

A jelentéstan egészen Platón *Theaitétoszáig* visszavezethető központi ideája az, hogy ismerni egy dolgot nem más, mint képesnek lenni számot adni a dolog részeiről. Az elképzelés modern megfogalmazását rutinszerűen Fregének szokás tulajdonítani (annak ellenére, hogy Janssen (2001) kimutatja, hogy ez bizonyos fokig félreértelmezése Frege vonatkozó gondolatainak) és a Kompozicionalitás elveként szokás rá hivatkozni:

> Egy összetett kifejezés jelentését annak szerkezete és az összetevőinek jelentése határozza meg.

Amit az elv megkövetel, az egy olyan algoritmus, amely képes szintaktikailag elemezni az adott kifejezést — azaz meghatározni a kifejezés összetevőit és szintaktikai szerkezetét — és ezt magukra az összetevőkre is képes rekurzíven alkalmazni, egészen addig, amíg el nem ér a tovább már nem elemezhető atomi egységekhez. Ahhoz, hogy megkapjuk a kifejezés jelentését, tehát egy további erőforrásra is szükségünk van: valamiféle szólistára vagy szótárra (ezt *lexikonnak* fogjuk hívni), ahol az atomi egységek jelentései tárolódnak. Első közelítésben — mert a dolgok ördögien bonyolulttá válnak, amint beleveszünk a részletekbe — azt az elméletet, amely az atomi egységek jelentését írja le, *lexikai szemantikának*, míg azt az eméletet, amely azt írja le, hogy hogyan építhetünk nagyobb egységeket kisebbekből (és megfordítva: hogyan bonthatjuk fel a nagyobb egységeket kisebbekre), *kompozicionális szemantikának* fogjuk nevezni.
A *szemantikai interpretáció* fogalma alatt egy olyan mechanizmust értünk, amely kiszámítja a nyelvi kifejezések jelentését, míg a *generálás* fogalma a fordított eljárást takarja, vagyis azt, amely valamilyen jelentéshez előállít egy jólformált kifejezést, amely éppen ezzel a jelentéssel bír. Ez a két fogalom együttesen alkotja a ⟨kifejezés, jelentés⟩ párokból álló *interpretációs relációt*. Míg a szemantika elméletét leginkább az interpretációs és a generáló algoritmusok technikai részletkérdései foglalkoztatják, nem árt észben tartani, hogy a legfontosabb gyakorlati érdekek sokkal inkább a végeredményhez, mintsem az előállítás folyamatához kapcsolódnak, és már csak emiatt is megérheti a néha hibás eredményt adó rendszereket előnyben részesíteni azokkal szemben, amelyek azon az áron szolgáltatnak hibátlan eredményeket, hogy az idő nagy részében nem hajlandóak semmilyen kimenetet adni, nehogy tévedjenek. Ez a jelenség — amelyet *a hibahajlandóság optimalizálási problémája* (*error-reject tradeoff*) néven ismerünk — minden területen előfordul, nemcsak szemantikai rendszerekben, hanem a beszéd- és optikaikarakter-felismerő, gépi fordító, és általában is az összes, természetes nyelvi be- és kimenettel dolgozó algoritmus esetében megjelenik.
Azt az esetet, amikor különböző $m_1, m_2, \ldots, m_k$ jelentések tartoznak ugyanahhoz az e kifejezéshez, *többértelműségnek* nevezzük, s azt mondjuk, hogy az e kifejezés k-szorosan többértelmű. Azt az esetet, amikor különböző $e_1, e_2, \ldots, e_l$ kifejezésekhez ugyanaz az m jelentés tartozik, *szinonímiának* hívjuk. Mivel az olyan kifejezések, amelyek tökéletesen mentesek minden többértelműségtől, viszonylag ritkák, rendszerint alsó indexek használatával teszünk különbséget a különböző értelmek között: *atléta*$_1$ 'sportoló' és *atléta*$_2$ 'ujjatlan ruházati termék'.
Amikor végtelen számú kifejezéssel van dolgunk (s ez a helyzet mind a természetes nyelv, mind a matematika esetében), a kompozicionalitás elvének valamilyen formában történő használata megkerülhetetlen. Egy számtani példát véve, a $3 \cdot (8+1)$ kifejezésben a zárójelek által kijelölt sorrendben haladunk, és az összeadást a szorzás előtt végezzük el. A zárójelek persze legtöbbször elhagyhatóak, hiszen konvenció szerint a $\cdot$ erősebben köt, mint a $+$. Mivel azonban a természetes nyelvek nem adják meg nekünk a zárójelek használatának kényelmét (habár a mondatlejtés bizonyos formái elég jól megközelítik ezt), a szavak egyszerű egymás után fűzése önmagában is többértelműségek forrása lehet, és a szófüzérhez tartozó szintaktikai elemzési fát is figyelembe kell vennünk,

hogy eldönthessük, melyik a megfelelő *olvasat* (ahogy ezt a jelentéstanban nevezik): '(a turista a hegyen) az antennával' vagy pedig 'a turista (a hegyen az antennával)'. (Ha pedig egy olyan teljes mondatot veszünk, mint amilyen a *Megütötték a turistát a hegyen az antennával*, akkor még egy olvasatot elkülöníthetünk, mivel a *megütni valakit az antennával* szintén értelmes kifejezés.)

Amikor megpróbáljuk elképzelni a fenti kifejezések által leírt helyzeteket, észrevesszük, hogy a *turista az antennával* egy kisebb, kézben hordozható antennát sugall, míg a *hegy az antennával* inkább egy nagy adótorony képét vetíti elénk. Annak a kérdésnek a megtárgyalását, hogy vajon ez a nyilvánvaló méretkülönbség igazolhatja-e az $antenna_1$ és $antenna_2$ különbségtételt, elhalasztjuk egészen a 6.4. szakaszig, ahol is a *monoszémia* elve mellett fogunk érvelni: ez a módszertani elv azt mondja ki, hogy szótári jelentéseket csakis a legvégső esetben szabad szaporítani. Általánosabban fogalmazva, a Frege által a *Kontextualitás elvének* nevezett elv érvényességét nem nehéz kimutatni olyan esetekre való hivatkozással, ahol nem fűződik kétség a többértelműség jelenlétéhez. Vegyük például a *mező* szót, amelynek (legalább) két jelentése van: $mező_1$ 'sík gazdálkodási terület' és $mező_2$ 'a sakktábla egy helye, amelyen egy bábu állhat egyszerre'. Amikor azt mondjuk, *A bástya a király melletti mezőn áll*, nyilvánvalóan a $mező_2$-re gondolunk, és amikor azt mondjuk, *A traktor a falu melletti mezőn áll,* a $mező_1$-re. Még ha egy olyan művész, mint Christo képes is lenne egy egész mezőgazdasági mezőt egy óriási sakktábla részeként bemutatni, ez a lehetőség meg sem fordul a fejünkben a mondat hallatán, hacsak nem vagyunk rá speciálisan előfeszítve. Megfogalmazhatjuk tehát a *Kontextualitás elvét:*

> Soha ne kérdezd, hogy mi egy szó jelentése önmagában, csak azt, hogy mi egy adott mondaton belül.

A jelentésnek a szerkezettől függő eseteiben az indexelésnek mint egyértelműsítési módszernek nem sok értelme van. Még a zárójelek kijelölése is gyorsan értelmét veszti, mivel fontos esetekben a kompozicionális szerkesztés megszakított összetevőket is tartalmazhat: a *fel fogom hívni* kifejezést az igekötős *fel ... hívni* 'telefonálni' és a jövő időt kifejező ragozott *fog* segédige összefésülése hozza létre. Figyelemre méltó, hogy Frege (1879) már észrevette ezt a problémát a Fogalomírás című munkájában, talán mert a jelenség könnyen érzékelhető a német szórend sajátosságai miatt, és speciális kapcsolatképző jeleket is bevezetett a kezelésére. A modern jelentéstanban Bach (1981)-gyel kezdve több eljárást is kidolgoztak e problémakör kezelésére (lásd különösképpen Jacobson (2014) 5.5. szakaszában és utána), azonban egy igazán kifejező jelölés kidolgozása még várat magára.

Amennyiben csupán véges sok kifejezéssel van dolgunk — mint például az európai közlekedési táblák vagy a kínai írásjelek esetében — akkor a lexikon önmagában is teljesen elegendő lehet, habár még ezekben az esetekben is hasznos a könnyebb felidézés elősegítésére a jeleket összetevő részekre elemezni és megjegyezni néhány egyszerű szabályt; például hogy a figyelmeztető táblák háromszög, míg a tiltó táblák kör alakúak. Az ilyen *redundancia-szabályok* azonban nem kompozicionálisak, hiszen nem létezik

például olyan közlekedési tábla, mint amit a 1.1. ábra mutat, azzal a különbséggel, hogy

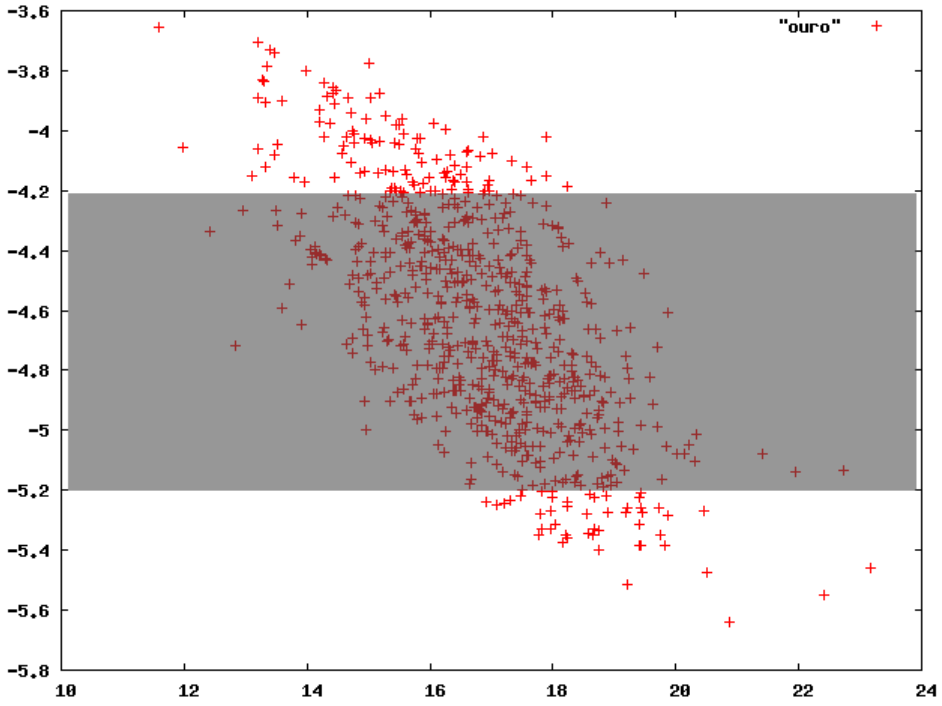
1.1. ábra Egy nem kompozicionális jel

kör és nem háromszög alakú. Valaki persze mondhatná, hogy elvileg egy ilyen közlekedési tábla létezhetne, és ha tényleg létezne, akkor azt jelentené, hogy 'gyerekkel tilos itt lenni', azonban ez meglehetősen furcsa kiindulási pont lenne bármilyen elmélethez.

1.2. A témaválasztásról

A legtöbb olvasó valószínűleg jól ismeri a Zipf-eloszlás fogalmát, ami kvantitatív formában azt állítja, hogy a szavak, szópárok (bigramok), és általánosan is, a szó n-esek eloszlása a hatványtörvény szerint alakul.
Ami talán kevésbé ismert, az a Zipf (1935)-ben tett kvalitatív megfigyelés, miszerint a leggyakoribb szavak egészében véve pont a legalapvetőbb, legelemibb nyelvi egységek, amelyeknek a története ráadásul kitűnően ismert. Zipf nem tudta számszerűen alátámasztani ezt a megfigyelését, mivel nem állt rendelkezésére egy formális jelentéselmélet, de amint rendelkezésünkre áll egy ilyen elmélet, máris felhasználhatjuk a bonyolultság és gyakoriság közötti negatív korrelációt a vizsgálat tárgyának kiválasztására: elsőként a legegyszerűbb és ezért leggyakoribb esetekkel kell foglalkoznunk, s csak akkor haladhatunk tovább a bonyolultabb és ezért ritkább esetek felé, amikor a gyakoribbakkal már elszámoltunk. Meg kell tanulnunk mászni, mielőtt megtanulunk járni, és ebben a könyvben gyakran célzatosan el fogunk tekinteni a futás és a korcsolyázás jelentette kihívásoktól, amíg tökéletesen el nem sajátítottuk az egyszerűbb mozgásformákat.

Az 1.2 ábra, amely itt még csak pusztán illusztrációként szolgál, a szavak bonyolultságát ábrázolja — amit hozzávetőlegesen a szó szótári meghatározásának hosszával tekintünk arányosnak (a pontosabb definíciót ld. a 202. oldalon 6.3 definíció alatt) — a szógyakoriság logaritmusának függvényében. Maguk a szavak a Longman Defining Vocabularyból kerültek ki, míg a szógyakoriságok a Google 1T korpuszból. Bár a bonyolultság és a log-gyakoriság közötti negatív korreláció szépen látható, ahhoz, hogy kvantitatív módon kezelni tudjuk Zipf feljebb tárgyalt kvalitatív megfigyelését, szükségünk lesz még jó néhány eszközre, egyrészt annak érdekében, hogy kiterjeszthessük a diagramot az alapszókincsen túlra, másrészt pedig hogy a bonyolultsági mértéket szavakról egész *konstrukciókra* általánosíthassuk — mind a két feladattal foglalkozni



1.2. ábra A szókincs bonyolultsága a gyakorisággal összevetve

fogunk a könyvben. A szürke sáv alatti terület, amely csak a számottevő gyakorisággal (több, mint 10m) rendelkező szavakat tartalmazza, analóg azzal, amit a Google „aluláteresztő szemantikának” (low pass semantics) nevez (ld. Pereira 2012).
Ebben az alsó tartományban (az „alul” a bonyolultság skáláján értendő, nem a gyakoriságén) a rendszert a valóság dolgaihoz és azok tulajdonságaihoz kapcsoljuk. Az ilyen entitások jelenleg elérhető legnagyobb nyilvános strukturált gyűjteménye, a Freebase, bőven 45 millió feletti számú entitás csomópontját tartalmazza (ezeket *topiknak* hívják), és körülbelül 1,9 milliárd *tényt* tárol az entitások közötti címkézett élek formájában, olyanokat mint pl. *Foglalkozás(Leonardo, matematikus)*. Az ilyen tényeknek a természetes nyelvi szövegekből való kivonatolása (ld. reláció-kivonatolás), illetve hogy hogyan dönthető el, hogy két nyelvi kifejezés, mint például *Leonardo* és *da Vinci* vajon ugyanarra a valóságbeli entitásra vonatkozik-e (ld. koreferencia-feloldás), az aluláteresztő szemantika központi kérdései közé tartoznak, és emellett olyan nagyszámú, jelenleg futó kutatási projekt tárgyát képezik, amelyeket a szabványosított adathalmazok és kiértékelési eljárások vagy az ezekkel kapcsolatos *nyilvános versenyfeladatok* (shared tasks) elérhetősége tesz lehetővé. Összességében azt mondhatjuk, hogy felszíni módszerek segítségével is már igen sok esetet helyesen tudunk kezelni, talán úgy 80%-ot, de ahhoz nem kevés világismeretre van szükség, hogy rájöjjünk, hogy a matematikus Leonardo az nem más, mint Leonardo Fibonacci és nem a magától értetődőnek tűnő Leonardo da Vinci.