

Pécsi Tudományegyetem
Bölcsészet- és Társadalomtudományi Kar
Nyelvtudományi Doktori Iskola
Alkalmazott Nyelvészeti Doktori Program

Hoss Alexandra

Nyelvtechnológiai eszközök és eljárások alkalmazhatósága autizmusban

Doktori (Phd) értekezés tézisei

Témavezető: dr. Viszket Anita, PhD



Pécs
2022

Tartalomjegyzék

1. Bevezetés, a dolgozat felépítése	2
2. Az autizmus spektrum elméleti megközelítése.....	3
2.1 Az autizmus korszerű fogalma, epidemiológiája, etiológiája és a komorbid állapotok ..	4
2.2 Az autizmus kognitív (pszichológiai) magyarázatai.....	6
2.3 Az autizmus viselkedési képe és a nyelvi, nyelvhasználati, illetve kommunikációs sajátosságok	8
3. Számítógépes nyelvészet, természetesnyelv-feldolgozás, nyelvtechnológia.....	10
3.1 A természetesnyelv-feldolgozás (rész)feladatai	10
3.2 A nyelvtechnológia módszerei és a nyelvi modellek.....	12
3.3 Alkalmazási területek.....	14
4. Autizmus és természetesnyelv-feldolgozás: szisztematikus áttekintés.....	16
5. A kutatás	19
5.1 A kutatás célja, hipotézisei, módszere	19
5.2 Eredmények	21
5.3 Az eredmények megvitatása	22
5.4 Kiegészítő vizsgálatok, kiegészítő magyarázatok az eredményekhez.....	24
6. Összegzés.....	28
Hivatkozások.....	30

1. Bevezetés, a dolgozat felépítése

A robbanásszerű technikai fejlődés, az információ, az adatok és szövegek mennyiségének növekedése és elérhetősége révén jelentős mértékű fejlődés tapasztalható a számítógépes nyelvészet, a természetesnyelv-feldolgozás területén és a nyelvtechnológiai eszközök és eljárások létrehozásában. Ez egyúttal azt is eredményezi, hogy a rendelkezésre álló eszközök és eljárások korábban viszonylag szűk körű felhasználása mindinkább kiszélesedik, hozzáférhetőségük növekszik. A mesterséges intelligencia, a neurális hálók és a mélytanulás területén az elmúlt egy-két évben olyan mértékű előrelépés történt, ami a számítógépes nyelvészeti, nyelvtechnológiai megoldásokban, a nyelvmodellekben és az alkalmazási lehetőségekben egyaránt jelentős előrelépést és változást idézett elő: az új, mind hatékonyabb és jobban teljesítő architektúrák révén a nyelvmodellek is egyre finomodtak, egyre több, korábban komoly nehézséget okozó részfeladatban nyújtva javuló teljesítményt. A különféle eszközök és eljárások használata és finomhangolása szintén egyszerűsödött. Mindezek egyre inkább ráirányítják a figyelmet a nyelvtechnológiai eszközök klinikai és pedagógiai alkalmazhatóságára, mint például Hoffmann és munkatársai (2017) eljárása, ami az enyhe kognitív zavar beszéd alapú felismerésének eljárását mutatja be, vagy a komplex kommunikációs igényű személyek gördülékenyebb kommunikációjának elősegítését célzó PictoBERT (Pereira és mtsai, 2022) alkalmazás.

Ma már nagy számban állnak rendelkezésre olyan nemzetközi szakirodalmak és kutatási eredmények, amelyek a depresszióval, a szorongással vagy az öngyilkossággal összefüggő nyelvi markerek feltárásával, elemzőrendszerek és automatizált szöveganalitikai eszközök implementálásával és alkalmazásával foglalkoznak. Klinikai területen a nemzetközi kutatásokban régóta alkalmaznak nyelvtechnológiai eljárásokat és eszközöket: a pszichológiai, pszicholingvisztikai indíttatású vizsgálatokban az egyik ilyen gyakran használt eszköz a *Linguistic Inquiry and Word Count* (Boyd és mtsai., 2022; Pennebaker és mtsai, 2007). A mentális egészséggel, pszichés zavarokkal kapcsolatos tematika a hazai kutatásokban, tudományos érdeklődésben is jelen van: megjelenik például a Magyar Számítógépes Nyelvészeti Konferencia tematikájában.

Disszertációmban – az előbbieken ismertetettekhez kapcsolódva – azt a kérdéskört járom körül, hogy a jelenleg rendelkezésre álló nyelvtechnológiai eszközök és eljárások bármelyikét alkalmazták-e az autizmus spektrum vizsgálatában, specifikusabban azt, hogy alkalmazásukkal az autizmussal élő személyek nyelvi, nyelvhasználati sajátosságairól mit tudhatunk meg így. A dolgozat exploratív jellege azt a célt is szolgálja, hogy a feltárt ismeretek,

eredmények mentén a nyelvészeti megközelítésen túl a klinikai hasznosíthatóság kérdéskörét megvizsgáljam: megalapozva egy jövőbeli kutatás elméleti kereteit és kiindulópontját. A céllal összhangban a dolgozat 2. fejezetében az autizmussal kapcsolatos korszerű, bizonyítékokon alapuló ismereteket és tudást tekintem át az állapot komplex, heterogén és spektrum jellegének megismerése, megismertetése érdekében az autizmus többszintű értelmezésének modellje mentén, ami nemcsak a tudományos, hanem a klinikai gyakorlatban is releváns. A 3. fejezet rövid betekintést nyújt a számítógépes nyelvészet, természetesnyelv-feldolgozás és nyelvtechnológia területébe, kitérve egy-egy konkrét példán (alkalmazáson) keresztül az alkalmazási területekre, valamint a dolgozat pilot-vizsgálatában a gépi elemzéshez használt NarrCat (Ehmann és mtsai, 2014) rendszer ismertetésére. A 4. fejezet egy szisztematikus áttekintés folyamatát és eredményeit mutatja be: a dolgozat kérdésfeltevéséhez illeszkedve azokat az információkat, amelyek a nyelvtechnológiai eszközök és eljárások alkalmazásával és az autizmus spektrum állapot nyelvi, nyelvhasználati sajátosságaival kapcsolatosak. A dolgozat 5. fejezetben egy kismintás, esettanulmány jellegű saját kutatást ismertetek, amelyben a 2. és 4. fejezet szakirodalmi alapján megfogalmazott hipotézisek teszteléséhez gépi elemzést (NarrCat) alkalmazok, problémafeltáró jelleggel kiegészítve kézi elemzéssel. Az összegzésben (6. fejezet) az előző fejezetek alapján megfogalmazható tanulságokat és következtetéseket foglalom össze, amelyek meglátásom szerint egy jövőbeli kutatás megtervezésének kiindulópontjaként szolgálhatnak.

2. Az autizmus spektrum elméleti megközelítése

A fejezet célja, hogy röviden bemutassa az autizmussal kapcsolatos jelenlegi konszenzuális tudást. Ennek érdekében először ismertetem az autizmus fogalmának változását az első leírásoktól, a medikális szemlélettől a napjainkban zajló paradigmaváltásáig, amelyben a participativitás és a neurodiverzitás elmélete és szemléletmódja egyaránt fontos (Fletcher-Watson, 2019; Pellicano, 2020). A többszintű modell logikáját követve előbb a már bizonyított biológiai háttérrel (etiológia és komorbid állapotok, epidemiológia sajátosságok), majd a kognitív pszichológiai magyarázatok kérdéskörét és végül pedig a viselkedéses szinten megnyilvánuló sajátosságokat mutatom be. A még napjainkban is vezető magyarázatként szolgáló kognitív hipotézisek (tudatelméleti, végrehajtó funkciók, gyenge centrális koherencia) ismertetésekor törekedtem a viselkedéses képpel, hangsúlyosan pedig a nyelvi, nyelvhasználati sajátosságokkal összefüggő bemutatásukra annak érdekében, hogy megalapozzam egyrészt az

5. fejezetben leírt kismintás vizsgálatom hipotéziseit, másrészt egy jövőbeli multidiszciplináris kutatás elméleti áttekintését vagy legalábbis annak kiindulópontját.

2.1 Az autizmus korszerű fogalma, epidemiológiája, etiológiája és a komorbid állapotok

Az autizmus jelenlegi korszerű fogalmának és szemléletmódjának kialakulásához számos tényező járult hozzá: így az autizmus és az autista személyek és családjaik iránti élénk tudományos és közéleti érdeklődés, az autista személyek érdekvédelmi tevékenysége által megvalósuló társadalmi szemlélet- vagy tudásformálás, az önérvényesítés megerősödése vagy a fogyatékoságtudományban bekövetkező paradigmaváltás. A szemléletmód változása és a „Semmit rólunk nélkülünk!” elv érvényesülése révén a személyek hatással lehetnek a társadalomban, kultúrában magukról kialakított képre. A paradigmaváltás és az azzal együttjáró társadalmi változások a tudomány területén is éreztetik hatásukat: aktívan alakítják az autizmusról való gondolkodást, az autizmus fogalmának meghatározását és a kutatások irányát, tematikáját.

Az autizmus spektrum¹ jelenlegi korszerű definíciója és megközelítése mint a fejlődés egy lehetséges, sajátos változatát határozza meg az autizmust, amiben a tudományos ismeretek és az érintett személyek állaspontjának megjelenése egyaránt megfigyelhető. A kezdeti leírásokban is megjelenő medikális szemléletű megközelítés az autizmus megismerése és megértése szempontjából ugyanakkor megkerülhetetlen, hiszen hagyományosan gyermekpszichiátriai diagnózisról van szó, valamint a tudományosan bizonyított eredményeken alapuló leírásokkal ez a medikális szemlélet megteremtette az autizmus-tudományt (Pellicano & den Houting, 2022). Bár az autizmus mai fogalmának kialakulását az utóbbi néhány évtized intenzív kutatási eredményei, empirikus adatai és tapasztalatai határozták meg, az autizmus mégsem tekinthető modernkori jelenségnek (bővebben lásd például Frith, 1991).

Lorna Wing (1996) írja le a ma már tudományosan igazolt, a klinikai és pedagógiai gyakorlat tapasztalataiban is realizálódó spektrum-felfogást, ami a jelenleg érvényben lévő DSM-5 (APA, 2013) klasszifikációs rendszerében a nomenklatura szintjén is hivatalosan bevezetésre került. Ehhez terminológiájában és szemléletmódjában is illeszkedik a hazánkban a szakértői bizottságokban és az egészségügyben egyaránt 2022 óta érvényben lévő BNO-11 (WHO, 2018, 2022). Mindkét kategóriarendszer az idegrendszer fejlődési zavarai közé sorolja az autizmust, valamint korábbi verziójukhoz képest megújult szemléletet képviselnek például

¹ A dolgozatban az *autizmust* az *autizmus spektrum*, *autizmus spektrum zavar*, *autizmus spektrum állapot* kifejezések szinonimájaként használom, habár az egyes megfogalmazások különböző preconcepciókat és konnotációkat sejtetnek.

az alkategóriák (támogatási szükséglet, nyelvi és intellektuális képességek színvonala, valamint az adaptív viselkedések minősége) vagy az állapot megközelítése (diádikus, multidimenzionális spektrum) kapcsán.

Az erőteljes gyermekpszichiátriai motiváltság megjelenik az autizmus mai konszenzuális meghatározásában: ez egy olyan neurokognitív fejlődési zavarként definiálja az autizmust, ami az etiológia, a kognitív háttér és a viselkedéses kép tekintetében egyaránt nagyfokú heterogenitást és komplexitást mutat (EMMI, 2020; Győri, 2012). Élethosszig fennálló állapotról van szó, amelyben már a korai időszakról kezdve azonosítható a fejlődés menetét eltérő volta. A spektrum-felfogás a több területen jelentkező heterogenitásból és a dinamikusan változó klinikai kép multidimenzionális értelmezéséből fakad, amelyben (legalább) hét fő dimenzió mentén ragadhatók meg az inter- és intraindividuális eltérések (Csepregi & Stefanik, 2012; EMMI, 2020; Stefanik & Prekop, 2015). A DSM-5 (APA, 2013) és a BNO-11 (WHO, 2018) az eltérő fejlődési mintázatokat két fő viselkedéses területen azonosítja: (1) a kölcsönös szociális interakciók és reciprok kommunikáció tartósan fennálló, több kontextusban jelen lévő minőségi eltéréseiben (az interakciók kezdeményezésének és fenntartásának állandóan jelen lévő minőségi eltérése), valamint (2) a rugalmas viselkedésszervezés nehézségeiben, repetitív viselkedésmintázatokban és az érdeklődés vagy tevékenységek szűk körű, repetitív, minőségében eltérő voltában. Ez a két fő viselkedéses területhez kapcsolódó tünetcsoport az „autisztikus diád”, ami lehetővé teszi az autizmus koherens értelmezését. A korszerű diád-szemlélet ugyanakkor legitimizálja a szociális és a kommunikációs területek szükségszerű és evidens egybefonódását, a szociális-kommunikációs nehézségek egységes szemléletét.

Az autizmussal kapcsolatos epidemiológiai adatok alakulása egyértelmű összefüggést látszik mutatni a folyamatosan gyarapodó ismereteken alapuló fogalom- és szemléletmód-változással, valamint az állapot természetének megismerésével. Jelenleg az utóbbi évtizedek nemzetközi kutatási adatai alapján az autizmus előfordulási gyakorisága meghaladja az 1 százalékot, és folyamatos növekedést mutat 1960-as és 1970-es évekből származó kezdeti adatokhoz képest (EMMI, 2020; Fombonne, 2020; Stefanik & Prekop, 2015). Még napjainkban is egyértelműen tapasztalható, hogy a fiúk, férfiak esetében közel négyszer gyakoribb az autizmus előfordulása (EMMI, 2020; Fombonne és mtsai, 2022; Rivet & Matson, 2011).

Az autizmus biológiai meghatározottságát támogató faktorok közé tartoznak: a korábban említett nemi különbségek, a társuló komorbid állapotok, valamint a biológiai háttér meglétét igazolják a különböző iker- és családvizsgálatok eredményei is (EMMI, 2020; Győri, 2012; Rutter & Thapar, 2014; Stefanik & Prekop, 2015). Az autizmus az egyik

legerőteljesebben genetikailag determinált (gyermek)pszichiátriai állapot, komplex genetikai háttérmechanizmusokkal és többnyire poligénes öröklődésmenettel (EMMI, 2020; Győri, 2012; Rutter, 2011). Több évtizede próbálják azonosítani az autizmushoz kapcsolódó környezeti faktorokat (mint például a magasabb apai életkor, pre- és perinatális tényezők) is, hiszen az ikervizsgálatokban sem találtak 100 százalékos konkordanciát (Karimi és mtsai, 2017; Lyall és mtsai, 2014).

Az autizmusra jellemző neurobiológiai sajátosságok között feltártak már jellegzetes, komplex és heterogén mintázatokat az idegrendszer fejlődésében és szerveződésében egyaránt. Ezek között vannak fokális, globális és konnektív jellemzők (Casanova, 2014; Győri, 2012).

Az autizmushoz társuló leggyakoribb állapotok az intellektuális képességzavar mellett az epilepszia, a fragilis X szindróma, a sclerosis tuberoza és bizonyos pszichiátriai zavarok (hangulati és szorongásos), a figyelemhiányos hiperaktivitás zavar, az alvással vagy étkezéssel kapcsolatos zavarok, illetve főként felnőttkorban az OCD (obsessive-compulsive disorder, vagyis obszesszív-kompulzív zavar), a bipoláris zavar vagy a skizofrénia (EMMI, 2020, Volkmar és mtsai., 2014). Bishop (2014) a nyelvi zavar kapcsán veti fel annak lehetőségét, hogy mint komorbid állapot megjelenhet autizmusban.

2.2 Az autizmus kognitív (pszichológiai) magyarázatai

Az autizmus természetének és viselkedéses képének mélyebb megértéséhez a kognitív pszichológia a megismerés területén nyújt adekvát magyarázatokat, ami nélkülözhetetlen a diagnosztika és a beavatkozás, a klinikai és pedagógiai eljárások, módszerek, eszközök megtervezéséhez, megvalósításához. Emellett nemcsak a kutatások tárgyaként, hanem lehetséges interpretációs keretként is szolgálhat a mélyebb megismerés és a megértés szempontjából. Érvényes ez a kommunikációs és nyelvi sajátosságok területére is, ahogy az a dolgozat 5. fejezetében megjelenik a gépi és kézi elemzéskor vizsgált kategóriákban. A kezdeti törekvések ellenére egyik kognitív elmélet kapcsán sem sikerült kellően meggyőző bizonyítékokat felsorakoztatni annak primér volta mellett. Sokkal komplexebb a feltételezhető kognitív pszichológiai háttér, így a korábban versengő elméletek ma már egymás mellett, egymást kiegészítve nyújtanak magyarázatokat

Az autizmus természetének és a viselkedéses képének megértése szempontjából a *naiv tudatelméleti képesség* (theory of mind, röviden ToM) az egyik jelentős magyarázóerővel bíró hipotézis alapja. A tudatelmélet egy olyan innáta alapokon nyugvó, intuitíven, automatikusan, akarattól függetlenül, gyorsan, erőfeszítés nélkül működő humánspecifikusnak tekintett képesség, amely mentális állapotok tulajdonítását teszi lehetővé másoknak és saját magunknak (Győri, 2003). A mentális állapotok (vélekedések, vágyak, szándékok, érzelmek) attribúciója

segít a különféle társas helyzetek, így a kommunikációs interakciók megértésében, magyarázatában, mások viselkedésének bejósolásában. Lehetővé teszi az adott szociális szituáció és interakció értelmezését, ezáltal az annak megfelelő viselkedést és a másik fél szempontjait is figyelembe vevő reakciók, válaszok kialakítását (Frith, 2002; Baron-Cohen & Bolton, 2000). A tudatelméleti képesség vizsgálatában alkalmazott feladatok többnyire a mentális állapotokkal kapcsolatos kifejezéseket tartalmazzak, ezáltal a nyelvi képességek (elsősorban a megértés) szintjétől nem függetleníthetők. Még azoknál a feladatoknál sem, amelyek nem direkt módon a nyelven keresztül kívánják feltérképezni a mentalizáció működésének sajátosságait. Éppen ezért az egyik általánosan megfogalmazott kritika ezekkel a feladatokkal kapcsolatban, hogy igénylik a nyelvi kompetenciát, sokszor még a nem verbális változatok is (Korkmaz, 2011). A tudatelméleti képesség vizsgálatában a nyelvi kompetenciára direkt módon és erőteljesen épít például Happé (1994b) és Tager-Flusberg (1992) vizsgálata.

A tudatelméleti hipotézissel kapcsolatban felmerülő, elsősorban a viselkedésszervezés területét érintő kérdések vezettek a *végrehajtó működések* eltérésének vizsgálatához autizmusban (Györiné, 2005; Ozonoff, South & Provençal 2005). Ez a rendkívül komplex, prefrontális kérgi területekhez kötődő kognitív képesség-halmaz teszi lehetővé a célorientált, adaptív viselkedés megtervezését, megszervezését, fenntartását és kivitelezését, szükség esetén rugalmas módosítását a környezet folyamatos monitorozásával, valamint az ehhez szükséges kognitív mechanizmusok működtetését, összerendezését (Györi, 2008). A célirányos, adaptív működés érdekében a végrehajtó rendszer irányítja, felügyeli és összerendezi a jövőbeli cél elérése érdekében megvalósítandó viselkedéseket, valamint az azok kivitelezéséhez szükséges kognitív folyamatokat (Györi, 2014). Ez a több komponensű (például tervezés, prepotens válasz gátlása, kognitív flexibilitás, szeriális szervezés, munkaemlékezet, monitorozás) problémamegoldó rendszer minden komplex emberi viselkedés, a viselkedéses és gondolati rugalmasság és mindezek adaptív kivitelezésének a háttérben fellelhető, így a nyelvi és kommunikációs megnyilatkozásokéban is. A végrehajtó működések eltéréseivel kapcsolatos hipotézis 1991-től vált igazán az autizmussal kapcsolatos kutatások és a kauzális modell lényegi elemévé Ozonoff és munkatársai (1991a,b) tevékenysége nyomán.

A *gyenge centrális koherencia* magyarázatát az erősségekkel, az egyenetlen képességprofilal, illetve a szenzoros észlelés és feldolgozás területén tapasztalható eltérésekkel kapcsolatos megfigyelések igénye motiválta, amikre az előbbi két magyarázat csak kevésbé, esetlegesen vagy részlegesen tudott magyarázatot szolgáltatni (Frith, 1991). A centrális koherencia olyan feldolgozási mód, tendencia vagy kognitív stílus, ami a különböző beérkező részinformációkat az adott kontextusnak megfelelő jelentésteli egészekké szervezi

(Frith, 1991; Pellicano, 2011). A hétköznapi társas és kommunikációs szituációi és interakciói alapvetően kontextusérzékeny megértésen és viselkedésen alapulnak: ez igaz a nyelvi képességgel összefüggő aspektusokra is. A nyelvi képességre fókuszálva a koherenciateremtés gyengeségének hipotézisében rejlő magyarázóerőt vizsgálja például Frith és Snowling (1983), Happé (1997), Jolliffe és Baron-Cohen (1999) és Snowling és Frith (1986) tanulmánya.

2.3 Az autizmus viselkedéses képe és a nyelvi, nyelvhasználati, illetve kommunikációs sajátosságok

A viselkedéses kép áttekintésekor az érvényben lévő klasszifikációs rendszerek, azaz a DSM-5 (APA, 2013) és a BNO-11 (WHO, 2018, 2022) által képviselt szemléletet, az ún. autisztikus diád megközelítést követtem. Ez az eltérő fejlődési mintázatokat két fő viselkedéses területen azonosítja: (1) a kölcsönös szociális interakciók és reciprok kommunikáció tartósan fennálló, több kontextusban jelen lévő minőségi eltéréseiben (az interakciók kezdeményezésének és fenntartásának állandóan jelen lévő minőségi eltérése), valamint (2) a rugalmas viselkedésszervezés nehézségeiben, repetitív viselkedésmintázatokban és az érdeklődés vagy tevékenységek szűk körű, repetitív, minőségében eltérő voltában.

Az autizmussal összefüggő fejlődési eltérések már korai életkortól jelen vannak, bár sokszor csak akkor manifesztálódnak, amikor a személy a képességszintjét meghaladó elvárásokkal szembesül (WHO, 2018). Az autizmusra jellemző tünetek legmarkánsabban a 4-5 éves kör közötti időszakban jelentkeznek, éppen ezért kiemelt jelentőséggel bír mind hossz-, mind keresztmetszeti szempontból az ebben az életkori sávban tapasztalható viselkedéses kép megismerése (Csepregi & Stefanik, 2012). 5 éves kor után a viselkedéses kép egyre divergensebbé, heterogénebbé válik. A serdülő és felnőtt kor viselkedéses képe átfed a korábbiakkal, ám a fiziológiai változások, társas-társadalmi elvárások változása, fokozódása sok esetben másodlagos, következményes állapotok kialakulásához vezethet (Csepregi & Stefanik, 2012). Az életkor előrehaladtával, fiatal felnőtteknél alapvetően javulás figyelhető meg az elemi, formális szociokommunikációs készségekben, bár néhány esetben (kisebb számban) romlás következik be (Győriné, 2005). A tanult kompenzációs és megküzdési stratégiák segítenek abban, hogy bizonyos helyzetekben a személy nehézségei rejtve maradjanak (Chawarska és mtsai., 2014; Csepregi & Stefanik, 2012; Howlin, 2014). Jelenleg az idős korra vonatkozó viselkedéses képet leíró ismeretek, az idős korról való tudás még nagyon korlátozott (Howlin, 2014).

Az autizmus nyelvi, nyelvhasználati és kommunikációs sajátosságaira vonatkozóan megállapítható, hogy a dinamikusan változó klinikai kép és a csoporton belüli variabilitás

ellenére a kommunikációban, illetve általában a nyelv kommunikatív használatában, valamint pragmatikai és sokszor szemantikai aspektusaiban feltárhatók olyan eltérések, amelyek sok esetben a diagnosztika szempontjából is relevánsak, miközben a nyelv más strukturális, formális aspektusaiban, a szintaxisban, a morfológiában és a fonológiában relatív érintetlen képességek mutatkoznak, azaz a szűkebb értelemben vett (klinikai) nyelvi zavar nem általános sajátossága az állapotnak (Csepregi & Stefanik, 2012; Győri, 2014; Kim és mtsai, 2014; Tager-Flusberg, 2000, Wetherby, 2008). Nincs egy, az autizmusra jellemző nyelvi és kommunikációs profil, univerzálisnak tekinthető lista a kommunikációs viselkedésekre nézve, ahogy valid alcsoportok létrehozása és megkülönböztetése sem lehetséges a jelenleg rendelkezésre álló kommunikációs és nyelvi adatok alapján: így például a beszélő és nem beszélő csoportokra való redukció, megkülönböztetés is problémás lehet, hiszen ezekben az esetekben is jelentős a heterogenitás, ami megnyilvánul például a kommunikációs viselkedések nagyon eltérő mintázataiban. A nyelvi képességek mellett egyéb dimenziók, úgy mint például az életkor, a személyiség vagy intellektuális képességek, tovább árnyalják az eltéréseket, azaz függetlenül a beszéd színvonalától a kommunikációs sajátosságok rendkívül változatos és sokszínű variációi állhatnak fenn (EMMI, 2020; Győri, 2008; Győri, 2014; Kim és mtsai, 2014; Tager-Flusberg, 2000).

A nyelvi képesség alapján a spektrum egyik felén elhelyezkedő *nem vagy alig beszélő* (ún. *minimally verbal*) személyektől a jó nyelvi képességekkel rendelkező, nyelvi zavart nem mutató, fluensen beszélő, a formális nyelvet késés nélkül elsajátító személyekig terjed a nyelvi képességek színvonala (Tager-Flusberg, 2006). Az autizmussal élő személyek szűkebb csoportját, körülbelül 30 %-át alkotó nem vagy alig beszélő személyek és komplex kommunikációs igényeiknek a megismerése mind a mai napig alulreprezentáltak a kutatásokban (Tager-Flusberg & Kasari, 2013). Az autizmussal kapcsolatos kutatásokban főként az ún. magasan funkcionáló autizmussal élő személyek vesznek részt. Ez egyrészt a klinikai és gyakorlati szempontok mentén, a beavatkozás és támogatás oldaláról közelítve nem előnyös és nem is célravezető, másrészt összefügg a nyelvi profil megalkotásának problematikájával (Győri, 2014). Kim és munkatársai (2014) szerint ugyanakkor ez nem jelent problémát, mert komplex kommunikációs igényű személyeknél feltételezhetően hasonló sajátosságok érvényesülnek. Az egységes profil alkotását, meghatározását tovább árnyaló tényező még például, hogy bármely intellektuális képességszint mellett fennállhat autizmus (WHO, 2018, 2022).

Éppen ezért a dolgozat kérdésfeltevésében, azaz a nyelvtechnológiai eljárások alkalmazásának és alkalmazhatóságának vizsgálatában sem egy egységes nyelvi profil

felállítására kell törekedni, sokkal inkább olyan sajátosságok feltárására, amelyek például a szűrés, a diagnosztika területén hasznosítható, objektíven mérhető paramétereket nyújthatnak, mint például a lexikon különböző szemantikai kategóriájába tartozó szócsoportok kvantitatív eltérései. Az objektív mérést nemcsak a számszerűség, a statisztikai alapok, hanem az automatizálás, így a nyelvtechnológiai eljárások bevonása is lehetővé teszi. Ezeknek az eljárásoknak az alkalmazása emellett azért is előnyös lehet, mert viszonylag gyors eredményeket, információkat nyújthatnak a vizsgált aspektusokról.

3. Számítógépes nyelvészet, természetesnyelv-feldolgozás, nyelvtechnológia

A számítógépes nyelvészet, a természetes nyelvfeldolgozás és a nyelvtechnológia kifejezések, habár gyakran szinonimaként jelennek meg, a nagymértékű átfedés ellenére sem tekinthetők ekvivalensnek: az egyes terminusok sajátos jelentéstartalommal is bírnak. Ehhez illeszkedve a disszertáció 3. fejezete rövid betekintést nyújt a számítógépes nyelvészet, a természetes nyelvfeldolgozás és nyelvtechnológia területébe, kitérve egy-egy konkrét példán (alkalmazáson) keresztül az alkalmazási területekre, valamint a dolgozat pilot-vizsgálatában a gépi elemzéshez használt NarrCat (Ehmann és mtsai, 2014) rendszer ismertetésére. A vázlatos történeti áttekintést követően a nyelvtechnológiai (rész)feladatok, valamint azoknak a pszicholingvisztikai perspektívából történő meghatározását mutattam be, kitérve az eljárások és alkalmazások teljesítményének kiértékelésében meghatározott két fő célra, a *fedés (recall)* és a *pontosság (precision)* maximalizálására, valamint az ezek mentén meghatározott F-mértékre. Ez a különböző nyelvtechnológiai megoldások (ki)értékeléséhez alkalmazott teljesítménymutató nemcsak technológiai, hanem például a klinikai alkalmazhatóság szempontjából is rendkívül fontos és informatív. A mutató a távlati célként megjelenő autizmus spektrumra irányuló, nyelvtechnológiai eszközökkel támogatott objektív, kvantifikált szűrés és diagnosztika esetében is kiemelt jelentőségű a validitás és reliabilitás tekintetében.

3.1 A természetesnyelv-feldolgozás (rész)feladatai

Annak érdekében, hogy a mindennapokban használt alkalmazások megfelelően működjenek, különböző természetesnyelv-feldolgozási (rész)feladatokat kell sikeresen megoldani. Ilyen, a szöveg normalizálásával és előfeldolgozásával kapcsolatos (rész)feladatok például a tokenizálás, az ún. stop-szavak eltávolítása, a lemmatizálás, a part-of-speech taggelés vagy a névelem-felismerés, aminek magyar nyelvi vonatkozásairól részletes áttekintés nyújt Simon (2013). További, a természetesnyelv-feldolgozáshoz kapcsolódó feladat a koreferencia feloldása, a szintaktikai elemzés vagy a szentiment- és érzelemelemzés (Jurafsky, & Martin é.n). Az egyes feladatok végrehajtására nincsenek univerzális, minden nyelven ugyanúgy

használható nyelvtechnológiai eljárások vagy eszközök: nyelvfüggő megoldások léteznek. Magyar nyelvre rendelkezésre álló eszközök például az e-magyar vagy emtsv (Simon és mtsai, 2020), a huSpaCy (Orosz és mtsai, 2022). Az előbbieken ismertetett nyelvtechnológiai feladatok kissé másként jelennek meg, ha a pszicholingvisztika oldaláról közelítjük. A dolgozat kérdésfeltevésében megjelenik a nyelvtechnológia eljárások, eszközök alkalmazhatóságának ilyen irányú, szűkebben pedig pszichodiagnosztikai motiváltságú felhasználási lehetősége. A pszicholingvisztikai szempontok mentén megfogalmazott természetesnyelv-feldolgozási részfeladatok (Babarczy és mtsai., 2014): a beszédfelismerés, a beszéd-szintézis, az elemzés vagy parsin, a szemantikai elemzés, a (nyelv)generálás és az inferencia.

A nyelvtechnológiai feladatok közül a dolgozathoz tematikájában szorosabban illeszkedik a szentimentelemzés. Ez a különböző megnyilatkozások affektív tartalmának, az azokban megbúvó érzelmi viszonyulásnak a feltárására irányul. A szentimentelemzés során a megnyilatkozást létrehozó személy véleményét, attitűdjét, érzelmeit, hangulatát, értékítéletét tükröző nyelvi elemek, azaz a szentimentek detektálása, kivonatolása és feldolgozása történik jellemzően két, illetve három polaritás mentén: pozitív, negatív, semleges (Liu, 2020; Szabó, 2018). Módszertanát tekintve ez alapvetően kétféle, szótáralapú és gépi tanulós megoldással történhet (Liu, 2020; Szabó, 2018). Szabó (2018) dolgozatában a szentimentértékek magyar nyelvű detektálásához ismertet egy hibridnek tekinthető megoldást annak érdekében, hogy az érték módosulás, így értékvesztés vagy értékváltás nyelvi jelenségét kezelni tudja nyelvtechnológiai eljárás segítségével, ezáltal növelve a szentimentérték meghatározásának pontosságát. A szentimentelemzés során az elemzendő egység nagysága, mennyisége alapján három feldolgozási szint határozható meg (Liu, 2020; Szabó, 2018): a dokumentum, a mondat és az aspektus. A magyar nyelvű vizsgálatokban általában a szentimentelemzés alá sorolják az érzelemelemzést is, amelynek során az előbbihez képest egy sokkal árnyaltabb kategóriarendszer mentén (általában 6-8 érzelmi kategória) történő klasszifikáció valósul meg annak érdekében, hogy a szövegben rejlő emotív tartalmakat kinyerjék és a hordozott érzelmi tartalom alapján kategorizálják (Drávucz és mtsai, 2017). A kezdeti időszakban a szentimentelemzés főként ipari, gazdasági, marketing jellegű alkalmazásokban jelent meg, ma már sokkal szélesebb körben alkalmazzák. Kutatási szempontból ígéretes perspektívát kínál a gyermeknyelvi, illetve speciális populációk vizsgálatában, habár ilyen irányú vizsgálatok meg viszonylag kis számban találhatók. Gyermeknyelvi kutatásokban a relevanciáját és aktualitását többek között az adja, hogy már a korai időszakról fennáll az érzelmi és nyelvi kompetencia, a szociokommunikációs készségek és képességek fejlődésének kapcsolata, ami aztán a teljes életúton átívelve a társas interakciók, a társadalmi beilleszkedés sikeressége és a mentális

egészség szempontjából egyaránt meghatározó szereppel bír (Conte és mtsai, 2019; Ramirez és Ruetti, 2021): a szentiment- és érzelemelemzés ennek megismeréséhez, nyomonkövetéséhez nyújthat automatizált és kvantifikálható megközelítést. Kifejezetten igaz ez bizonyos neurodiverz fejlődési állapotok, így az autizmus spektrum esetén: az előző fejezetben részletesen ismerttettem az ehhez szorosan kapcsolódó sajátosságokat és a lehetséges magyarázatokat (például a mentalizációs nehézségeket és annak nyelvi vetületei), ezáltal is megalapozva a dolgozat kérdésfeltevését, valamint a távlati célt. Az affektív és evaluatív nyelvi jelenségek vizsgálata nemcsak elméleti, hanem klinikai alkalmazhatósággal összefüggő relevanciával bírhat, ami a diagnosztikai eszközök és a beavatkozás területén egyaránt értelmezhető.

3.2 A nyelvtechnológia módszerei és a nyelvi modellek

A nyelvtechnológiai megoldások megvalósításakor klasszikusnak tekinthető metodológia a szabályalapú megközelítés, ami a nyelvi képességgel kapcsolatban már rendelkezésre álló tudás alapján szabályok alkotásával és alkalmazásával építi fel az elvégzendő feladatokat kielégítő rendszereket. Előnyeként fogalmazható meg, hogy a fejlesztő sokkal nagyobb kontrollal rendelkezik a rendszer felett, ami egyúttal hátrány is abban a tekintetben, hogy jelentős szakértelmet kíván.

Az 1990-es években mind inkább előtérbe kerülő statisztikai alapú módszertan kiindulópontja a nagy mennyiségű természetes nyelvi input, azaz korpusz. A korpuszokból automatikus módon mintákat nyer ki a rendszer, amikhez aztán különböző statisztikai eljárások, valószínűségek mentén különböző értékeket rendel. Ezeknek az értékeknek a segítségével újabb szövegek feldolgozásakor vagy akár szövegek létrehozásakor a valószínűségek maximalizálására törekedve mintaillesztést végez, ezáltal az új adatokra vonatkozó predikciókat alkotva. Előnye lehet a flexibilitás, a sorrendezhetőség vagy rangsorolhatóság az elemzésekhez rendelt valószínűségek mentén, illetve a rejtett, új információk kinyerése. A rendszer megfelelő működéséhez ugyanakkor jelentős mennyiségű annotált adatra van szükség: ez szakértelem, illetve idő- és erőforrásigényessé teszi.

A statisztikai metodológiához szorosan kapcsolódva, szintén az 1990-es évek vége felé egyre élénkebb érdeklődés mutatkozott a gépi tanulással történő megoldására. A gépi tanulás, a neurális háló és a mélytanulás egy folyamatosan egyre jobban szűkülő módszertani megoldást jelent. Az alapvető cél a minta-, illetve szabály(osság)felismerés, megfelelő prediktív teljesítmény elérése, ami megvalósulhat *felügyelt/ellenőrzött (supervised learning)*, *felügyelet/ellenőrzés nélküli (unsupervised learning)* és *félleg felügyelt (semi-supervised learning)* módon egyaránt, valamint *megerősítéses tanulás (reinforcement learning)* révén is

(Jurafsky & Martin, é.n.). A neurális hálózatok valójában olyan gépi tanulási algoritmusok, amelyeket az emberi idegrendszer funkcionális és strukturális felépítése ihletett. A kezdeti architektúráknál ezzel összhangban – illeszkedve a konnekcionista megközelítésekhez – a hálózatok alapegységeként a mesterséges neuronokat képelték el, amik között különböző súlyú kapcsolatok jöttek létre: ezek a súlyozott kapcsolatok jelentették a hálózat tudását. Ma már sokkal inkább olyan hálózatként határozzák meg a neurális hálózatot, ami kisebb számítási egységek hálózatából áll (Jurafsky & Martin, é.n.; Nemeskey, 2020) A hálózat egységei, neuronjai között a súlyok beállítása az előbbieken ismertetett módon gépi tanulási algoritmusok segítségével történik. Ma már sokkal komplexebb felépítésű neurális hálókat alkalmaznak, amelyeknél a bemeneti és a kimeneti réteg között számos további *rejtett réteg* (*hidden layer*) található rejtett csomópontokkal. Működés módjukat tekintve ezek lehetnek *előrecsatolt* (*feedforward*), *rekurrens* vagy *visszacsatolt* (*recurrent*), illetve különböző *transzformer alapú* (*transformer based*) architektúrák.

Az alkalmazások háttérében álló nyelvmodellek az előbb említett módszertani megközelítések alapján valósíthatók meg. A nyelvmodellek kapcsán általában elmondható, hogy a szavak reprezentálása két különböző módszerrel valósulhat meg (eltérő matematikai számítások mentén): az egyik a folytonos tér, a másik a diszkrét technika. A modellek egyik klasszikus csoportját a statisztikai modellek képezik, amik valószínűségi alapú modellekkel dolgozva a természetes nyelvben fellelhető szabályosságok feltárását kísérik meg (Nemeskey, 2020). A valószínűségi nyelvmodellek széles körben elterjedt formái az n-gramm modellek (például uni-, bi- vagy trigramm) (Jurafsky & Martin, é.n.; Nemeskey, 2020). A szavak reprezentálásában használt diszkrét nyelvmodellezési technika révén alkalmasak a szavak közötti hasonlóságok felhasználására a predikciók létrehozásakor.

A diszkrét szóreprezentálással összefüggő komoly korlátok a neurális hálós nyelvmodellek létrejöttében fontos szereppel bírtak. Már a korai neurális hálós modelleknél folytonos módon vektorokban reprezentálódnak a szavak. A hasonló kontextusban előforduló szavakkal kapcsolatos jobb általánosítási képesség, valamint a nagyobb mennyiségű előzmény kezelése és a pontosabb szópredikció a neurális modellek három fő olyan jellemzőjeként említhető, amelyek egyértelmű előnyt jelentenek a korábban bemutatott n-gramm modellekkel szemben. Hátrányuk ugyanakkor még ma is, hogy lassabb a tanítási folyamat és komplex, kevésbé interpretálható rendszerek. A neurális hálós megoldások téren áttörésnek tekinthető a word2vec (Mikolov és mtsai, 2013). A széleskörű alkalmazást a rekurrens neurális hálók alapuló nyelvmodellezés megjelenése és elterjedés tette lehetővé, ugyanakkor napjainkban a különböző transzformer alapú megoldásokkal létrehozott nyelvmodellek fejlesztése és

alkalmazása kerül előtérbe. Előnyük, hogy kontextualizált előrejelzést adnak, azaz a szavak közötti kontextális kapcsolatokat használják fel predikciók létrehozásához. Két fő típusa különböztethető meg a transzformer alapú nyelvmodelleknek az alkalmazott rétegek típusa alapján: a balról jobbra és a kétirányú model. A ma széles körben elterjedt neurális hálós előtanított modellek, így például a GPT-3 (Floridi & Chiriatti, 2020) vagy a BERT (Devlin és mtsai, 2019) a különböző nyelvtechnológiai és nyelvvel kapcsolatos feladatok széles körében nyújtanak kiemelkedően jó teljesítményt, ahogy például neurális hálózatokra támaszkodó spaCy elemzőlánc (Honnibal, 2015) is. Mindegyik esetében megtalálhatók a magyar nyelvi modellek is, így például a huBERT (Nemeskey, 2021), HILBERT (Feldmann és mtsai, 2021), huSpaCy (Orosz és mtsai, 2022), de fejlesztés alatt áll a PULI, is ami alapvetően egy GPT alapú magyar modell.

3.3 Alkalmazási területek

A dolgozat 3. fejezetének zárásaként néhány, a nyelvtechnológiai eljárásokkal összefüggő alkalmazási területet érintettem röviden, kiemelve egy-egy olyan példát vagy rendszert, ami az alkalmazhatóságot szemlélteti. A példák megválasztásánál törekedtem arra, hogy kapcsolódjanak a dolgozat témájához. Így például a klinikai alkalmazás esetében Linguistic Inquiry and Word Count (továbbiakban LIWC), valamint NarrCat (<https://webnarrcat.ttk.hu>), azaz narratív kategóriális tartalomelemző rendszer. A NarrCat rendszert alkalmaztam a disszertáció 5. fejezetében bemutatott vizsgálat gépi elemzéséhez, mert lehetővé teszi a nyelvhasználat lexikai sajátosságainak kvantitatív elemzését.

A nyelvtechnológiai eljárásoknak az egyik klasszikus és talán legáltalánosabb felhasználása, a hatalmas adattömegeket feldolgozó információkinyerés széles körben jelen van: az ipar, a gazdaság, a politika és a szolgáltatási területén is. Számos más eljárással, mint például a szentimentelemzéssel kombinálható, ezáltal komplexebb, árnyaltabb információk kinyerését teszi lehetővé. Jordan és munkatársai (2019) például a 2016-os amerikai elnökválasztás meglepő eredményének megértéséhez, a politikai vezetőkre jellemző mögöttes pszichés sajátosságok nyelvhasználatban megragadható jellemzőinek feltárásában alkalmazták, felhasználva a LIWC rendszert. Egyre gyarapodnak a hazai kutatások és kutatási eredmények is, így például a társadalomtudomány, politika és ösztársadalmi jelenségek vizsgálatában (bővebben lásd Sebők, 2016; <https://poltextlab.tk.hu/en>; <https://rc2s2.elte.hu>).

Az információkinyerés klinikai területen is jelentős szerepű: megjelenik például a „társalgó rendszerek”, a chatbotok klinikai alkalmazásában. De a világ népességének időskorú tagjait érintő jelentős probléma, a kognitív képességek hanyatlásával (is) járó különféle demenciáknak a korai felismerése, valamint a mielőbbi beavatkozás miatt is kiemelt szereppel

bír. A számítógépes nyelvészet és nyelvtechnológia jelenléte ezen a területen is egyre hangsúlyosabbá válik (lásd például Hoffmann és mtsai, 2017). Számos más klinikai kórkép, mentális zavar, valamint a mentális egészség kapcsán jelennek meg hazai és nemzetközi beszámolók, tanulmányok (lásd például Bagi és mtsai, 2019; Calvo és mtsai, 2017; Gosztolya és mtsai, 2022; Liu és mtsai, 2021; Vetráb és mtsai, 2022;). A klinikai alkalmazás vagy alkalmazhatóság egy sajátos megvalósulási formája a pszichológiai jelenségek nyelvi formákkal való összefüggéseinek feltárása, ami a szűrés, a diagnosztikai folyamat és a beavatkozások tervezése és monitorozása szempontjából is hasznos információkat nyújthat. Gyakorlati megvalósulása a nyelvi formák és pszichológiai összefüggéseik lehorgonyzását szószámláláson alapuló szövegelemzéssel megvalósító LIWC program (Tausczik & Pennebaker, 2010). Az angol nyelvű szövegek elemzésében a pszichológiai alapú természetes nyelvfeldolgozás „gold standardjeként” is definiált alkalmazásnak 2022-ben adták ki ötödik verzióját (LIWC-22) (részletes leírásért lásd például Boyd és mtsai, 2022). A pilot-vizsgálatban alkalmazott NarrCat, azaz narratív kategoriális tartalomelemző rendszer fejlesztését a LIWC ihlette. A NarrCat (Ehmann és mtsai, 2014) rendszerét jelen vizsgálatba annak érdekében vontam be, hogy vizsgálatom módszertanában konzisztens legyen a szisztematikusan áttekintésben (a dolgozat 4. fejezete) szereplő több tanulmánnyal, amelyek a nyelvhasználat lexikai sajátosságait nézték. A rendszert azonban ennél jóval komplexebb kérdések megválaszolására és jelenségek feltárására alkalmazzák a tudományos narratív pszichológiai kutatásokban.

A gyermeknyelvi, illetve speciális populációktól származó nyelvi mintákból felépített korpuszok vagy adatbázisok, az azokon végzett gépi és kézi elemzések, nyelvészeti kutatások a számítógépes nyelvészet és a nyelvtechnológia kínálta lehetőségek alkalmazásának sajátos megvalósulását képezhetik, amik részben klinikai, részben kutatási motivációk és célok mentén szerveződhetnek. Ma már nagy számban állnak rendelkezésre különféle általános és specifikus korpuszok, így például a CHILDES, azaz Child Language Data Exchange System (MacWhinney, 2000), Gyermeknyelvi beszédAdatBázis és Információtár, azaz a GABI (Bóna, 2017), Magyar Óvodai Nyelvi Korpusz, azaz MONYEK (Mátyus & Orosz, 2014) vagy a Beszélt nyelvi Adatbázis (BEA, Gósy és mtsai, 2012; Grácsi és mtsai, 2020), a Tini BEA (Gyarmathy & Neuberger, 2015) és a KorSzak (al)korpuszai (Baumann és mtsai, 2020). Jó kiindulópontként szolgálhatnak különböző gépi elemzésekhez, segíthetnek abban is, hogy koncepcionális és módszertani megfontolásokat nyújtsanak speciális, atipikus fejlődésmentű gyermekektől származó nyelvi minta gyűjtéséhez korpusz vagy adatbázisépítési céllal.

Klinikai, pedagógia és társadalmi szempontból egyaránt kiemelt jelentősége van az akadálymentesítésének, amelynek egyik lehetséges formája az asszisztív technológiák alkalmazása. Az asszisztív technológiák egy speciális csoportját a különböző augmentatív és alternatív kommunikációs (AAK) eszközök képezik, amelyek kiemelt jelentőséggel bírnak a komplex kommunikációs igényű személyek kommunikációjának vagy akár nyelvvelsajátításának támogatásában. A nem vagy alig beszélő autizmussal élő személyek az AAK-t használók egy speciális csoportját képezik. Számukra is érdekes lehet a Pereira és munkatársai (2022) által létrehozott transzformer alapú neurális hálós architektúrát alkalmazó BERT nyelvi modellen alapuló rendszer (PictoBERT), ami a felhasználó számára releváns piktogramok felkínálásával a gyors, hatékony kommunikációt támogatja. Számos további, a hétköznapi életben már alapvetőnek tekintett alkalmazási területe és módja van a nyelvtechnológiai eszközöknek és eljárásoknak. Ilyen például a helyesírásellenőrzés, a gépi fordítás, az automatikus feliratozás, az információkeresés, amelyek evidens módon hordozzák magukban az edukációs jellegű felhasználást, de bizonyos esetekben az akadálymentesítés lehetőségét is.

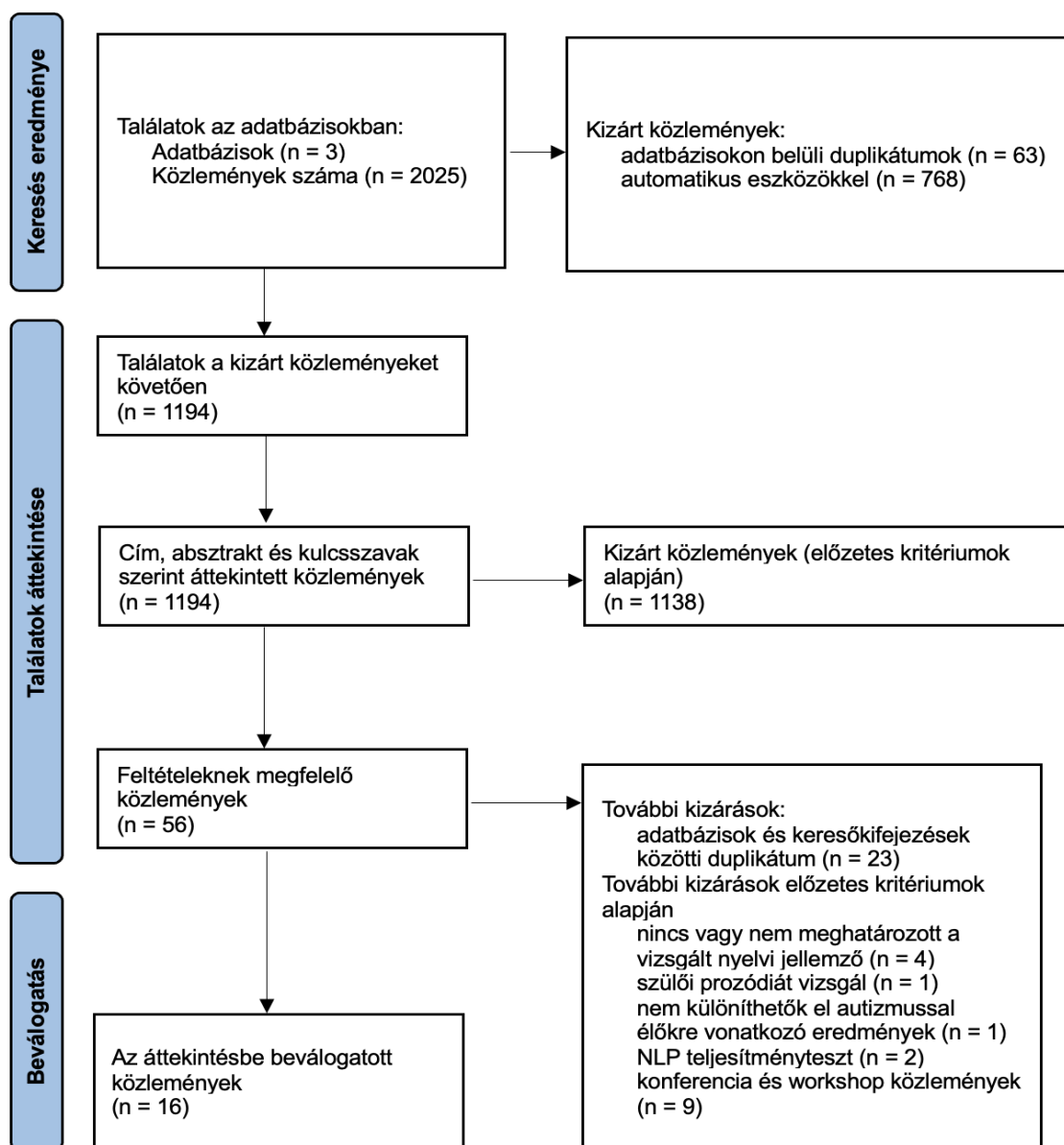
4. Autizmus és természetesnyelv-feldolgozás: szisztematikus áttekintés

A dolgozat 4. fejezete egy szisztematikus áttekintés folyamatát és eredményeit mutatja be: a dolgozat kérdésfeltevéséhez illeszkedve azokat az információkat, amelyek a nyelvtechnológiai eszközök és eljárások alkalmazásával és az autizmus spektrum állapot nyelvi, nyelvhasználati sajátosságaival kapcsolatosak. A kutatómunka és az áttekintés során megismert tanulmányok egyúttal kiindulópontként szolgáltak a disszertáció 5. fejezetében ismertetett pilot-vizsgálat hipotéziseinek megfogalmazásához, valamint az alkalmazott gépi elemző megválasztásához.

A szisztematikus kutatómunka során áttekintettem azokat a szakirodalmakat, amelyekben az autizmussal élők nyelvi, nyelvhasználati jellemzőinek megismerését nyelvtechnológiai eszközökkel és eljárásokkal próbálják megvalósítani. Az áttekintés egyrészt hiánypótlónak tekinthető, hiszen jelenleg magyar nyelven nincs olyan tanulmány, ami ezt a problémakört ismertetné. A szisztematikus áttekintés készítésekor ilyen jellegű angol nyelvű tanulmányról sem volt tudomásom. Másrészt szeretnék rávilágítani arra, hogy bár jelenleg kevésbé kutatott területről van még szó, a technológia rohamos fejlődésének, a mind szélesebb körű és könnyebb hozzáférhetőségének köszönhetően az ilyen jellegű kutatások gyakorlati, empirikus megvalósíthatóságának jövőbeli kedvező irányú változása feltételezhető. Áttekintésem szűkebben leginkább a „scoping review” típusába sorolható, hiszen az alapvető

célom a kutatások mennyiségének és irányainak feltárása azzal a céllal, hogy egy, a dolgozat keretein túlmutató, jövőbeli hazai kutatás kereteit és főbb irányvonalait felvázoljam.

Ennek érdekében három nagy online adatbázisban (Scopus, Web of Science, EBSCO) azonosított 2025 közleményből (a duplikátumok kiszűrését követően 1962), meghatározott keresőkifejezések mentén végrehajtott, előre definiált beválasztási és kizárási kritériumok alapján kiszűrt 16 tanulmányt tekintettem át (lásd 1. ábra). A beválogatási feltételek között szerepelt, hogy 2000. január és 2022. október között megjelent, angol és magyar nyelvű, teljes szöveggel elérhető tanulmányok és könyvek, könyvfejezetek legyenek. Fontos szempont volt, hogy ezek a munkák lektorált közlemények legyenek, vagyis ún. peer reviewed folyóiratokban jelenjenek meg.



1. ábra PRISMA folyamatábra a szakirodalmi keresés, áttekintés és kiválasztás folyamatáról

A tanulmányok feldolgozása, áttekintése során több általános módszertani szempontból releváns aspektust is megvizsgáltam, így például az autizmussal élő személyek azonosításának módját, a kontrollcsoport alkalmazását, a csoportok illesztési szempontjait, a minta elemszámát. A szisztematikus áttekintésbe beválogatott munkákkal kapcsolatban általánosságban megfogalmazható, hogy alapvetően empirikus jellegűek és a minta viszonylag kis elemszámú, ami az eredmények és a következtetések kapcsán egyaránt limitációként fogalmazható meg. Az áttekintett tanulmányok alapján jól látszik, hogy az elemzések alapjaként szolgáló minták döntő többségében egyértelműen komplexebb nyelvi megnyilatkozások, narratívák és/vagy társalgási helyzetek. Számos esetben standard mérőeszköz alkalmazásából származó szövegekről van szó. A vizsgált nyelvi szint és jelenségek nagyrészt a szemantika, lexiko-szemantika, esetleg a pragmatikai területéhez kapcsolódnak, illeszkedve a korábban már leírt nyelvi és kognitív sajátosságokhoz. A nyelvtechnológiai eljárások és eszközök kapcsán megfogalmazható, hogy módszertanilag széles kört ölelnek fel, ugyanakkor döntő többségében az előbb említett nyelvi jellemzőket megragadni képes vektortér-reprezentációs, valamint szótár alapú megoldások dominálnak. Újdonságértékük nem annyira a nyelvi vonatkozások, mint inkább a nyelvtechnológiai metodikában, illetve annak alkalmazási vetületeiben ragadható meg.

A tanulmányokkal kapcsolatban összességében elmondható, hogy a nyelvi sajátosságok elemzéséhez nyelvtechnológiai eszközöket alkalmazó vizsgálatok motivációja nagyrészt klasszifikációs és detekciós. Ezt jelentős mértékben a jövőbeli klinikai alkalmazhatóság, a hasznosulás indokolja, ami egyúttal a jelenlegi idő- és erőforrások szempontjából költséges prevenció és diagnosztikus folyamatok költséghatékonyabbá tételét is jelentheti. A direkt módon a nyelvi markerekre fókuszáló vizsgálatok is, mint például a női fenotípus nyelvi sajátosságainak feltárását célzó, részben klinikai, diagnosztikus céllal keresik a megfogalmazott kérdésekre a választ. Általában megfogalmazható, hogy mindegyik tanulmány esetében direkt vagy közvetett módon jelen van az objektifikált, automatizált diagnosztikai relevancia igénye: akár mint korai szűrés, akár mint kiegészítő eljárás. Az áttekintett munkák nagy részében elsősorban szóbeli produkciók elemzéséről van szó, amelynek ingeranyaga sok esetben az ADOS (Autism Diagnostic Observation Schema) (Lord és mtsai, 2012) életkor és nyelvi szint alapján meghatározott moduljának aktivitásaihoz kapcsolódik. A különböző online közösségek, szociális média felületek szintén értékes információkkal szolgálhatnak a nyelvhasználati és azon túlmutató jelenségekről, ugyanakkor a tényleges diagnózissal kapcsolatos kétségek az ilyen jellegű munkák esetében óvatosságra intenek.

5. A kutatás

A dolgozat 5. fejezetben egy kismintás, esettanulmány jellegű saját kutatást ismertetek, amelyben a 2. és 4. fejezet szakirodalmi alapján megfogalmazott hipotézisek teszteléséhez gépi elemzést (NarrCat) alkalmazok, problémafeltáró jelleggel kiegészítve kézi elemzéssel. A pilot vizsgálatomban a disszertáció fő célkitűzéséhez kapcsolódóan azt mutatom meg, hogy a nyelvi gyakorisági adatokon alapuló megközelítéssel mennyire tárható fel, hogy az autizmussal élő személyek mutatnak-e eltérést az értékelést, affektivitást és kognitív jelentést inherensen, lexikai szinten tartalmazó szavak használati gyakoriságában a nemben és életkorban illesztett neurotipikus személyekhez képest. A dolgozatom fontos újdonsága, hogy mindezt magyar nyelvű adatközlőkkel vizsgálom. A vizsgálat idejében hasonló jellegű, az autizmussal élő személyek körében végzett NLP-eszközt alkalmazó, publikált korábbi munkáról nem áll rendelkezésemre információ, így abból a feltételezésből indultam ki, hogy ilyen jellegű magyar nyelvű adatok még nincsenek.

A hosszabb távú kutatásom első lépéseként a vizsgálatot írott nyelvű szövegeken végeztem. Összességében megállapítható, hogy nagyon alacsony számban vannak az írott nyelvre vonatkozó eredmények, ráadásul, ha figyelembe vesszük a szisztematikus áttekintésből kizárt tanulmány(ok) következtetéseit (például Newton és mtsai, 2009) is, akkor az írott nyelvre vonatkozó eredmények ellentmondásosak. Ugyanazon tesztekkel elvégzett írásbeli és szóbeli vizsgálatok fényt deríthetnek ezekre az ellentmondásokra: a pilot vizsgálatomban ennek a tervezett munkának az első részét, az írott nyelven való tesztelést végeztem el.

5.1 A kutatás célja, hipotézisei, módszere

A vizsgálatom fókuszában tehát az autizmussal élő személyek írásbeli nyelvhasználatában megfigyelhető lehetséges megkülönböztető szerepű, interindividuais eltéréseinek és csoportsjátosságainak az azonosítása áll egy specifikus nyelvtechnológiai eszköz, a NarrCat alkalmazásával. A disszertációm fő kérdésfeltevése ugyanakkor ennél sokkal tágabb (exploratív jellegű): arra keresem a választ, hogy a rendelkezésre álló különféle nyelvtechnológiai eszközök és eljárások alkalmazhatóak-e az autizmus spektrum vizsgálatában, és ha igen, milyen alkalmazási eljárások tekinthetők hiteles eredményt nyújtó vizsgálatnak, és milyen információkkal szolgálhatnak ezek a kutatások az autizmussal élő személyek nyelvi, nyelvhasználati sajátosságairól.

A megfogalmazott hipotéziseim a szisztematikus áttekintés és további szakirodalmi kutatómunka során azonosított tanulmányok nemzetközi (döntően angol nyelvű) adatain,

valamint az autizmussal kapcsolatos, a 2. fejezetben megfogalmazott ismereteken alapulnak (Boorsie és mtsai, 2019; Chojnicka, & Wawer, 2020; De Bruyne és mtsai, 2018; Gaigg és mtsai, 2018; Gormley és mtsai, 2021; Happé, 1993, 1994a, 1994b; Kumar és mtsai, 2016; Lee és mtsai, 2018; Milosavljevic és mtsai, 2016; Nguyen és mtsai, 2013, 2014, 2015; Quill & Stansberry Brusnahan, 2017; Renteria-Vazquez és mtsai, 2021; Tager-Flusberg, 1992, 1999; White és mtsai, 2009). Ezek alapján három fő hipotézist fogalmaztam meg:

H1: Az autizmussal élő személyek nyelvi megnyilatkozásaiban a szentimentek, az értékeléssel kapcsolatos kifejezések kisebb arányban jelennek meg.

H2: Az autizmussal élő személyek nyelvi megnyilatkozásaiban kisebb arányban jelennek meg az érzelmekkel kapcsolatos kifejezések.

H2.1.: Az autizmussal élő személyek esetén a negatív affektivitást, jelentéstartalmat hordozó nyelvi megnyilatkozások nagyobb arányban vannak jelen

H2.2: A neurotipikus kontroll személyek esetén a pozitív érzelmi töltetű nyelvi elemek nagyobb arányban jelennek meg a megnyilatkozásokban.

H3: Az autizmussal élő személyek nyelvi megnyilatkozásaiban kisebb arányban jelennek meg a kognícióval kapcsolatos kifejezések.

A vizsgálatban résztvevő személyekkel kapcsolatban közös elvárás volt, hogy 16 éves életkorukat betöltött, magyar anyanyelvű fiatalok és felnőttek legyenek. Ezt egyrészt az indokolta, hogy a történetalkotási képességek a szakirodalom szerint ebben az életkorban eléri a felnőttkori szintet, másrészt a késői diagnózis (főként a lányok, nők) szempontjából is érdekes az idősebb kor. A vizsgálat fókuszában – a DSM-5 és a BNO-11 által használt terminológiával – az autizmus spektrum zavar diagnózissal rendelkező személyek álltak. A vizsgálatban végül összesen 45 fő (N) vett részt: 23 fő autizmus spektrum zavar diagnózissal rendelkező személy és 22 fő neurotipikus fiatal és felnőtt. A két csoport illesztésénél a személyek nemet és életkorát vettem figyelembe. Ugyan a két csoport adatai nem teljesen kiegyensúlyozottak, az elvégzett statisztikai próbák eredményeit tekintve sem a nem ($\chi^2(1) = 0.237$; $p = 0.763$), sem pedig az életkor ($t(43) = -1.47$; $p = 0.148$) esetén nem mutatkozott szignifikáns különbség.

Az adatgyűjtéshez online kérdőív alkalmazását választottam annak érdekében, hogy minél több résztvevőt elérjek. Az eljárás kiválasztásakor fontos szempont volt az, hogy az autizmussal élő csoport tagjai esetében a személyes részvételből fakadó, a társas és ismeretlen helyzettel összefüggő szorongást lehetőleg ki tudjam zárni vagy minimalizáljam, ezáltal biztosítva az érzelmileg biztonságos környezetet, ezzel is facilitálva a jobb nyelvi, narratív teljesítményt. A kérdőív létrehozása és az adatok rögzítése az ingyenesen elérhető és kutatási

célokra is gyakran alkalmazott Google Forms űrlap segítségével történt. Az elemzések alapjául szolgáló kérdéseket, feladatokat tartalmazó blokkban rövid véleményt, illetve történetet kellett alkotnia a személyeknek egy hír, egy videó, egy kép, egy élettörténeti momentum és két étkezési szituáció kapcsán. Az űrlap segítségével gyűjtött szövegek tartalmi és nyelvi helyességét manuálisan ellenőriztem: ahol szükséges volt, a hibákat javítottam (például *álmosik-álmodik*, *gyönyörű-gyönyörű* vagy *buszkek-büszkék* helyett) annak érdekében, hogy az elemzőrendszer megfelelően tudja azonosítani a különböző modulokhoz, kategóriákhoz tartozó elemeket. A manuális ellenőrzést követően lefuttattam az elemzést a NarrCat rendszerében, és az így kapott adatokat használtam fel a statisztikai elemzésekhez. A statisztikai elemzés során a kérdőív kérdéseire adott válaszok eltérő száma és hossza miatt a NarrCat elemzőrendszer által kiadott elsődleges abszolút gyakorisági adatokból (darabszám) képzett relatív gyakorisági mutatókat vettem alapul az elemzésekhez. Ez utóbbi az abszolút gyakorisági adatok és a válaszok szószámban mért hosszának hányadosából képzett számokat jelenti.

5.2 Eredmények

Az első hipotézis vizsgálatához, amellyel azt vizsgáltam, hogy a szentimentek (a NarrCat kategóriarendszerében ehhez leginkább illeszkedő Értékelés modul által nyert adatok) mentén az értékeléssel összefüggő kifejezések megjelenési arányában mutatkozik-e különbség a két csoport között. Specifikusabban azt feltételeztem, hogy az autizmussal élő személyek esetén arányaiban kevesebb ilyen értékelő nyelvi eszköz jelenik meg. A statisztikai próba eredménye alapján ($U = 249$; $p = 0.928$) a két csoport nem mutatott szignifikáns különbséget. Külön-külön vizsgálva a negatív és pozitív értékeléssel összefüggő nyelvileg expliciten kifejeződő tartalmakat, szintén nem mutatkozott szignifikáns különbség a két csoport esetében (negatív $U = 249$; $p = 0.937$, pozitív $U = 245$; $p = 0.865$). Azaz az értékeléssel összefüggő kifejezések megjelenési arányában (szemben az első hipotézissel) nem mutatkozik különbség a vizsgált és a kontrollcsoport között.

A második hipotézis vizsgálatában az affektivitással, érzelmekkel kapcsolatos nyelvi elemek megjelenésében mutatkozó különbségeket vizsgáltam (a NarrCat Érzelem moduljának eredményei mentén). Azt feltételeztem, hogy az autizmussal élő személyek csoportjában kisebb számban jelenik meg az affektív tartalmak kifejezése, valamint az arányok tekintetében a negatív hangulatú, jelentéstartalmú kifejezések jelentősen magasabb számban lesznek jelen a kontrollcsoporthoz viszonyítva, míg utóbbi csoportban a pozitívak dominálnak. Az eredmények alapján az érzelmek összesített kifejeződésére vonatkozó adatok alapján nem mutatkozott szignifikáns különbség a két csoport esetében ($t(43) = -1.65$; $p = 0.106$), azaz az

érzelmeikkel összefüggő kifejezések megjelenési arányában nem mutatkozik különbség az autizmussal élő és a kontrollcsoport között. Sem az autizmussal élő csoport esetében a negatív érzelmeire ($t(43) = -1.665$; $p = 0.103$), sem a kontrollcsoport esetében a pozitívakra ($t(43) = -1.15$; $p = 0.258$) vonatkozó eredményekben nem mutatkozott szignifikáns különbség a két csoport között. Az adatok statisztikai elemzése a második hipotézis állításait nem igazolta.

A harmadik hipotézis a kognícióval, a kognitív folyamatokkal, funkciókkal és állapotokkal kapcsolatos kifejezések megjelenési arányának csoportközi vizsgálatára irányult (NarrCat Kogníció moduljának segítségével): azt feltételeztem, hogy az autizmussal élő személyek csoportjában alacsonyabb számban jelennek meg ezek a kifejezések a megnyilatkozásokban. A lefuttatott Student-féle t-próba eredménye alapján a két csoport között ebben az esetben sem mutatkozott szignifikáns különbség ($t(43) = -0.265$; $p = 0.792$).

5.3 Az eredmények megvitatása

A hipotézisekben megfogalmazottak szerint az előzetes feltételezésem az volt, hogy a vizsgálati csoportnál alacsonyabb számban, arányban lesznek jelen az értékeléssel, érzelmeikkel és kognícióval kapcsolatos nyelvi elemek, mint a kontroll csoportnál. Emellett úgy véltem, például Nguyen és munkatársai (2013, 2014), valamint Kumar és munkatársai (2016) korábbi eredményeit figyelembe véve, hogy az affektivitás területén a vizsgálati személyeknél a negatív jelentéstartalmak, míg a neurotipikus csoportban a pozitívak túlsúlya figyelhető meg. A rendelkezésre álló adatok statisztikai elemzése egyik hipotézisemet sem támasztotta alá, vagyis egyik vizsgált változó esetében sem tudtam igazolni szignifikáns csoportkülönbséget.

Azt, hogy a csoportok között nem mutatkozott szignifikáns különbség, számos tényező magyarázhatja. Az egyik tényező lehet, hogy az ismertetett és a hipotézisek megfogalmazásában felhasznált tanulmányok nagy része a szóbeli nyelvhasználattal összefüggő feladatokban, döntő többségében ADOS-aktivitásokban vizsgálta a szóhasználati sajátosságokat. Kivételt képez ez alól De Bruyne és munkatársai (2018) és Nguyen és munkatársai (2013, 2014, 2015) munkája, akik szintén írásbeli szövegalkotási helyzetekből és nem szóbeli megnyilatkozások átírataiból származó narratívák, társalgások LIWC-kategóriák mentén történő szóhasználati sajátosságainak vizsgálatával detektáltak eltérést egyes kategóriatípusok, így az affektivitás és kogníció kapcsán.

A tanulmányok nemcsak a vizsgált minták modalitását, de az életkori övezetet és az adatgyűjtéshez alkalmazott eljárást tekintve is heterogének és eltérnek a saját kutatásomtól. A szentiment, affektív és kognitív tartalmak szóhasználati sajátosságaiban tapasztalható eltéréseket az életkor és az azzal szervesen összefüggő tapasztalatok is befolyásolhatják (Conte

és mtsai, 2019; Ramírez és Ruetti, 2021), akárcsak a választott feladat típusa, ahogy az egyértelműen látszik például Chojnicka és Wawer (2020) kutatásából. Vizsgálatomban több különböző típusú, de minden esetben az érzelmekkel, értékelésekkel és kognícióval kapcsolatos nyelvi kifejezések előhívását és használatát erőteljesen facilitáló feladatot adtam a résztvevőknek, amik szintén magyarázhatják a szakirodalmi adatoktól való eltérő eredményeket. A szóhasználatban mutatkozó különbségek másik forrása lehet az életkor, azaz érdemes megvizsgálni és figyelembe venni a fejlődés sajátosságokat és vonatkozásokat. Annak, hogy az általam elemzett szövegekben nem mutatkozott különbség a két csoport között, egyik lehetséges magyarázata lehet, hogy döntő többségében felnőtt személyektől származó megnyilatkozásokkal dolgoztam: esetükben pedig a tapasztalatok, a fejlődés eltérő menetéből fakadó fiatalabb életkorban markánsabban megnyilvánuló eltérések kevésbé kifejezettek, főként olyan szituációkban, amik elegendő időt biztosítanak a nyugodt szövegalkotáshoz és mentesek a társas helyzeteket jellemző számos további szorongást okozó tényezőtől.

Az eredményeim leginkább Newton és munkatársai (2009) blogok szóhasználati sajátosságainak vizsgálata során nyert eredményeivel mutatnak egyezést. Az eredményeket azonban óvatosan kell kezelni, hiszen az online közeg egy minőségében teljesen más modalitását nyújtja a kommunikációnak. Ez a megnyilatkozási forma befolyásolhatja, elfedheti azokat a nehézségeket, amik a spontán kommunikációban, a társas interakciók során a szóbeli szövegalkotásakor könnyebben felszínre kerülnek. A saját eredményeim kapcsán a csoportok közötti teljesítménybeli kiegyenlítődés magyarázható akár ezekkel, az általuk disztálisnak nevezett helyzetből fakadó kedvező körülményekkel is, amik egyrészt csökkentik a szorongást és a stresszt a személyben, másrészt önszabályozott, hogy a személy mennyi időt és energiát fordít az egyes kérdések megválaszolására. Newton és munkatársai (2009) továbbá úgy vélik, hogy a proximális kontextusban (például laborban zajló vizsgálatok vagy maga a diagnosztikus folyamat) megfigyelhető kommunikációs nehézségek, és nyelvi, nyelvhasználati eltérések mint a vizsgálati helyzet melléktermékei értelmezhetők amiatt, hogy alapvetően társas interakciókban történő vizsgálatokról van szó. Ezt a megállapítását ugyanakkor érdemes némi fenntartással kezelni, hiszen a hétköznapi életben való részvétel, a társadalmi beilleszkedés mindig társas közegben, különböző társas interakciók során történik, így az ezekben a naturalisztikus helyzetekben nyújtott teljesítmény sokkal inkább mérhető az egyén és a társadalom, az életminőség, a pszichés jólét, a boldogulás és a jövőbeli perspektívák szempontjából. Emellett azt is fontos szem előtt tartani, hogy a jelenlegi viselkedés alapú diagnosztikus folyamat éppen ezekre a társas, szociokommunikációs nehézségekre fókuszál: a

standard diagnosztikus eszközök, eljárások alkalmazásával ezek mentén tudják a szakemberek igazolni vagy kizárni az autizmus meglétét.

5.4 Kiegészítő vizsgálatok, kiegészítő magyarázatok az eredményekhez

Egy jövőbeli kutatás módszertani megfontolásainak kezdő lépéseként, valamint a csoportkülönbségek hiánya mögötti lehetséges további magyarázatok feltárásának igénye által motiválva további, kiegészítő vizsgálatokat végeztem. Így például megvizsgáltam, hogy az autizmussal élő személyek csoportján belül a pozitív és negatív értékelő és érzelmi tartalmak megjelenésében tapasztalható-e szignifikáns különbség. Nguyen és munkatársai (2013, 2014, 2015) vagy Kumar és munkatársai (2016) munkájukban arról számolnak be, hogy a negatív hangulatú, negatív töltetű nyelvi elemek nagyobb arányban jelennek meg a vizsgálati csoport szövegeiben, mint a pozitívak. Az értékelő tartalmak esetében a csoporton belül nem mutatkozott szignifikáns különbség a pozitív és negatív polaritás tekintetében ($W = 151$; $p = 0.211$), míg az érzelmi töltet kapcsán szignifikáns különbség mutatkozott ($W = 235$; $p < 0.001$; Effect Size = 0.866). Az eredmények ugyanakkor a szakirodalmi adatoknak ellentmondóak, azaz a pozitív érzelmi töltet szignifikánsan nagyobb arányú megjelenését mutatja. Ez talán magyarázható a feladatok jellegével: sok esetben a fordulópontként elmesélt élettörténet valamilyen pozitív eseményhez, sokszor magához a diagnózishoz kötődik, vagy a videóról alkotott narratíváknál a vicces momentum, a kisfiú és Mr. Bean elégedettségének kifejeződése jelenik meg. Emellett a hivatkozott szakirodalmak esetében Nguyen és munkatársai (2013, 2014, 2015) vizsgálatai online közösségi bejegyzésekhez kapcsolódnak, amelyekben az autizmussal élő csoport esetében hangsúlyosan megjelennek a hétköznapi átható, az autizmussal összefüggő, abból fakadó társas nehézségek, szorongás és depresszió, így az alacsonyabb valenciájú elemek, a negatív jelentéstartalmak túlsúlya magyarázhatóvá válik ezekkel.

A statisztikai elemzés nem tárt fel szignifikáns különbséget a klinikai és a kontroll csoport megnyilatkozásainak szószáma tekintetében ($U=191$; $p=0.164$). Írásbeli szövegek elemzésekor De Bruyne és munkatársai (2018) is hasonló eredményekre jutottak, habár összességében a szakirodalomban diverz eredmények találhatók. Ezt azt is jól illusztrálja, hogy a kutatás kérdésfeltevése és az ahhoz illeszkedő módszertan szempontjából az adatgyűjtés módja kiemelt jelentőségű.

A gépi elemzés mellett manuális kódolást végeztem véletlenszerűen kiválasztott öt szövegen, amelyhez független annotálók segítségét kértem. Az annotáláshoz egy segédletet is összeállítottam, alapul véve Szabó (2018) munkáját, amelyben az annotálással kapcsolatos információkat, valamint az annotálás módját ismertettem példákon keresztül szemléltetve. Az

annotálóktól azt kértem, hogy jelöljék a pozitív és negatív szentimenteket, érzelmeket inherensen, lexikai szinten tartalmazó szavakat, kifejezéseket, valamint azokat az elemeket, amik hasonlóan inherensen hordozzák a kognitív jelentéstartalmat. Emellett azt is kértem, hogy a kontextusfüggő elemek is kerüljenek jelölésre, hiszen módszertanilag fontos kérdésre világítanak rá. A vizsgálat célja nem egy teljeskörű elemzés volt, sokkal inkább exploratív, problémafeltáró jellegű: egyrészt annak feltérképezése, hogy a NarrCat által megismert adatok és egy kézi elemzés során a gyakorisági adatokban tapasztalhatunk-e eltéréseket, és ha igen, vajon milyen tényezőkkel állhatnak kapcsolatban, másrészt rávilágítani módszertani kérdésekre, problémákra, amik például a szentiment- és érzelemelemzés kapcsán felmerülhetnek.

Az annotálók csak a nyers, előzetes adattisztításon, előfeldolgozáson átesett szövegeket kapták meg, a szövegekkel kapcsolatos további információkat, így például az életkort vagy a diagnózis meglétét nem ismerték, a szövegeket vakon annotálták. Mivel csak tendenciák feltérképezésére voltam kíváncsi, így az adatokat annyiban egyszerűsítettem, hogy a szentiment és affektív tartalmú kifejezéseket összesítettem egyrészt egy evaluatív kategóriába, másrészt ehhez kapcsolódóan a kettős polaritást megtartva negatív és pozitív kategóriákra. Összességében megállapítható, hogy a gépi elemzés során azonosított elemek száma jóval kisebb, mint a manuális annotálásakor jelölteké, azaz amennyiben nem ismerjük pontosan a nyelvtechnológiai rendszer háttérében működő eljárást, algoritmusokat és esetleg szótárakat, mindenképp célszerű lehet kiegészíteni kézi elemzéssel is a vizsgálatot, legalább véletlenszerűen néhány szöveg kapcsán.

A kézi és gépi azonosítás közötti jelentős eltérés számos kérdést és módszertani problémát, megfontolnivalót vet fel. Ilyen például, hogy a gépi rendszer mennyiben képes felismerni sajátos szavakat, neologizmusokat és idoszinkráziákat, amik például autizmussal élő személyek nyelvhasználatában megjelenhetnek, mint például a *butuli* kifejezés. A NarrCat szótáralapú és lokális nyelvtanokat működtető rendszere nem igazán alkalmas arra, hogy a mélyebb szöveggörnyezetet figyelembe vegye, így az azonosított elemek számában megfigyelhető különbség egyik magyarázata lehet a kontextusérzékenység is. Erre példa a sajátos beszélnyelvi fordulatként is értelmezhető az *egyszer, ha nézős* kifejezés, ami szintén inherensen hordozza a negatív értékítéletet, de abban az esetben, ha hiányzik a rendszer szótárából, nem kerül azonosításra. Ezek alapján megfogalmazható, hogy amennyiben egy szótáralapon működő rendszert alkalmazunk, kulcstényező a szótárban szereplő szavak és kifejezések jó megválasztása, a doménspecifitás kérdésköre (Liu, 2020; Szabó, 2018). Érdeemes az összeállított szólistákat kontextustól függetlenül még külön annotálókkal értékelteni, akik

valóban a szavak, kifejezések lexikailag hordozott értékminőségére vonatkozó ítéleteket fogalmazzuk meg. Aztán ehhez viszonyítva érdemes megfontolni a kontextusfüggő jelenségek kezelését, ami a területspezifitás kérdésével is szorosan összefügg. Fontos kitérni a nem szó szerinti jelentések, valamint a szentimentshifterek kezelésére is (Liu, 2020; Szabó, 2018). Ez manuálisan viszonylag könnyen kezelhető és felismerhető, azonban a gépi detektálásnak vannak nehézségei. A kognitív kifejezések felismerése és annotálása során sokkal kisebb eltérések láthatók mind a gépi elemzéssel, mind pedig az egyes annotálók jelölésével kapcsolatban. Megállapítható tehát, hogy a két kategória (evaluatív, kognitív) esetén nagyon eltérő tendenciák figyelhetők meg az elemzések alapján a gépi és kézi, valamint az annotálok esetében.

Az eredmények alapján megfogalmazható, hogy a jelentéssel összefüggő területeken a kontextusérzékenység a gépi elemzések során komoly problémát jelenthet, habár a mai, neurális alapú, mélytanulós, sokszor transzformer architektúrával megvalósított kontextuális és tartalmi beágyazásokkal élő nyelvmodellek ezeknek a problémáknak a kezelésében is új alternatívákat kínálnak, javítva a korábbi rendszerek és nyelvtechnológiai eszközök által nyújtott teljesítményt. A csoportok közötti különbségekre vonatkozóan az látszik ezekből az adatokból, hogy a kognitív kifejezések esetén a kézi elemzések eredményei alapján sem igazán mutatkozik releváns eltérés a két csoport között, ahogy az a gépi adatok statisztikai elemzésekor sem jelent meg. Ehhez képest, bár a NarrCat adatain végzett statisztikai próbák nem mutattak ki szignifikáns különbséget a szentiment és affektív jelentéstartalmú kifejezések gyakorisági adatain a vizsgált és a kontrollcsoport között, a manuális annotálás eredményei alapján nagyobb mértékű eltérések mutatkoznak, azaz relatív jelentősebb, komolyabb eltérést sejtetnek az adatok a két csoport között az evaluatív kategóriához tartozó nyelvhasználatában. Az eredmények eltéréseinek egyik magyarázata lehet, a korábban már említett kontextusfüggőség, ami tetten érhető az annotálók viszonyulásában is az adott témához, kifejezésekhez, ahogy azt a kézi elemzés tapasztalatai is alátámasztották. Másik magyarázatként megfogalmazható a különböző kategóriába sorolható kifejezések azonosítása során tapasztalható eltérő nehézségi szint, a könnyebb szótárazhatóság, azaz mintha a kognitív kifejezések nemcsak a kontextusfüggőség tekintetében, hanem mind a jelentésüket, mind a terjedelmüket (a többszavas kifejezések kevésbé jellemzőek) tekintve egyértelműbbek lennének.

A manuális elemzések eredményei, az annotálók között mutatkozó különbségek (például a kontextushatás érvényesülésének mértéke vagy a személyes tapasztalatokon alapuló érzelmi töltet mentén történő azonosítás) két, módszertani szempontból fontos kérdésre hívják

fel a figyelmet: egyrészt az annotálási segédlet összeállításának és az annotálók felkészítésének kiemelt szerepére a kézi elemzés folyamatában, másrészt szorosan ezzel összefüggésben az annotálók közötti egyetértés vizsgálatának kiemelt jelentőségére. A kutatás következő lépésként mindenképpen célszerű lenne az összes rendelkezésre álló szöveget kézi annotálással elemezni: egyrészt annak érdekében, hogy kiderüljön, az ezen a kis mintán az evaluatív elemek használatában mutatkozó csoportközi különbségek statisztikailag is igazolhatók-e, másrészt azért, hogy külön-külön is megvizsgáljam az értékeléssel és az affektivitással kapcsolatos kifejezések alakulását, az annotálásukban mutatkozó nehézségeket. A dolgozatom szempontjából fontos eredménynek tekintem azt a megállapítást, hogy a szótáralapú és a nem megfelelően megválasztott szövegeken, tanítóhalmazon betanított neurális modellek esetében is komoly általános felhasználási és alkalmazási korlátot jelent a doménspecifitás: különösen alapos megfontolást igényel mindezeknek a kérdéseknek a megfelelő kezelése a megbízhatóság miatt a klinikai gyakorlat, a klinikai alkalmazhatóság szempontjából.

A kutatás számos limitációval rendelkezik, így például a kis elemszám vagy az online kérdőíves módszer alkalmazása. Emellett a résztvevők közötti finomabb különbségek nem ragadhatók meg a nyelvileg expliciten kifejezett elemek gyakorisági alapú megközelítésével. A kutatás közvetlen folytatásaként mindezek alapján célszerű egyrészt a meglévő szövegek további kézi annotálását elvégezni, az így nyert adatokon is elvégezni a statisztikai elemzéseket, másrészt szóbeli megnyilatkozások átiratainak segítségével megvizsgálni a jelen kutatásban tesztelt hipotéziseket, valamint feltárni az írott nyelvi és a szóbeli adatok összehasonlító elemzése során mutatkozó esetleges különbségeket, ahogy arra Newton és munkatársai (2009) is felhívják a figyelmet. Harmadrészt mindenképp érdemes lehet megvizsgálni az adatok alakulását más, már rendelkezésre álló gépi nyelvfeldolgozó rendszert bevonva a vizsgálatba (például az emtsv-t), valamint feltérképezni a neurális háló alapú megoldásokban rejlő lehetőségeket egy jövőbeli gyakorlati alkalmazhatóság mentén. Ez természetesen csak kutatócsoportban és intézményi együttműködések keretén belül valósítható meg érdemben. Hosszabb távon az eredmények hasznosulása, a klinikai és pedagógiai relevanciája kiemelt szereppel bír. Különösen érdekes és értékes szegmensét képezi az alkalmazhatóságnak a szűrés és diagnosztika területe, hiszen napjainkban az autizmus diagnózisának megállapítása idő- és humán erőforrásigényes. Az ilyen eljárások kidolgozásához és alkalmazásához nem elegendők a kis mintás vizsgálati eredmények. Emellett, ahogy arra korábban is felhívtam a figyelmet, kiemelt jelentőséggel bír, hogy a vizsgált nyelvi megnyilatkozások milyen forrásból származnak és milyen témához kapcsolódnak.

6. Összegzés

Jelen dolgozat kérdésfeltevésében alapvetően elméleti indíttatású: annak feltérképezését kívánja megvalósítani, hogy a nemzetközi és hazai kutatásokban alkalmaznak-e, és ha igen, milyen nyelvtechnológiai eljárásokat az autizmussal élő személyek nyelvi, nyelvhasználati sajátosságainak mélyebb megismerése érdekében. A disszertáció exploratív jellegével szoros összhangban egy szisztematikus áttekintés és egy kis mintás esettanulmány jellegű vizsgálat eredményeit, valamint az eredményekhez vezető folyamatot és az így szerzett tapasztalatok mentén felmerülő kérdéseket, problémákat ismertettem. A dolgozat célja, mintegy nulladik lépésként, egy jövőbeli komplex kutatás irányvonalainak megalapozása. Ennek érdekében részletesebben áttekintettem az autizmussal kapcsolatos korszerű tudásunkat, hiszen csak e mentén fogalmazhatók meg releváns, a klinikai és a pedagógiai gyakorlat szempontjából is hasznosítható, hasznosuló kérdések. E kettős motiváció, a gyógypedagógusé és az alkalmazott nyelvészé képezi a dolgozat személyes indíttatását

Az autizmus megismerése és természetének megértése mellett fontos a számítógépes nyelvészet, a nyelvtechnológia alapfogalmainak, alapvető módszertani kérdéseinek és alkalmazhatóságának ismerete. Ez utóbbi különösen hangsúlyos, hiszen a fejlődési zavarokkal, így az autizmus spektrummal kapcsolatos kutatások eredményeinek hasznosulása, a klinikai és pedagógiai relevanciája kiemelt szereppel bír. Az autizmus jelenlegi, viselkedéses alapú diagnosztikáját jelentős mértékben befolyásolja a diagnosztikát végző személy szakértelme és korábbi tapasztalatai. Éppen ezért egy automatizált, objektív eszköz alkalmazásával a szubjektív tényezők szerepe csökkenthetővé válik, valamint a felismerés folyamata felgyorsítható.

Az autizmussal és a nyelvtechnológiával kapcsolatos elméleti ismeretek, valamint a szisztematikus áttekintés és a saját kismintás kutatásom eredményei alapján úgy vélem, hogy az elméleti indíttatású, mélyebb megértésre fókuszáló nyelvi, nyelvhasználati sajátosságokat feltáró kutatások eredményeit érdemes lenne összekötni a már rendelkezésre álló nyelvtechnológiai eszközökkel lefolytatott gépi elemzések eredményeivel, annak érdekében, hogy a klinikai alkalmazhatósággal kapcsolatos konkrét módszertani megfogalmazásokat lehessen tenni. Ilyen jellegű magyar nyelvű adatok a dolgozat írásának időpontjában nem álltak rendelkezésre. Éppen ezért munkám általános értelemben vett fő hozadékanak azt tekintem, hogy felhívja a figyelmet a fejlődési zavarok, specifikusan pedig az autizmussal élő személyek nyelvi megnyilatkozásainak automatizált, gépi elemzésében, a nyelvtechnológiai eszközök alkalmazásában rejlő potenciálra és aktuálisan tapasztalható hiányára.

A dolgozatban bemutatott kutatómunka folytatásaként egyrészt a szisztematikus áttekintés kibővítését (például a konferenciaközlemények és az áttekintés idejét követően publikált tanulmányok, könyvek, könyvfejezetek feldolgozását) tervezem, másrészt a hipotézisek teszteléséhez szóbeli megnyilatkozások vizsgálatát, valamint a rendelkezésre álló szövegek további manuális annotálását.

Ahogy az Newton és munkatársai (2009), valamint a saját kutatási adataim alapján látszik, a gépi elemzés nem tárt fel csoportközi különbségeket az értékelés, affektivitás és kogníció területéhez kapcsolódó szóhasználati gyakoriságban, aminek egyik lehetséges magyarázatát nyújtja a feladatok teljesítésének körülménye, az ún. disztális közeg. További szóbeli nyelvi produkciók elemzésével az a kérdés is vizsgálhatóvá válik, hogy a társas interakciót igénylő kontextus megléte vagy hiánya valóban megmutatkozik-e csoportközi különbségekben is vagy sem, illetve az autizmussal élő csoporton belüli teljesítménykülönbségben, specifikusabban a nyelvi, nyelvhasználati sajátosságok minőségi és mennyiségi eltéréseiben.

A kézi annotálás eredményei alapján úgy tűnik, hogy jelentős eltérés van az evaluatív, azaz értékelő és affektivitással kapcsolatos, valamint a kognícióhoz tartozó kifejezések gépi és kézi azonosításában. Ez egyrészt abban nyilvánul meg, hogy a kogníció területével kapcsolatos kifejezések esetében hasonló abszolút és relatív gyakorisági számokat találtam, másrészt az evaluatív kifejezések esetében nagyobb eltérést ugyanezekben a mutatókban a kétféle módszertant alkalmazva. Ugyanakkor a kézi elemzés alacsony elemszáma miatt mindenképpen további minta szükséges az így levont következtetés alátámasztásához. A gépi és a kézi elemzés abszolút és relatív gyakorisági adataiban mutatkozó nagyobb eltérés az evaluatív kategóriában összhangban van a számítógépes nyelvészeti, a nyelvtechnológiai szakirodalomban a természetesnyelv-feldolgozás kapcsán megfogalmazódó problémákkal, amelyek a szentiment- és érzelelemzés területét jelentős mértékben érintik, így például a doménspecifitás, illetve a kontextusérzékenység problémakörével.

A kutatás folytatásaként azonban fontosnak tartom a szövegek más elemzőláncsal, például az emtsv-vel vagy huSpaCy-vel való feldolgozását, neurális nyelvmodellek tesztelését is, hogy kiderüljön, a kérdésfeltevés, a vizsgált probléma szempontjából melyik megoldás alkalmazhatósága lehet kedvezőbb, informatívabb.

Hivatkozások

- American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5. kiadás). American Psychiatric Association.
- Babarczy, A., Gábor, B., & Serény, A. (2014). Nyelvtechnológia és a pszicholingvisztika számítógépes modelljei. In Pléh Cs. & Lukács Á. (szerk.), *Pszicholingvisztika 1-2* (pp. 411-441). Akadémiai.
- Bagi, A., Gosztolya, G., Szalóki, S., Szendi, I., & Hoffmann, I. (2019). Szkizofrénia azonosítása spontán beszéd temporális paramétereit alapján – egy pilot kutatás eredményei. In Berend G., Gosztolya G., Vincze V. (szerk.), *XV. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 189–201). Szegedi Tudományegyetem Informatikai Intézet.
- Baron-Cohen, S., & Bolton, P. (2000). *Autizmus*. Osiris.
- Baumann T., Majoros, J., Pelcz, K., Schmidt, I., Szita, Sz., & Vermeki, B. (2020). Bemutatkozik a Korpusznyelvészeti és Szakmódszertani Munkacsoport. *Hungarológiai Évkönyv 21*. 32-41.
- Bishop, D. V. M. (2014). Pragmatic language impairment: A correlate of SLI, a distinct subgroup, or part of the autistic continuum. In Bishop, D. V. M. & Leonard, L. B. (szerk.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome* (pp. 99-113). Psychology Press Ltd.
- Bóna, J. (2017). GABI – Gyermeznyelvi beszédatadtbázis a kutatásban. In Bóna J. (szerk.), *Új utak a gyermeknyelvi kutatásokban* (pp. 35-50). ELTE Eötvös Kiadó.
- Boorse, J., Cola, M., Plate, S., Yankowitz, L., Pandey, J., Schultz, R. T., & Parish-Morris, J. (2019). Linguistic markers of autism in girls: evidence of a “blended phenotype” during storytelling. *Molecular Autism*, 10(1), 1-12.
- Boyd, R., Ashokkumar, A., Seraj, S., & Pennebaker, J. (2022). *The Development and Psychometric Properties of LIWC-22*. University of Texas at Austin.
- Calvo, R.A., Milne, D.N., Hussain, M.S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering* 23(5), 649–685.
- Casanova, M. F. (2014). The neuropathology of Autism. In Volkmar, F. R., Paul, R., Klin, A. & Cohen, D. J. (szerk.), *Handbook of Autism and Pervasive Developmental Disorders, Diagnosis, Development, Neurobiology, and Behavior* (pp. 497-532). John Wiley Sons.
- Chawarska, K., Macari, S. L., Volkmar, F. R., Kim, S. H., & Shic, F. (2014). ASD in infants and toddlers. In Volkmar, F. R., Paul, R., Klin, A. & Cohen, D. J. (szerk.), *Handbook of*

- Autism and Pervasive Developmental Disorders, Diagnosis, Development, Neurobiology, and Behavior* (pp. 121-148). John Wiley and Sons.
- Chojnicka, I., & Wawer, A. (2020). Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction. *PLoS One*, *15*(3), e0229985. <https://doi.org/10.1371/journal.pone.0229985>
- Conte E., Ornaghi V., Grazzani I., Pepe A., & Cavioni V. (2019). Emotion Knowledge, Theory of Mind, and Language in Young Children: Testing a Comprehensive Conceptual Model. *Frontiers in Psychology*, *10*. <https://doi.org/10.3389/fpsyg.2019.02144>
- Csepregi, A., & Stefanik, K. (2012). *Autizmus spektrum zavarral élő gyermekek, tanulók komplex vizsgálatának diagnosztikus protokollja*. Educatio.
- Csepregi, A. & Stefanik, K. (2019). Javaslatok az autizmuspektrum-zavar kisgyermekkorai figyelmeztető jeleinek azonosításához. In Csepregi A. (szerk.), *Ajánlások: A fejlődési zavarok és az érzékszervi fogyatékoságok korai szűréséhez, állapotmegismeréséhez* (pp. 215-262). Családbarát Ország Nonprofit Közhasznú Kft.
- De Bruyne, L., Verhoeven, B., & Daelemans, W. (2018). Stylometric text analysis for dutch-speaking adolescents with autism spectrum disorder. *Computational Linguistics in the Netherlands Journal*, (8), 3-23.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA. <https://doi.org/10.48550/arXiv.1810.04805>
- Drávucz, F., Szabó, M., & Vincze, V. (2017). Szentiment- és emóciósótárak eredményességének mérése emóció- és szentimentkorpuszokon. In: Vincze V. (szerk.), *XIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 228-239). Szegedi Tudományegyetem Informatikai Intézet.
- Ehmann B., Csertő I., Ferenczhalmy R., Fülöp É., Hargitai R., Kővágó P., Pólya T., Szalai K., Vincze O. & László J. (2014). Narratív kategóriális tartomelemzés: a NARRCAT. In Tanács A., Varga V. & Vincze V. (szerk.), *X. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2014* (pp. 136-147). Szegedi Tudományegyetem Informatikai Tanszékcsoport.
- Emberi Erőforrások Minisztérium (2020). *Az Emberi Erőforrások Minisztériuma egészségügyi szakmai irányelve az autizmusról/autizmus spektrum zavarokról*. Emberi Erőforrások Minisztériuma.

- Feldmann, Á., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang, Z. Gy., & Váradi, T. (2021) HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In: Berend G., Gosztolya G. & Vincze V. (szerk.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 29-36). Szegedi Tudományegyetem TTIK.
- Fletcher-Watson, S., & Happé, F. (2019). *Autism. A New Introduction to Psychological Theory and Current Debate*. Routledge.
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681-694.
- Fombonne, E. (2020). Epidemiological controversies in autism. *Swiss Archives of Neurology, Psychiatry and Psychotherapy*, 171(01).
- Fombonne, E., MacFarlane, H., Salem, A. C., & Zuckerman, K. E. (2022). Epidemiological surveys of ASD: Current findings and new directions. In Matson, J. L. & Sturmey, P. (szerk.), *Handbook of Autism and Pervasive Developmental Disorder: Assessment, Diagnosis, and Treatment* (pp. 135-184). Springer International Publishing.
- Frith, U. (1991). *Autizmus. A rejtély nyomában*. Kapocs Kiadó.
- Frith, U. (2003). *Autism: Explaining the enigma*. Blackwell Publishing.
- Frith, U., & Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *British Journal of Developmental Psychology*, 1(4), 329-342.
- Gaigg, S. B., Cornell, A. S., & Bird, G. (2018). The psychophysiological mechanisms of alexithymia in autism spectrum disorder. *Autism*, 22(2), 227-231.
- Gormley, E., Ryan, C., & McCusker, C. (2022). Alexithymia is associated with emotion dysregulation in young people with autism spectrum disorder. *Journal of Developmental and Physical Disabilities*, 34(1), 171-186.
- Gósy, M., Gyarmathy, D., Horváth, V., Grácsi, T., Beke, A., Neuberger, T., & Nikléczy, P. (2012). BEA: Beszélt nyelvi adatbázis. In: Gósy M. (szerk.), *Beszéd, adatbázis, kutatások* (pp. 9-24). Akadémiai Kiadó.
- Gosztolya, G., Tóth, L., Svindt, V., Bóna, J., & Hoffmann, I. (2022). Sclerosis Multiplex hangalapú felismerése akusztikai alapú beágyazások használatával. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 151-160). Szegedi Tudományegyetem, Informatikai Intézet.
- Grácsi, T. E., Huszár, A., Krepsz, V., Száraz, B., Damásdi, N., & Markó, A. (2020). Longitudinális korpusz magyar adatközlőkről. In Berendi G., Gosztolya G., & Vincze V. (szerk.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 103-114). Szegedi Tudományegyetem Informatikai Intézet.

- Gyarmathy, D., & Neuberger, T. (2015). Egy hiánypótló adatbázis: a Tini BEA. In: Gósy, M. (szerk.), *Beszéd kutatás* (pp. 209-221). MTA Nyelvtudományi Intézet.
- Györi, M. (2003). A neurokognitív fejlődés moduláris zavarai. az autizmus. In Gulyás B., Pléh Cs. & Kovács Gy. (szerk.), *Kognitív idegtudomány* (pp. 738-759). Osiris.
- Györi, M. (2008). Viselkedéskontroll és megismerés: a végrehajtó működések. In Csépe V., Györi M. & Ragó A. (szerk.). *Általános pszichológia 3. Nyelv, tudat, gondolkodás* (pp. 322-336). Osiris.
- Györi, M. (2012). A neurokognitív fejlődési zavarok viselkedésgenetikája. In Bereczkei T. & Hoffmann Gy. (szerk.), *Gének, gondolkodás, személyiség. Bevezetés a humán viselkedésgenetikába* (pp. 237-273). Akadémiai Kiadó.
- Györi, M. (2014). Nyelv, kommunikáció és megismerés atipikus mintázatai és kapcsolatai autizmus spektrum zavarokban. In Pléh Cs. & Lukács Á. (szerk.), *Pszicholingvisztika 1-2* (pp.1345-1382). Akadémiai.
- Györiné Stefanik K. (2005). *Az iskoláskorra vonatkozó prognózis kérdése autizmusban – a korai képességmintázat és a naiv tudatelméleti vonatkozású viselkedések szerepe* [Doktori (PhD) disszertáció, Eötvös Loránd Tudományegyetem Pszichológiai Doktori Iskola].
- Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2), 101-119.
- Happé, F. G. (1994a). *Autism. An introduction to psychological theory*. University College of London Press.
- Happé, F. G. (1994b). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental disorders*, 24(2), 129-154.
- Happé, F. G. (1997). Central coherence and theory of mind in autism: Reading homographs in context. *British Journal of Developmental Psychology*, 15(1), 1-12.
- Hoffmann, I., Tóth, L., Gosztolya, G., Szatlóczki, G., Vincze, V., Kárpáti, E., Pákáski, M., Kálmán, J. (2017). Beszédfelismerés alapú eljárás az enyhe kognitív zavar automatikus felismerésére spontán beszéd alapján. In Bánréti Z. (szerk.), *Általános Nyelvészeti Tanulmányok XXIX: Kísérletes nyelvészet* (pp. 385–405). Akadémiai.
- Honnibal, M. (2015). Introducing spaCy. Letöltve 2022 november, <https://explosion.ai/blog/introducing-spacy>
- Howlin, P. (2014). Outcome in Adult with Autism Spectrum Disorders. In Volkmar, F. R., Paul, R., Klin, A. & Cohen, D. J. (szerk.), *Handbook of Autism and Pervasive*

- Developmental Disorders, Diagnosis, Development, Neurobiology, and Behavior* (pp. 97-117). John Wiley and Sons.
- Jolliffe, T., & Baron-Cohen, S. (1999). A test of central coherence theory: linguistic processing in high-functioning adults with autism or Asperger syndrome: is local coherence impaired?. *Cognition*, 71(2), 149-185.
- Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, 116(9), 3476-3481.
- Jurafsky, D., & Martin, J. H. (éd.n.). *Speech and Language Processing* (3. kiadás, tervezet). Letöltve 2022 november, <https://web.stanford.edu/~jurafsky/slp3/>
- Karimi, P., Kamali, E., Mousavi, S. M., & Karahmadi, M. (2017). Environmental factors influencing the risk of autism. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 22:27. <https://doi.org/10.4103/1735-1995.200272>
- Kim, S. H., Paul, R., Tager-Flusberg, H., & Lord, C. (2014). Language and communication in autism. In Volkmar, F. R., Paul, R., Klin, A., & Cohen, D. J. (szerk.), *Handbook of Autism and Pervasive Developmental Disorders, Diagnosis, Development, Neurobiology, and Behavior* (pp. 230-263). John Wiley and Sons.
- Korkmaz, B. (2011). Theory of mind and neurodevelopmental disorders of childhood. *Pediatric Research*, 69(8), 101-108.
- Kumar, M., Gupta, R., Bone, D., Malandrakis, N., Bishop, S., & Narayanan, S. S. (2016). Objective Language Feature Analysis in Children with Neurodevelopmental Disorders During Autism Assessment. In *Interspeech* (pp. 2721-2725).
- Lee, M., Martin, G. E., Hogan, A., Hano, D., Gordon, P. C., & Losh, M. (2018). What's the story? A computational analysis of narrative competence in autism. *Autism*, 22(3), 335-344.
- Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- Liu, Z., Proctor, L., Collier, P. N., & Zhao, X. (2021). Automatic diagnosis and prediction of cognitive decline associated with alzheimer's dementia through spontaneous speech. In *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (pp. 39-43). IEEE. <https://doi.org/10.1109/ICSIPA52582.2021.9576784>
- Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule* (2. kiadás). Western Psychological Corporation.

- Lyall, K., Schmidt, R. J., & Hertz-Picciotto, I. (2014). Maternal lifestyle and environmental risk factors for autism spectrum disorders. *International Journal of Epidemiology*, 43(2), 443–464.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3. kiadás). Lawrence Erlbaum Associates.
- Mátyus, K., & Orosz, Gy. (2014). MONYEK: Morfológiailag egyértelműsített óvodai nyelvi korpusz. *Beszédkutató* 22. 237–245.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781. Letöltve 2022 november, <https://arxiv.org/abs/1301.3781>
- Milosavljevic, B., Carter Leno, V., Simonoff, E., Baird, G., Pickles, A., Jones, C. R. G, Erskine, C., Charman, T., & Happé, F. (2016). Alexithymia in adolescents with autism spectrum disorder: Its relationship to internalising difficulties, sensory modulation and social cognition. *Journal of Autism and Developmental Disorders*, 46(4), 1354-1367.
- Nemeskey, D.M. (2020). *Natural Language Processing Methods for Language Modeling* [Doktori disszertáció, Eötvös Loránd Tudományegyetem].
- Nemeskey, D. M. (2021). Introducing huBERT. In: Berend G., Gosztolya G. & Vincze V. (szerk.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 3-14). Szegedi Tudományegyetem TTIK.
- Newton, A. T., Kramer, A. D., & McIntosh, D. N. (2009). Autism online: a comparison of word usage in bloggers with and without autism spectrum disorders. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 463-466).
- Nguyen, T., Duong, T., Phung, D., & Venkatesh, S. (2014). Affective, linguistic and topic patterns in online autism communities. In *Web Information Systems Engineering–WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part II 15* (pp. 474-488). Springer International Publishing.
- Nguyen, T., Duong, T., Venkatesh, S., & Phung, D. (2015). Autism blogs: Expressed emotion, language styles and concerns in personal and community settings. *IEEE Transactions on Affective Computing*, 6(3), 312-323.
- Nguyen, T., Phung, D., & Venkatesh, S. (2013). Analysis of psycholinguistic processes and topics in online autism communities. In *2013 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
- Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., & Farkas, R. (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In Berend G., Gosztolya G. &

- Vincze V. (szerk.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 59-73). Szegedi Tudományegyetem TTIK.
- Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991a). Executive function deficits in high-functioning autistic individuals: Relationship to theory of mind. *Journal of Child Psychology and Psychiatry*, 32(7), 1081-1105.
- Ozonoff, S., Rogers, S. J., & Pennington, B. F. (1991b). Asperger's syndrome: Evidence of an empirical distinction from high-functioning autism. *Journal of Child Psychology and Psychiatry*, 32(7), 1107-1122.
- Ozonoff, S., South, M., & Provençal, S. (2005). Executive functions. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (szerk.), *Handbook of autism and pervasive developmental disorders: Diagnosis, development, neurobiology, and behavior* (pp. 606–627). John Wiley and Sons.
- Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). *Linguistic Inquiry and Word Count (LIWC)* [Computer software]. LIWC Inc.
- Pellicano, E. (2011). Psychological models of autism: An overview. In Roth, I. & Rezaie, P. (szerk.), *Researching the Autism Spectrum: Contemporary Perspectives* (pp. 219-265). Cambridge.
- Pellicano, E. (2020). Commentary: Broadening the research remit of participatory methods in autism science—a commentary on Happé and Frith (2020). *Journal of Child Psychology and Psychiatry*, 61(3), 233-235.
- Pellicano, E., & den Houting, J. (2022). Annual Research Review: Shifting from 'normal science' to neurodiversity in autism science. *Journal of Child Psychology and Psychiatry*, 63(4), 381-396.
- Pereira, J. A., Macêdo, D., Zanchettin, C., de Oliveira, A. L. I., & do Nascimento Fidalgo, R. (2022). Pictobert: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202, 117231. <https://doi.org/10.1016/j.eswa.2022.117231>
- Quill, K. A., & Stansberry Brusnahan, L. L. (2017). *Do-watch-listen-say: Social and communication intervention for autism spectrum disorder*. Paul H Brookes Publishing.
- Ramírez, V. A., & Ruetti, E. (2021). Language modulation on emotional valence tasks in preschoolers. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*. <https://doi.org/10.1007/s12144-021-01868-4>
- Renteria-Vazquez, T., Brown, W. S., Kang, C., Graves, M., Castelli, F., & Paul, L. K. (2021). Social inferences in agenesis of the corpus callosum and autism: Semantic analysis and topic modeling. *Journal of Autism and Developmental Disorders*, 52(2), 1-15.

- Rivet, T. T., & Matson, J. L. (2011). Review of gender differences in core symptomatology in autism spectrum disorders. *Research in Autism Spectrum Disorders*, 5(3), 957-976.
- Rutter, M. (2011). Progress in Understanding Autism: 2007–2010. *Journal of Autism and Developmental Disorders*. 41(4), 395-404.
- Rutter, M., & Thapar, A. (2014). Genetics of autism spectrum disorders. In: Volkmar, F. R., Paul, R., Klin, A. & Cohen, D. J. (szerk.), *Handbook of Autism and Pervasive Developmental Disorders, Diagnosis, Development, Neurobiology, and Behavior* (pp. 411-424.). John Wiley and Sons.
- Sebők M. (szerk.) (2016). *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban*. L'Harmattan Kiadó.
- Simon, E. (2013). *Approaches to Hungarian Named Entity Recognition* [Doktori disszertáció, Budapesti Műszaki és Gazdaságtudományi Egyetem].
- Simon, E., Indig, B., Kalivoda, Á., Mittelholcz, I., Sass, B., & Vadász, N. (2020). Újabb fejlemények az e-magyar háza táján. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 29-42). Szegedi Tudományegyetem TTIK.
- Snowling, M., & Frith, U. (1986). Comprehension in “hyperlexic” readers. *Journal of Experimental Child Psychology*, 42(3), 392-415.
- Stefanik, K., & Prekop, Cs. (2015). Autizmus spektrum zavarok. In Balázs J. & Miklósi M. (szerk.), *A gyermek- és ifjúkor pszichés zavarainak tankönyve* (pp. 61-67). Semmelweis Kiadó.
- Szabó, M. (2018). *A szentimentérték módosulásának a problémája a magyar nyelvű szövegek szentimentelemzésében, különös tekintettel az értékvesztésre és az értékváltásra* [Doktori értekezés, SZTE BTK NYDI].
- Tager-Flusberg, H. (1992). Autistic children's talk about psychological states: Deficits in the early acquisition of a theory of mind. *Child Development*, 63(1), 161-172.
- Tager-Flusberg, H. (1999). A psychological approach to understanding the social and language impairments in autism. *International Review of Psychiatry*, 11(4), 325-334.
- Tager-Flusberg, H. (2000). Understanding the language and communicative impairments in autism. In Glidden, L. M. (szerk.), *International review of research in mental retardation* (pp. 185-205). Academic Press.
- Tager-Flusberg, H. (2006). Defining language phenotypes in autism. *Clinical Neuroscience Research*, 6(3-4), 219-224.

- Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism Research*, 6(6), 468-478.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24-54.
- Vetráb, M., Egas López, J. V., Balogh, R., Imre, N., Hoffmann, I., Tóth, L., Pákáski, M., Kálmán, J., & Gosztolya, G. (2022). Enyhe kognitív zavar automatikus felismerése szekvenciális autoenkóder használatával. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 175-184). Szegedi Tudományegyetem Informatikai Intézet.
- Volkmar, F. R., Rowberry, J., Vinck-Baroody, O. D., R Gupta, A., Leung, J., Meyers, J., & Wiesner, L. A. (2014). Medical care in autism and related conditions. In: Volkmar, F. R., Paul, R., Klin, A. & Cohen, D. J. (szerk.), *Handbook of Autism and Pervasive Developmental Disorders, Diagnosis, Development, Neurobiology, and Behavior* (pp. 532-555.). John and Wiley Sons.
- Wetherby, A. M. (2008). Understanding and Measuring Social Communication in Children with Autism Spectrum Disorders. In Charman, T. & Stone, W. (szerk.), *Social and communication development in autism spectrum disorders: Early identification, diagnosis, and intervention* (pp. 3-35). Guilford Press.
- White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, 80(4), 1097-1117.
- Wing, L. (1996). *The autistic spectrum: A guide for parents and professionals*. Constable.
- World Health Organization (2018). *ICD-11 for mortality and morbidity statistics (ICD-11 MMS)* Letöltve 2022 november, URL: <https://icd.who.int/browse11/l-m/en>
- World Health Organization (2022). *BNO-11: Mentális zavarok*. Animula.

Internetes hivatkozások

NarrCat <https://webnarrcat.ttk.hu>

PoltextLab <https://poltextlab.ttk.hu/en>

ELTE Research Center for Computational Social Science <https://rc2s2.elte.hu>