

DOKTORI (Ph.D.) ÉRTEKEZÉS

Üveges István

## **Közérthetőség és automatizáció**

- kísérletek a jog, természetesnyelv-feldolgozás és informatika határán

Elméleti nyelvészet doktori program

Nyelvtudományi Doktori Iskola

Szegedi Tudományegyetem

Témavezető:

Dr. Vincze Veronika

Tudományos főmunkatárs

HUN-REN SZTE Mesterséges Intelligencia Kutatócsoport

Szeged, 2024

*“Clear writing from your government is a civil right.”*

Albert Arnold Gore Jr., 1998

Former Vice President of the United States

## **Köszönetnyilvánítás**

Köszönettel tartozom témavezetőmnek, Dr. Vincze Veronikának, hogy graduális hallgatóként elindított azon az úton, amely során betekintést nyerhettem a tudományos kutatásba a számítógépes nyelvészet – természetesnyelv-feldolgozás területén. Köszönöm, hogy az évek során számtalan tanáccsal, biztató szóval (és sok esetben implementációs segédlettel) segítette munkámat, és amiért mindvégig biztatott.

Köszönöm az SZTE BTK Általános Nyelvészeti Tanszék oktatóinak a fáradhatatlan lelkesedést mindazon előadások és szemináriumok alkalmával, amelyeken az évek során részt vehettem.

Köszönöm Dr. Tribl Norbertnek, hogy a dolgozat jogi témájú kérdéseivel bármikor bizalommal fordulhattam hozzá, és hogy nem volt olyan kérdés, amelyre ne válaszolt volna készségesen, illetve Dr. Pongó Tamásnak az eredmények értékelése során nyújtott fáradhatatlan segítségéért.

Hálával tartozom a Nemzeti Adó- és Vámhivatal Közérthetőségi programjában dolgozóknak, amiért munkámat segítették az általuk kezelt dokumentumok rendelkezésre bocsátásával. Nélkülük a dolgozat jelentős része nem készülhetett volna el.

Végül, de nem utolsó sorban örök hálával tartozom szüleimnek, akik megismertették velem a tudomány szeretetét és megértéssel, türelemmel és szeretettel támogattak azon az úton, amely a dolgozat elkészületéhez vezetett. El nem múló odaadásuk átsegített a nehézségeken és mindig emlékeztetett a céljaimra. A dolgozatot nektek ajánlom.

Ezúton nyilatkozom, hogy az értekezés saját, önálló munkám.

Szeged,

2024. 02. 11.

Üveges István

# Tartalom

|  |    |
|--|----|
| Rövidítések jegyzéke.....  | 1  |
| 1.Bevezetés.....   | 2  |
| 1.1 Kutatási kérdések és hipotézisek .....                                 | 6  |
| 1.2 Az értekezés szerkezete.....   | 9  |
| 1.3 Adatok elérhetősége .....  | 11 |
| 2. A pontos megfogalmazás kritériuma.....                                  | 12 |
| 2.1 Normavilágosság.....   | 13 |
| 2.2 Kinek szól a norma? .....  | 16 |
| 2.3 A jogértelmezés szerepe.....   | 17 |
| 2.3.1 A jogi norma természete .....  | 18 |
| 2.3.2 Jogértelmezési módszerek.....  | 20 |
| 2.3.3 Ki „jogosult” a jogot értelmezni? .....                              | 22 |
| 2.4 Szaknyelviség.....   | 23 |
| 2.4.1 Nyelvváltozatok .....  | 24 |
| 2.4.2 A szaknyelviség által támasztott kritériumok .....                   | 25 |
| 2.4.3 Szaknyelviség és közérthetőség kollíziója .....                      | 27 |
| 2.4.4 Normavilágosság és/vagy közérthetőség?.....                          | 27 |
| 3. Megközelítési módok a <i>közérthetőség</i> fogalmával kapcsolatban..... | 29 |
| 3.1 Pszicholingvisztika.....   | 31 |
| 3.1.1 A megértés folyamata és tényezői.....                                | 31 |
| 3.1.2 A megértést nehezítő pszicholingvisztikai tényezők .....             | 33 |
| 3.1.3 Pszicholingvisztika a tárgyalóteremben .....                         | 35 |
| 3.2 Egyszerű lexikai alapú módszerek .....                                 | 38 |
| 3.2.1 Az érthetőség mérése képletekkel: olvashatósági formulák.....        | 38 |
| 3.2.2 A megközelítés kritikái .....  | 40 |
| 3.3 Plain Language .....   | 41 |
| 3.3.1 Történeti előzmények.....  | 41 |
| 3.3.2 A Plain Language Movement térnyerése .....                           | 44 |
| 3.3.3 Mozgalomból kutatási terület.....                                    | 45 |
| 3.3.4 Federal Plain Language Guidelines.....                               | 47 |
| 3.4 A „Jog és nyelv” irányzat.....   | 49 |

|   |     |
|---|-----|
| 3.5 Korpusznyelvészet .....   | 51  |
| 3.5.1 Az „adatvezérelt fordulat” .....                              | 51  |
| 3.5.2 Miskolc Jogi Korpusz.....                                     | 53  |
| 3.6 Plain Language: ISO-standard .....                              | 56  |
| 4. A közérthetőségi törekvések eddigi eredményei.....               | 58  |
| 4.1 Egyesült Államok .....  | 58  |
| 4.2 Európai Unió .....  | 61  |
| 4.2.1 Uniós szint.....  | 62  |
| 4.2.2 Tagállami szint .....   | 66  |
| 4.2.2.1 Svédország .....  | 66  |
| 4.2.2.2 Norvégia .....  | 68  |
| 4.2.2.3 Dánia .....   | 70  |
| 4.2.2.4 Finnország .....  | 70  |
| 4.2.2.5 Németország.....  | 71  |
| 4.3 Magyarország .....  | 75  |
| 4.3.1 A Kúria joggyakorlat-elemző munkacsoportjának véleménye ..... | 75  |
| 4.3.2 A „Jog és nyelv” irányzat néhány hazai eredménye .....        | 78  |
| 4.3.3 Nemzeti Adó- és Vámhivatal – közérthetőségi program I. ....   | 80  |
| 4.3.4 További eredmények .....                                      | 81  |
| 4.4 Összegzés .....   | 82  |
| 5. A közérthetőség kritikái .....                                   | 86  |
| 5.1 A probléma leegyszerűsítése .....                               | 86  |
| 5.2 A közérthetőség fogalmának alulspecifikáltsága .....            | 87  |
| 5.3 A pragmatikai kontextus figyelmen kívül hagyása .....           | 91  |
| 5.4 A közérthetőség helye a jogi domén szövegeiben.....             | 94  |
| 5.5 Gyakorlati lehetőségek – elméleti összefoglalás.....            | 95  |
| 6. Gépi tanítási kísérletek előkészítése.....                       | 99  |
| 6.1 Felügyelt és felügyelet nélküli gépi tanítás .....              | 101 |
| 6.2 Korpusz közérthetőségi osztályozáshoz .....                     | 102 |
| 6.2.1 Nemzeti Adó- és Vámhivatal – Közérthetőségi Program II. ....  | 103 |
| 6.2.2 Bemeneti fájlok .....   | 104 |
| 6.3 Adatelőkészítés.....  | 105 |
| 6.3.1 Mondatszegmentálás .....                                      | 105 |
| 6.3.2 Szétválasztás „eredeti” és „átfogalmazott” mondatokra .....   | 108 |
| 6.3.3 Adattisztítás és előfeldolgozás .....                         | 110 |

|  |     |
|--|-----|
| 6.4 A végleges korpusz .....   | 111 |
| 7. Modelltanítás SVM-mel .....   | 113 |
| 7.1 Eredmények.....  | 114 |
| 7.2 Következtetések .....  | 116 |
| 8. Modelltanítás a fastText alkalmazásával.....  | 117 |
| 8.1 Eredmények.....  | 118 |
| 8.2 Következtetések .....  | 125 |
| 9. Modelltanítás BERT-tel .....  | 126 |
| 9.1 Finomhangolás .....  | 126 |
| 9.2 Adatok előkészítése.....   | 127 |
| 9.2 Eredmények.....  | 129 |
| 10. Gépi tanulási kísérletek összefoglaló értékelése .....                                 | 132 |
| 10.1 Pragmatika.....   | 136 |
| 10.2 Az LLM-ek „pragmatikai kompetenciája” .....   | 139 |
| 10.3 Kitekintés: generatív mesterséges intelligencia .....                                 | 144 |
| 11. Szakértői kiértékelés .....  | 150 |
| 12. Szabálykészlet .....   | 156 |
| 12.1 Áttekintés .....  | 156 |
| 12.2 Lexikai szabályok.....  | 158 |
| 12.3 Szintaktikai szint .....  | 167 |
| 12.4 Bekezdésekre vagy a teljes szövegre vonatkozó ajánlások.....                          | 170 |
| 13. A felület és működés bemutatása .....  | 174 |
| 13.1 Extraktív kivonatolás.....  | 176 |
| 13.2 További implementációs részletek .....  | 177 |
| 13.3 Az alkalmazás kimenetének áttekintése .....   | 184 |
| 13.4 A fejlesztéssel kapcsolatos összefoglaló .....  | 185 |
| 14. A hipotézisek értékelése és a kutatási kérdésekre adható válaszok összefoglalása ..... | 186 |
| 15. Összefoglalás.....   | 191 |
| Függelék .....   | 194 |
| Hivatkozott esetek és jogszabályok.....  | 208 |
| Irodalomjegyzék.....   | 209 |

## Rövidítések jegyzéke

|          |   |
|----------|---|
| AB       | Alkotmánybíróság  |
| FPLG     | Federal Plain Language Guidelines                                 |
| MJK      | Miskolc Jogi Korpusz  |
| ML       | Machine Learning – <i>gépi tanulás</i>                            |
| NLP      | Natural Language Processing - <i>természetesnyelv-feldolgozás</i> |
| PLM      | Plain Language Movement   |
| Szja tv. | 1995. évi CXVII. törvény  |
| SzT      | Szeged Dependency Treebank  |
| NB       | Naive Bayes   |
| SVM      | Support Vector Machine – <i>szupport vektor gép</i>               |
| LR       | Logistic Regression   |
| LVC      | Light Verb Contructions – <i>funkcióigék</i>                      |
| regex    | REGular Expression – <i>reguláris kifejezés</i>                   |
| TB       | jogi terminológiai adatbázis                                      |
| ES       | Extractive Summarization – <i>extraktív kivonatolás</i>           |
| AS       | Abstractive Summarization – <i>absztraktív kivonatolás</i>        |
| LLM      | Large Language Model – <i>nagy nyelvi modell</i>                  |
| sota     | State-of-the-art – <i>legmodernebb, korszerűbb</i>                |



## 1.Bevezetés

A Belügyminisztérium Önkormányzati Koordinációs Irodája 2015 és 2019 között egy kiemelt projektet<sup>1</sup> indított, melynek címe „Önkormányzati fejlesztések figyelemmel kísérése II.” (ÖFFK II.) volt. Ennek fő célja az önkormányzatok hatékony működésének és azok fenntarthatóságának növelése volt. Ezt a célt egyúttal elsődleges prioritásként határozták meg a 2021-től 2027-ig tartó időszakra vonatkozóan is.

A projekt keretében több kérdésre is választ kerestek. Ilyen volt például, hogy a hivatali szereplők által adott információk (köz)érthetősége mennyiben befolyásolja a beléjük vetett bizalom fokát, vagy éppen, hogy a kapott tájékoztatás érthetősége hogyan korrelál a lakosság tájékozódási kedvével.

A projekt során utóbbi körében (reprezentatív mintán) végzett felmérés is készült (Budai 2017). Ez elemezte többek között a lakosságnak a hivatali működéssel való elégedettsége és hivatali a kommunikációs formák összefüggéseit, ezen kommunikációs formák és a lakosság a hivattal szembeni bizalmi szintjének összefüggését, valamint a tájékozódás és információszerezés gyakorlatát. A felmérés keretében a válaszadóknak egy ötfokozatú skálán kellett értékelniük megadott szempontokat. Ezek között szerepeltek olyan kérdések is, amelyek a közhivatalokkal kapcsolatos általános elégedettséget, az egyes közszerelőket kapcsolatos bizalmi szintet, valamint fenti kettőnek a kapott információk közérthetőségével való összefüggését értékelték. Az adatok tükrében a tanulmány több, a közérthetőséget érintő megállapítást is tesz (uo. 5-7):

- „a tájékoztatás eredményessége ... szorosan összefügg a hivatal megítélésével. Minél érthetőbbek az önkormányzati közlemények, információk annál magasabb értékeket adtak a válaszadók arra a kérdésre, hogy »az önkormányzat jól végzi-e a dolgát«”,
- „az információk (köz)érthetősége is befolyásolja a bizalom fokát. Érthetetlen közleményeknél a polgármesterrel szembeni bizalmi érték 2,2, míg a teljesen érthető közleményeknél 4,18. Önkormányzati képviselők esetében ugyanezek az értékek 2,0, illetve 3,96”<sup>2</sup>,

---

<sup>1</sup> KÖFOP-2.3.4-VEKOP-15-2016-00002

A programról lásd bővebben: <https://bm-oki.hu/?projectMainType=OFFK2>

<sup>2</sup> Ahol az 1-es érték jelentette az alsó korlátot (egyáltalán nem bízunk meg), az 5-ös pedig a felső korlátot (teljes mértékben megbízunk).

- „ugyancsak meghatározó a tájékozódás gyakoriságánál az önkormányzati közlemények, információk közérthetősége. Az érthetetlen információk megölik a tájékozódási hajlandóságot (4,44) – mintegy leszoktatják a lakosságot az érdeklődésről –, míg a jól érthető kommunikátumok meghozzák a kedvet a tájékozódáshoz (2,87)”<sup>3</sup>.

A vizsgálat lezárásaként készült kutatási jelentés<sup>4</sup> megállapítja, hogy egy tetszőleges hivatal kommunikációs szempontból akkor tekinthető sikeresnek, ha a lakosok számára érthetően és minél többet kommunikál, valamint, hogy az ügyintézés sikerességét az anyagi helyzet mellett főként az információk mennyisége és közérthetősége határozza meg.

Habár az önkormányzati működés a teljes hivatalos / jogi doménben zajló írott kommunikációnak csak egy szeletét adja, mégis mindezek alapján feltételezhetnénk, hogy a magyar jogrendszerben (a jogforrások szintjén) kiemelt helyet, de legalábbis említést kap a hivatalok közérthető kommunikációja iránti elvárás. Erről azonban szó sincs. A dolgozat lezárásakor Magyarországon nincs hatályban olyan jogszabály, amely a közhivatalok számára előírná, hogy tájékoztatásaikban törekedjenek a laikus olvasó nézőpontjának figyelembevételére, más szóval, hogy azokban közérthetően fogalmazzanak. Hasonlóképpen olyan általánosan elfogadott segédanyag vagy ajánlásrendszer sem létezik, amely számukra ebben segítséget nyújthatna.

Ennek ellenére mind a közigazgatásban, mind pedig az igazságszolgáltatásban alkalmazott nyelvezet kapcsán megemlíthetők olyan kutatások, vizsgálatok és kezdeményezések, amelyek valamiképpen számot vetettek már a közérthetőség kérdésével.

Talán egy egyik legfontosabb megállapítás szerint „az átlátható hatósági eljárás része a jogtudományi értelmezés szerint a megfelelő kommunikáció. Ez magában foglalja a társadalom és az érintettek informálását szolgáló, következetes, érthető és nyílt érintkezést ... a hatósági működéssel, valamint a szervezeti felépítéssel kapcsolatosan.” (Tóth 2023: 26) Átláthatóság alatt itt az az alapelv értendő, amelybe a demokratikus jogállami követelmények (pl. tisztességes eljárás) ügyféli oldala is beleértendő, ezért pedig ennek

---

<sup>3</sup> Az adatok a következő skála szerint adott válaszok átlagaként adódtak 1810 válaszadó visszajelzései alapján a lakosság közügyekkel kapcsolatos tájékozódási gyakorisága tekintetében: 1 – havonta többször, 2 - havonta legfeljebb egyszer, 3 - évente többször, 4 - évente legfeljebb egyszer, 5 – soha.

<sup>4</sup> Online elérhető: <https://docplayer.hu/108554227-Kutatasi-jelentes-kofop-vekop-onkormanyzati-fejlesztések-figyelemmel-kiserese-ii-kutatas-ii.html> (A link ellenőrizve: 2021.08.15.)

kapcsolata a közérthetőséggel megkerülhetetlen. A témáról például Hohmann (2022) írt bővebben a közigazgatás hatósági jogkörben történő eljárásai kapcsán.

Az igazságszolgáltatásban alkalmazott nyelvhasználat kutatása modern értelemben már az 1980-as években megkezdődött. Ez először csak az írott szövegekre korlátozódott, majd a 2000-es évek elején a beszélnyelvi adatok is vizsgálata is megindult (Vinnai 2014a). A Miskolci Egyetemen immár évtizedek óta számos kutatási program zajlik a jogi nyelv és a laikusok interakciójának elemzése kapcsán, elég csak a jogász és nyelvész szakemberek közreműködésével lezajlott „Nyelvhasználat a jogi eljárásban. Nyelvi átfordítás és ténykonstitúció a jogi tényállás megállapításának folyamatában”, valamint „A tisztességes eljárás nyelvi aspektusai. A jogi nyelv hatása a joghoz való hozzáférés érvényesülésére” című (egyenként több éves) kutatásokra gondolni.

A Kúriai határozatokat 2013-ban joggyakorlatelemző munkacsoport vizsgálta a szövegezés és a dokumentumszerkesztés közérthetőséget támogató aspektusainak felderítése érdekében (Orosz: 2014). Az Országos Bírósági Hivatal döntése szerint 2017 a „közérthetőség éve” volt a magyar bíróságokon, amelynek keretében az ország valamennyi bíróságán kötelező képzés is zajlott „Közérthetőség és szakszerűség a bírósági jogalkalmazásban” címmel.

A hivatali / közigazgatási nyelv tekintetében kiemelendő, hogy a Nemzeti Adó- és Vámhivatal saját közlése alapján mintegy 10 éve foglalkozik a közérthető kommunikáció kérdésével. A cél egységes programmá 2019-ben állt össze, amikor is a NAV vezetője szervezeti célként fogalmazta meg a hivatal kommunikációjának közérthetőségét, amely után megkezdődött a szervezett közérthetőségi munka nyelvész szakértők bevonása mellett.

Az említett kutatások és programok (néha implicit módon) igyekeznek számot vetni az olyan kérdésekkel, mint például a „közérthető” fogalmának tisztázása, definíciós kísérlete, annak eldöntése, hogy létezik-e (létezhets-e) objektív kritériumrendszer ennek meghatározására, hogyan kapcsolódik a számítógépes nyelvészet és a vele többé-kevésbé szinonim természetesnyelv-feldolgozás (Natural Language Processing, a továbbiakban NLP) a problémához, és hogy ez utóbbi milyen hozzájárulást tehet a mind jobban a digitalizáció irányába tartó hivatali kommunikáció érthetőbbé tételére kapcsán?

Utóbbi azért is különösen fontos, hiszen a globalizáció és a technológiai fejlődés következményeként megnövekedett információmennyiség a társadalmak kvázi „eljogiasodásához” vezet(ett): a jog egyre több élethelyzetre alkalmaz specifikus szabályokat, amely szabályok mennyiségileg is egyre hangsúlyosabban vannak jelen a

hétköznapokban<sup>5</sup>. Ennek fényében növekvő jelentőséggel lép színre az az igény, hogy az állampolgárok minél többet és minél egyszerűbben megérthessenek jogi normákból, melyek számukra kötelező magatartási szabályokat állapítanak meg, illetve adott esetben az őket érintő konkrét döntésekből.

Habár a Kúrián lezajlott említett projekt során informatikusok bevonása is megtörtént, a közérthető fogalmazás automatizációval történő támogatása kapcsán nem valósult meg átfogó vizsgálat. Nem készült olyan (nyilvánosan elérhető) szoftver sem, amely a digitalizáció által nyújtott előnyöket kihasználva, valamint a bírósági szervezetrendszer általánosan jellemző feszített munkatempót megkönnyítendő támogatni tudná a közérthetőségi szempontok érvényre jutását az eljárások során. Ilyen rendszer egyébként a hivatali dokumentumok felülvizsgálatára sem áll rendelkezésre.

Éppen ezért a dolgozat többek között a fenti kérdéseket is körbejárva arra igyekszik választ keresni, egyben pedig erre egy gyakorlati megoldást kidolgozni, hogy miképpen lehetséges a jogi domén szövegeit „hatékonyan közérthetőbbé tenni”?

A kutatás tudományos jelentősége a témafelvetés azon aspektusában rejlik, mely szerint egy olyan absztrakt szemantikai-pragmatikai kérdésre, mint a hivatalos szövegeknek az átlagos befogadó szempontjából vett érthetősége, a dolgozat elsősorban a természetesnyelv-feldolgozás (számítógépes nyelvészet) eszközkészletével igyekszik megoldást találni. Egy ilyen vállalás esetén elkerülhetetlenül számolni kell a rendelkezésre álló természetesnyelvi elemzők aktuális limitációival, azzal a körülménnyel, hogy sok nyelvi jelenség még a legmodernebb technológiák alkalmazásával sem fogható meg hatékonyan automatikus eszközökkel, valamint a közérthetőség kérdését jellemző általános adathiánnyal.

Ez utóbbi főként arra a körülményre utal, hogy míg olyan alapfeladatok esetében, mint például a szófaji címkézés, a névelemfelismerés, vagy éppen a morfológiai elemzés, ma már elemzőeszközök széles tárháza áll rendelkezésre legyen szó akár nyelvmodellekről vagy

---

<sup>5</sup> A jelenséget jól példázza az az összehasonlítás is, amit Vinnai (2017a: 47) a hazánkban hatályban levő jogszabályok számosságára vonatkozóan, az azok nyelvi megformálása jutó átlagos idő kapcsán közöl. Mint írja; „1938-ban a magyar parlament 39 törvényt fogadott el, míg 2013-ban 254-et, és azóta is évente több mint 200-at. Magyarországon 2018 januárjában 1718 törvény, 308 törvényerejű rendelet, 2228 kormányrendelet és 3346 miniszteri rendelet volt hatályban (ez összesen kb. 7600 jogszabály), s ehhez adódik még hozzá kb. 150 ezer önkormányzati rendelet, több mint 30 ezer hatályos uniós jogi aktus, melyekből évente 2-3000 újabb keletkezik. Ha a nyomtatott Magyar Közlöny terjedelmének változásait vizsgáljuk, azt találjuk, hogy míg 1998-ban 8800 oldal jelent meg, 2016-ban már 35 620 és 68 931 oldal melléklet”.

különböző programnyelveken implementált elemzési láncokról<sup>6</sup>. Saját gépi tanuló modellek építéséhez is rendelkezésre állnak korpuszok a legkülönfélébb kutatási kérdések esetén, legyen szó akár politikai szövegek szentiment- és érzelemelemzéséről (Ring et al. 2023), kérdés-válasz benchmark adatbázisokról (pl. Novák – Novák 2023) vagy például szövegek absztraktív összefoglalóit elkészítő modellek kiértékeléséhez használható korpuszokról (pl. Barta et al. 2023).

Ha azonban olyan korpuszt keresünk, amely közérthetőségi szempontú szövegválogatást tartalmaz, akkor sem magyar, sem pedig (a szerző által ismert) egyéb nyelven nem találunk rendelkezésre állót.

Az algoritmikus megvalósíthatósághoz ezen felül szükséges az elérni kívánt cél pontos meghatározása, illetve olyan célok kitűzése, amelyek nem csak nyelvtudományi szempontból megalapozottak és jogtudományi szempontból elfogadhatóak, de a természetesnyelv-feldolgozás (~számítógépes nyelvészet) jelenleg rendelkezésre álló eszközkészletének képességeit figyelembe véve reális fejlesztési célként is megragadhatók.

A fentiek megoldásához elengedhetetlen egyrészt a közérthetőség kérdését nyelvészeti és jogi szempontból tárgyaló (magyar és nemzetközi) szakirodalom eredményeinek szintetizálása, másrészt olyan informatikai eszközrendszer implementálása, amely képes a megfogalmazott feladatot a jelen technológiai korlátok mellett is megfelelően megoldani. Tekintettel arra, hogy ez utóbbi esetén az elmúlt évtizedben a gépi tanulási és mélytanuló algoritmusok alkalmazása (pl. Devlin et al. 2018, Liu et al. 2019, Peters et al. 2018, Radford – Narasimhan 2018 stb.) vált dominánssá, így ezek alkalmazásához megfelelő mennyiségű tanítóadatot is biztosítani kell.

## **1.1 Kutatási kérdések és hipotézisek**

A dolgozat létrejöttét elsősorban a már említett alapvetések, illetve a közérthetőségi igényeknek a jogrendszer egyes szövegeire vonatkozó érvényesítési lehetőségeinek meghatározása és az ilyen szempontokat tekintetbe vevő (mintegy, intralingvális átfordításon átesett) szövegek létrehozását támogatni képes gyakorlati alkalmazási lehetőségek kimunkálása indokolta.

---

<sup>6</sup> Előbbire jó példa a HuSpaCy keretrendszerben, vagy önálló modellként is használható neurális hálóval képzett nyelvmodellek sora (<https://huggingface.co/huspaCy>), utóbbira a magyarlanc Java nyelven írt elemző (<https://rgai.inf.u-szeged.hu/magyarlanc>).

Ahogy Tóth (2022: 20) megfogalmazza: „a közérthetőség olyan tulajdonsága valamely mondatnak, tantételnek, értelmi vagy erkölcsi igazságnak, szabálynak, amelynél fogva azok jelentése, tartalma mindenki által könnyen felfogható”. Mindemellett utalást tesz a Plain Language Movement nemzetközi hatású mozgalomának eredményeire (Üveges 2021) és arra a körülményre, hogy az irányzat képviselői kritikai észrevételek mellett olyan konkrét javaslatokkal is szolgálnak, amelyek alkalmazása elősegítheti a laikusok szövegértését a jogi nyelvvel való találkozás során. Dobos Csillát idézve (Tóth 2022: 20) felvázolja továbbá az érthetőség alappilléreit, amelyeket

- az egyszerűbb szöveg könnyebb érthetőségében,
- a tagolt és rendezett szöveg könnyebb átláthatóságában,
- a rövidebb, tömörebb szöveg feldolgozási erőfeszítésének csökkenésében, valamint
- az érzelmi motivációra utaló nyelvi-stilisztikai eszközök alkalmazásának célszerűségében határoz meg.

Mindemellett említést tesz a jogi nyelv rétegzettségéről, amely előre vetíti, hogy a dolgozat által támogatni kívánt közérthetőségi erőfeszítések nem juthatnak egységesen érvényre a hivatali / jogi domén teljes spektrumában.

A 2010. évi CXXX. törvény előírásaként is megfogalmazza, hogy a szabályozási tartalom a jogszabály címetjei számára egyértelmű legyen. Tóth (2020: 21) említi továbbá a korábbi jogalkotási törvényt is, amely még nevesítette a közérthető megfogalmazás kritériumát a következők szerint: „...jogszabályokat a magyar nyelv szabályainak megfelelően, világosan és közérthetően kell megszövegezni.”<sup>7</sup>.

A fentiek a következő hipotézisek megfogalmazásához vezettek:

- i. A hazai és/vagy a nemzetközi szakirodalom alapján a közérthetőségnek nemcsak definíciója létezik, de meghatározható nyelvi jellemzők konkrét csoportja is, amely rontja egy szöveg könnyű érthetőségét.
- ii. A jogi nyelvnek létezik olyan rétege, amelyben a közérthetőség, mint célkitűzés a gyakorlatban is érvényesíthető.

---

<sup>7</sup> 1987. évi XI. törvény 18. § (2) bekezdés

A hazai Jog és nyelv kutatások, mint a magyarországi kontextusban leginkább témába vágó kutatási irányzat szakirodalmának rövid előzetes áttekintése (pl. Vinnai 2017b, 2018, Minya – Vinnai 2018) a következő kutatási kérdésekhez vezetett:

- I. A szaknyelvek szoros elvárásrendszerben működnek, a szakma művelői joggal követelik meg tőlük a pontosságot, és a szakmai korrektséget. A közérthetőség ezzel látszólag ellentétesen hat; a szöveget a befogadó (laikus) nézőpontjához igazítja. Hogyan egyeztethető össze ez a két szempont, ráadásul az utóbbi prioritizálása mellett a jogi doménben?
- II. Ebből következőleg mi a jogtudomány álláspontja egy ilyen intralingvális átfordításról hazai és nemzetközi kontextusban?
- III. Melyek a jogi domén azon rétegei, amelyekben a közérthető fogalmazás primátusa legitimálható és megvalósítható?
- IV. Mely preferált és diszpreferált nyelvi jellemzők jelennek meg a közérthetőséget tárgyaló hazai és nemzetközi szakirodalomban?

A rendelkezésre álló (főleg a Plain Language Movement külföldi eredményeihez köthető) irodalom áttekintését követően világossá vált, hogy az egyes hivatalok tájékoztató anyagai jelenthetik azt a médiumot, amely a jogalkotó (illetve az állami szervezet apparátus) legszélesebb érintkezési felületét jelentik a laikus befogadóval.

A Nemzeti Adó- és Vámhivatal Közérthetőségi Munkacsoportjában dolgozó nyelvész szakértőkkel történt kapcsolatfelvételt követően rendelkezésre állt olyan adatbázis, amely a gyakorlati munka során közérthetőségi szempontokat figyelembe véve átfogalmazott szövegeket tartalmazott. Mivel az adatbázis mérete elégségesnek tűnt gépi tanulási technikák alkalmazására, a következő hipotézisek igazolása vált szükségessé:

- iii. Megfelelő tanítóadatok birtokában lehetséges gépi tanult modellel közérthető, és átalakításra szoruló szövegek automatikus szétválogatása (klasszifikáció).
- iv. Egy ilyen modell köré lehetséges olyan szoftvert építeni, amely a szöveg fogalmazóját (egyfajta speciális helyesírás-ellenőrzőként) közérthetőségi javaslatokkal tudja támogatni.

A probléma megoldását az alábbi kérdések megválaszolásának igénye motiválta:

- V. Az egyes (felügyelt) gépi tanulási algoritmusok közül melyik működik kellően megbízhatóan, hogy a szakértők munkáját érdemben támogathassa?

- VI. Milyen tervezési elvek mentén implementálható egy közérthetőségi ellenőrző / javaslattevő szoftver?
- VII. Ha a szakirodalomban léteznek az egyes nyelvi szintekhez kötődő, konkrét javaslatok a közérthetőségre vonatkozóan, akkor ezek közül melyeket, és hogyan lehet algoritmizálni?

## 1.2 Az értekezés szerkezete

A dolgozat felépítése a kutatás folyamatát tükrözi, tekintetbe véve a témaválasztás három tudományterületen (nyelvtudomány, jogtudomány, informatika) átívelő jellegét is. Az egyes kutatási kérdéseket a hozzájuk kapcsolódó módszerek bemutatásával egyidejűleg válaszolom meg a kidolgozás során. Értekezésem kezdeti szakaszában elsődlegesen a téma kapcsán felmerült jogtudományi kérdésekkel foglalkozom behatóbban. Ezt követően sorra veszem, hogy az egyes nyelvtudományi irányzatok hogyan vélekednek a közérthető kommunikáció jellemzőiről, illetve bemutatom, hogy a Plain Language Movement, mint a témát érintő legbefolyásosabb, de nem tisztán tudományos irányzat hogyan hatott a hivatali kommunikációt meghatározó (akár jogalkotói) megfontolásokra a létrejötté óta eltelt több, mint fél évszázad során. A kapcsolódó kritikai észrevételek összegzését követően mind fejlettségben, mind erőforrásigényben eltérő gépi tanuló algoritmusok alkalmazása és kiértékelése kapcsán szerzett tapasztalatokat ismeretek, majd pedig kitérek azon kézi szabályrendszer bemutatására, amely az elkészített alkalmazásban végül helyt kapott.

Az értekezés a következő főbb egységekre tagolható. A második fejezet ismerteti a jogtudomány szempontjából a jogi szövegek újrafogalmazása kapcsán felmerülő nehézségeket, kiemelt hangsúlyt fektetve a közérthetőségi törekvések alkalmazási területének meghatározására. A fejezet egyebek mellett kitér a normavilágosság kritériumára, a jogi normák címzettjeinek körét érintő eltérő álláspontok gyakran összeütközésbe kerülő természetére, valamint a jogi nyelv szaknyelvi jellegéből adódó feltétlen restrikciónkra (I., II. kutatási kérdések).

A harmadik fejezet az egyes nyelvtudományi irányzatok és vizsgálati módszerek fontosabb hozzájárulását tartalmazza a dolgozat témájához. A pszicholingvisztika kapcsán az értelmezési folyamatok szerepe, valamint a kapcsolódó tárgyalótermi kutatások eredményei kerülnek ismertetésre. A probléma történeti dimenzióját főként a tengerentúli korabeli megközelítési módok bemutatása, valamint a Plain Language Movement kezdeteinek leírása



jelenti. Utóbbi kapcsán fontos szerep jut annak az együttműködési folyamatnak, amely a civil szféra képviselői, a nyelvtudomány, valamint a jogtudomány képviselői között létrejött és lehetővé tette például a közérthető megfogalmazás 2010-es törvényi erőre emelkedését az Amerikai Egyesült Államokban. A korpusznyelvészet kapcsán szó esik a Miskolc Jogi Korpusz létrejöttéhez vezető évtizedes hazai kutatásról, annak eredményeiről és továbbéléséről. A negyedik fejezet ehhez kapcsolódóan a kutatás nemzetközi kontextusának aktualitásairól ad számot, feltérképezve az USA-beli és az Európai Unió országaiban végbemenet változásokat, amelyek a közérthető hivatali kommunikáció megvalósulásának irányába mindeddig hatottak (IV. kutatási kérdés).

Az ötödik fejezet a dolgozat során korábban ismertetett elvek és irányzatok kapcsán leggyakrabban felmerült kritikai észrevételek összegzésére hivatott, valamint áttekinti a közérthető fogalmazás propagálása kapcsán eddig elért eredményeket, illetve azok alkalmazhatóságát a hazai kontextusban (III. kutatási kérdés).

A hatodik fejezet a gépi tanítási fázis bemeneteként szükséges korpusz háttérét, összeállítási munkálatait, valamint az elkészült korpusz legfontosabb jellemzőit ismerteti a szövegek forrására, előfeldolgozására, tipikus felépítésére, valamint a kialakított adatbázis fontosabb leíró statisztikai jellemzőire összpontosítva.

A hetedikől a kilencedik fejezetig a tesztelt gépi tanulási megoldások bemutatása történik meg. Az alkalmazott algoritmusok általános ismertetésén túl minden fejezetben helyt kap az elért eredmények kvantitatív és kvalitatív értékelése is (V. kutatási kérdés). Ehhez szorosan kapcsolódva a tizedik fejezet összefoglalóan áttekinti a gépi tanulási modellek építése során szerzett tapasztalatokat, továbbá ismerteti a folyamat során szerzett főbb tapasztalatokat. Itt kerül röviden tárgyalásra a közérthetőség problémájának néhány pragmatikai vonatkozása, különös tekintettel a nagy nyelvi modellek jelenleg ismert korlátaira a pragmatikai következtetések reprezentálása terén.

Tekintettel arra, hogy a szakirodalomban mindeddig (a szerző legjobb tudomása szerint) hasonló gépi tanítási kísérletre nem volt példa, az eredmények kontextusba helyezését a tizenegyedik fejezet jogász szakértők bevonásával készült kézi validálási folyamat keretében végzi el.

A tizenkettedik fejezet részletesen bemutatja a dolgozat első öt fejezetének ismeretében felállított szabályrendszert, amely a gépi tanult modellel előszűrt szövegegységek esetében képes a felhasználó számára konkrét javaslatokkal támogatni a céldokumentumok

fogalmazási munkálatait. Az egyes alfejezetek témacsoportok szerint rendezve taglalják a szakirodalomban felmerült lehetséges javaslatokat, illetve rövid összegzést adnak azok megvalósíthatóságáról konkrétan a magyar nyelv esetében (VI., VII. kutatási kérdések).

Az értekezés érdemi zárásaként az elkészített konkrét alkalmazás fontosabb implementációs részleteinek bemutatására, valamint az alkalmazás használatának ismertetésére kerül sor, végül pedig az egyes kutatási kérdések és hipotézisek tekintetében összegzem a disszertáció eredményeit.

Összefoglalóan tehát az értekezés a természetesnyelv-feldolgozás<sup>8</sup> eszköztárának egy absztrakt nyelvészeti (szemantikai-pragmatikai) problémára történő alkalmazását kívánja implementálni és verifikálni. Ehhez kvantitatív adatelemzési és kiértékelési technikákat éppúgy alkalmaz, mint az egyes érintett tudományterületek megállapításainak kvalitatív szempontú szintézisét. Mindezzel a nemzetközi szintéren évtizedek óta zajló tudományos diskurzushoz kíván hozzájárulni interdiszciplináris nézőpontból, az eredmények praktikus alkalmazhatóságát mindvégig szem előtt tartva.

### 1.3 Adatok elérhetősége

Az értekezésben bemutatott fontosabb erőforrások a reprodukálhatóság érdekében megtalálhatók a következő linkeken:

- A fontosabb Python nyelven írt kódok, valamint a működő alkalmazás elérhető github repozitóriumba szervezve a következő címen: <https://github.com/uveges/Comprehensibility>
- Az alkalmazás által jelenleg használt huBERT alapú modell (huBERTPlain) letölthető a HuggingFace model hub-ról: <https://huggingface.co/uvegesistvan/huBERTPlain>
- Az értekezéshez készített korpusz nyilvánossá tételéhez szükséges egyeztetés a kézirat lezárta kor még folyamatban van.

---

<sup>8</sup> A kifejezésben szereplő természetes nyelv a mesterséges nyelvekkel történő szembeállításra utal. Természetes nyelven olyan nyelveket értünk, amelyeket emberi közösségek használnak, azok szabályszerűségei a történelem során spontán, vagy tudatos folyamatok (pl. nyelvtervezés) által változnak. A mesterséges nyelvek (pl. az eszperantó, de ide tartozik valamennyi programozási nyelv is) ezzel szemben tudatos tervezés eredményeképpen jöttek létre.

## 2. A pontos megfogalmazás kritériuma

Érdekes adat, hogy a 2022-ben rendelkezésre álló adatok alapján Tóth (2022) a központi jogszabályok számát mintegy tízezerre becsüli. Elemzésében főként arra keresi a választ, hogy a közérthetőség követelménye miként szerepel a hazai országos normákban (azaz ide nem értve például az önkormányzati normákat). Ahogyan írja, ebben az időpontban mindösszesen 116 olyan jogszabály, 11 alkotmánybírósági határozat és alig 47 közjogi szervezetszabályozó jogi forrás létezett, amelyben a közérthetőségre kötelezés egyáltalán megjelent valamilyen megfogalmazásban. Ez első hallásra jelenőrsnek tűnhet, de utána számolva hamar kiderül, hogy ez a teljes említett joganyagnak alig mintegy 2%-a.

A fentiek mellett Tóth (2022: 25-26) további, a dolgozat szempontjából kritikusan fontos megállapításokat is tesz. Ezek egyike szerint Magyarországon a jogalkotással foglalkozó szereplők számára nem léteznek az egyes jogágakra összpontosító közérthetőségi előírások. Ennél talán még fontosabb az a leírás, mely szerint a „...megtalált közérthetőségi szabályok nyolcvan százaléka a jogalkotó helyett a jogalkalmazóra hárítja a közérthető jogi szövegtípusok (tájékoztató, hirdetmény, ügyfél kitanítása, kötelezettségekre felhívás, határozat) előállítását.” Emellett Tóth álláspontja szerint, ahol elő is fordul a közérthetőség, mint elvárás, ott sem az adott téma alapjogi súlya, sokkal inkább valamilyen uniós, vagy emberjogi kötelezettség áll ennek háttérében.

Mindez összefoglalóan arra utal, hogy a közérthetőség terminusa nem képezi szerves részét a hazai jogszabályoknak, az erre való törekvés intézményes formában mindössze marginálisan van jelen, betartásához pedig sok esetben semmilyen érdemleges szankció nem fűződik.

Ennek ellenére a hozzá legközelebb álló elv, vagyis az érthetőség és a világos, egyértelmű megfogalmazás határozott elvárása kiemelt fogalom valamennyi jogág esetén.

Közérthető megfogalmazás alatt ellenben a legegyszerűbb közelítésben olyan (szöveg – mondat) szerkesztést érthetünk, amelyet mindenki gyorsan és könnyen megért. A két elvárás között felfedezhetők azonosságok és eltérések is, így e fejezetben (a normavilágosság vizsgálatán keresztül) elsődlegesen a következő kérdésekre keresem a választ:

- Hogyan és milyen szemszögből értelmezi a jogalkotó a jogszabályok pontos megfogalmazását?

- Milyen nézőpontból értékeli e követelmény teljesülését?
- Mik az e mögött álló legfontosabb érvek?
- Ezek alapján pedig tekinthető-e a normavilágosság egy közérthetőségi igény elvi alapjának?

## 2.1 Normavilágosság

A jogalkotási törvény (Jat. 2.§ (1) bek.) a következők szerint fogalmaz: „a jogszabálynak a címzettek számára egyértelműen értelmezhető szabályozási tartalommal kell rendelkeznie”. A fenti kijelentés egyben annak a követelménynek is explicit deklarációja, melyre a jogelmélet normavilágossággént hivatkozik, és amely olyan nagyobb (olykor magas absztrakciós szinten elhelyezkedő és nehezen körülhatárolható) kritériumrendszerek egyik eleme, mint az „alkotmányosság” vagy a „jogállamiság”.

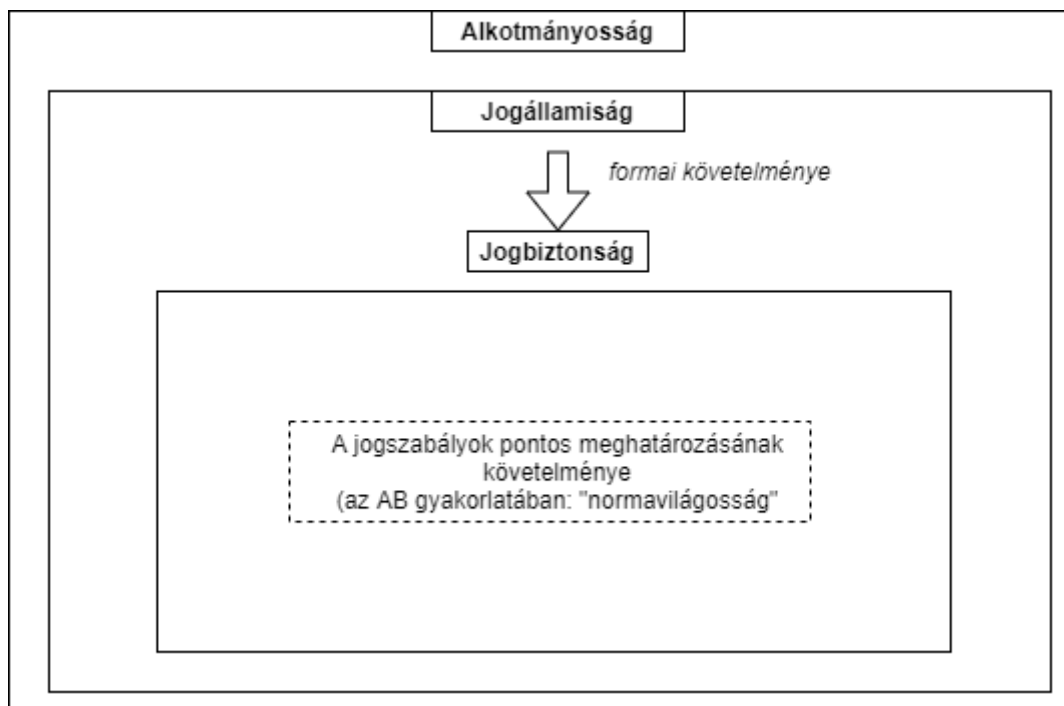
Ezek közül az előbbit olyan elvi követelmények együtteseként lehet meghatározni, amelyek az ideális demokratikus alkotmány tartalmát és megvalósulását jellemzik, a jogállamiság pedig valójában az egyik ilyen követelményként értelmezendő (Papp et al. 2007: 29).

A jogállamiság önmaga szintén egy olyan ernyőfogalom, amely több, azt definiáló kritérium együtt jelenléteként építhető csak fel egyértelműen. Az angolszász büntetőjog kapcsán Ficsor (2018: 40-41) az ilyen elvek ismérveként általánosan a jogbiztonság, kiszámíthatóság előmozdítását tekinti, és konkrétumként nevezi meg például a visszaható hatályú jogalkotás tilalmát (azaz, hogy például a bírák csak olyan jogszabályt alkalmazhatnak, amely az elkövetés idején hatályban volt) és (Horder 2016: 66 nyomán) a normák pontos megfogalmazásának követelményét (principle of maximum certainty). Ezzel kapcsolatban kifejti, hogy utóbbi „azt várja el a jogalkotótól, hogy a jogszabályok szövegét oly módon fogalmazza meg, hogy a laikus állampolgárok számára is világos, érthető legyen, milyen magatartásokat kívánnak meg tőlük” (Ficsor 2018: 41).

A jogállamiság és a jogbiztonság elve, habár előbbi pontos meghatározására egységesen elfogadott és koherens kritériumrendszer mindmáig nem született, mégis az egész jogrendszerre ható alapelveként (esetenként „eszményként”) nevezhető meg, amely áthatja valamennyi jogágat (büntetőjog, alkotmányjog, polgári jog stb.). A jogszabályok pontos meghatározottságának követelménye voltaképpen a jogbiztonság egyik megvalósulási formája – követelménye.

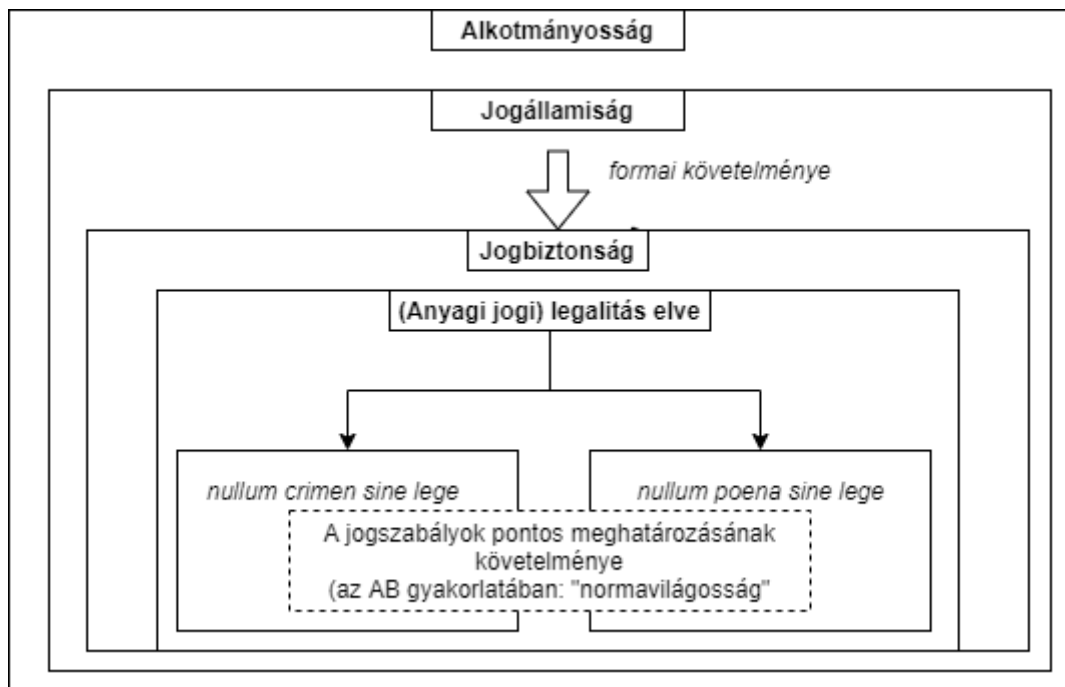
A büntetőjogi jogág esetében a pontos meghatározottság ennél egy fokkal még jobban tagolható elvek láncolatának végén helyezkedik el. A kapcsolódó meghatározások teljes hálójának felrajzolásához ez esetben fontos még megemlíteni a(z anyagi jogi) legalitás elvét, amely szintén hagyományosan a jogállamiság alappilléreként jelenik meg – konkrétan a jogállamiság formai követelményének, vagyis a jogbiztonságnak a megtestesülése (Nagy 2014: 69) – a magyar (büntető) jogrendben pedig speciálisan a „*nullum crimen sine lege* és a *nulla poena sine lege* maximák kompozíciója”-ként áll elő (Karsai 2015: 42). Ezen az egybekapcsolódó jogelvek<sup>9</sup> közül az első kimondja, hogy „csak az a cselekmény minősülhet bűncselekménynek, amit a törvény már az elkövetése előtt annak nyilvánít”, a második pedig deklarálja, hogy „az elkövetővel szemben csak olyan büntetés (szankció) alkalmazható, amit törvény ír elő és amely már az elkövetés idején ismert” (Nagy 2014: 69).

Fontos tehát leszögezni, hogy míg a jogbiztonság elve független a vizsgált jogágtól, azaz „mindenhol jelen van”, ezzel szemben a legalitás elve speciálisan a büntetőjogi jogág sajátja, az abban kialakult zárt logikai rendszer része (Nagy 2014: 65 – 69).



1. ábra: A normavilágosság (~ a jogszabályok pontos meghatározottságának követelménye) kapcsolata a bennfoglaló fogalmakkal a magyar jogrendben **általános**an.

<sup>9</sup> Magyarország Alaptörvényének XXVIII. cikk (4) bekezdése a következő megfogalmazás szerint tartalmazza a két jogelvet: „Senki nem nyilvánítható bűnösnek, és nem sújtható büntetéssel olyan cselekmény miatt, amely az elkövetés idején a magyar jog vagy – nemzetközi szerződés, illetve az Európai Unió jogi aktusa által meghatározott körben – más állam joga szerint nem volt bűncselekmény.”



2. ábra: A normavilágosság (~ a jogszabályok pontos meghatározottságának követelménye) kapcsolata a bennfoglaló fogalmakkal a **büntetőjogi** jogágban.

Nagy (1995) a fenti két jogelvnek<sup>10</sup> négy részét különbözteti meg, avagy négy alágra osztja fel, amelyek konkrétan:

- a visszaható hatályú jogalkotás tilalma,
- az írott jog elsődlegessége,
- az analógia tilalma,
- valamint a jogszabályok pontos meghatározásának követelménye.

A jelen dolgozat szempontjából ez utóbbi tekinthető kulcsfontosságúnak abból a szempontból, hogy a közérthetőség igénye / elvárásamiképpen ágyazódik be a jogtudományba. A tárgyalt fontosabb fogalmak viszonyrendszerét az 1. ábra szemlélteti a jog egészére vonatkozóan, a 2. ábra pedig speciálisan a büntetőjogi jogágra vonatkozó kiegészítésekkel.

A gyakorlatban a jogi normák pontos meghatározásának követelménye az Alkotmánybíróság (a továbbiakban: AB) határozataiban úgy jelenik meg, mint a

<sup>10</sup> *nullum crimen sine lege*, valamint *nulla poena sine lege*

„normavilágosság elve”, és az AB gyakorlatában az az elvárás értendő alatta, mely szerint „a jogszabályoknak érthetőeknek és világosaknak kell lenniük” (Ficsor 2018: 44).

A fogalom egy meghatározását Trócsányi – Schanda (2014: 33) közli, mely szerint: „A formális jogállamiság egyik alappillére a jogbiztonság. Az Alkotmánybíróság szerint a jogbiztonság megköveteli azt, hogy a jog egésze, egyes részterületei és az egyes jogszabályok világosak, egyértelműek, működésüket tekintve kiszámíthatóak és előreláthatók legyenek a norma címzettjei számára (*normavilágosság elve*).”

Fontos tehát kiemelni, hogy habár a normavilágosság mint konkrét terminus jellemzően csak büntetőjogi kontextusban használatos, azonban mindazon elvek, amelyek érvényre jutását biztosítani hivatott, egységesen kulcsfontosságúak a jog valamennyi területén. Álláspontom szerint a jogtudományban használatos terminusok közül a normavilágosság kritériuma az, amely a közérthetőség kritériumával leginkább rokonítható. A két fogalom lehetséges viszonyának meghatározásához további kérdések megválaszolásán keresztül juthatunk el.

## **2.2 Kinek szól a norma?**

Ha a normavilágosság szempontjából az AB gyakorlatát tekintjük, akkor érdemes felidézni két korai határozatot, amelyek indoklásában a testület utal annak követelményére. Az egyik ilyen a 9/1992. (I. 30.) AB határozat, melynek 3. pontjában a következő olvasható: „A jogállam nélkülözhetetlen eleme a jogbiztonság. A jogbiztonság az állam – s elsősorban a jogalkotó – kötelességévé teszi annak biztosítását, hogy a jog egésze, egyes részterületei és az egyes jogszabályok is világosak, egyértelműek, működésüket tekintve kiszámíthatóak és előreláthatóak legyenek a norma címzettjei számára”. A 11/1992. (III. 5.) AB határozat IV. 5. pontjában pedig a következő megfogalmazás áll: „A jogbiztonság megköveteli a jogszabályok olyan világos és egyértelmű fogalmazását, hogy mindenki, akit érint, tisztában lehessen a jogi helyzettel, ahhoz igazíthassa döntését és magatartását, s számolni tudjon a jogkövetkezményekkel” (az idézetben kiemelés a szerzőtől).

Az olyan relatív fogalmak azonban, mint az érthetőség és a világosság kizárólag a befogadó szempontjából értelmezhetők, ha pedig egy közérthetőségi törekvést a normavilágosság fogalmával kívánunk alátámasztani, akkor emiatt a legfontosabb kérdéssé az válik, hogy kit tekintünk a jogi normák (elsődleges) címzettjének? Törvényszerűen más értelmezési helyzetben lesz pl.: egy ítélet, jogszabály, hivatali tájékoztató szöveg stb. olvasásakor egy

jogi végzettséggel rendelkező olvasó és egy, a jogtudományhoz nem értő laikus. Éppen ezért pedig a jogi szövegek általában vett határozottságának kérdése is több dimenzióban értelmezhető. Ez alapján tehát megkülönböztethetünk (minimálisan) két nézőpontot; a jogalkalmazásban részt vevő (és a szaknyelvet elsajátított) szakemberek, és a joggal vonatkozó szakképzettség megléte nélkül találkozó befogadó nézőpontját.

Ficsor (2018: 47) amellett érvel, hogy az AB gyakorlatában határozottan a jogalkalmazó perspektívájának hegemoniája érhető tetten, példaként pedig a 3258/2015. (XII. 22.) AB határozatot hozza. A kérdéses határozatban ([25] pont) az AB a következők szerint fejti ki a normavilágosság szempontjából általa tekintetbe vett befogadó kilétét: „...az alkotmánybíróvási gyakorlat abban is következetes, hogy a sérelmezett jogszabály tartalmának homályossága, ellentmondásossága, bizonytalansága csak abban az esetben alaptörvény-ellenes, ha a normatartalom tisztázatlanságának feloldására a bíróságok jogszabály-értelmezése már nem elegendő”<sup>11</sup>. A gyakorlat alátámasztására számos korábbi határozatot megemlít egészen 1993-ig visszamenőleg<sup>12</sup>, így pedig az AB-nek az általa felvázolt értelmezése a norma elsődleges címzettjére vonatkozóan tartósnak és koherensnek mondható.

A fentiekből az is következik, hogy a két lehetséges nézőpont közül az AB hangsúlyosan a bírák (azaz általánosságban véve a jogalkalmazó) nézőpontját tekinti mérvadónak az állampolgárokéval szemben. Ennek magyarázatához röviden át kell tekintenünk a jogértelmezés módszereit és szerepét a hazai büntetőjog példáján.

### **2.3 A jogértelmezés szerepe**

A túlnyomórészt általános érvényű megfogalmazott és nem pusztán egy-egy egyedi esetre irányadó jogi normák a jogalkalmazás során átesnek egy értelmezési folyamaton, amelynek eredményeképpen az absztrakt normatív tartalmat a jogalkalmazó átemeli, „lefordítja” a konkrét tényállásra alkalmazható formába. Ez a folyamat képezi a tulajdonképpeni bírói döntéshozatal alapját.

---

<sup>11</sup> A példában kiemelés a szerzőtől.

<sup>12</sup> 1263/B/1993. AB határozat ABH 1994, 672, 673–674.; 534/E/2001. AB határozat, ABH 2002, 1283, 1288–1289.; megerősítve: 4/2013. (II. 21.) AB határozat, Indokolás [59].



### 2.3.1 A jogi norma természete

Szilágyi (2006: 253) a jogi normát a jog legkisebb, még önmagában értelmes egységeként határozza meg, amelyre nézve érvényes, hogy önmagában egy (1) teljes, (2) értelmezhető, (3) követhető és (4) alkalmazható magatartásszabályt alkot. Ezen felül kifejti, hogy egy norma nem feltétlenül csak egyetlen jogszabályhelyen kerülhet megfogalmazásra, éppen ezért nem a nyelvi megfogalmazás egységeként, hanem gondolati egységként hivatkozik rá. A jogi normák (csakúgy, mint általánosan a társadalmi normák) lényegeként egy kettős tagolódást említ, amely szerint azok tartalmazznak egyrészt egy követendő magatartásra vonatkozó szabályt, másrészt pedig az erre adott (elvárt) közösségi reakciót leíró szabályt.

A nyelvi megfogalmazás kapcsán érdemes kiemelni, hogy a jogtétel (jogi mondat) nem alkot 1:1 kapcsolatot a jogi normával. Példának okáért a már nem hatályos 1989. évi III. törvény a gyülekezési jogról a következők szerint fogalmaz (a példában kiemelés ismét csak a jelen dolgozat szerzőjétől):

- 1) „2. § (1) A gyülekezési jog gyakorlása keretében békés összejövetelek, felvonulások és tüntetések (...) tarthatók, amelyeken a résztvevők véleményüket szabadon kinyilváníthatják.
- (2) A rendezvény résztvevői jogosultak a közösen kialakított álláspontjukat az érdekeltek tudomására hozni.
- (3) A gyülekezési jog gyakorlása nem valósíthat meg bűncselekmény elkövetésére való felhívást, valamint nem járhat mások jogainak és szabadságának sérelmével.”

Az 1) példában valójában 3 jogtétel ad az értelmezés során egyetlen normát, amely körülbelül így hangozhat: a gyülekezésen részt vevőknek joga van véleményüket kinyilvánítani, de ennek keretében mások jogait korlátozni, illetve másokat bűncselekmény elkövetésére felhívni nem lehet.

Tovább árnyalja a képet, hogy a jogi normát szerkezeti elemeit tekintve (klasszikus értelmezésben) három főbb részre szokás osztani, amelyek:

- a hipotézis: a magatartás körülményeinek, szituációjának a megfogalmazása, annak leírása, hogy milyen körülmények között kell vagy szabad egy adott magatartást tanúsítani,

- a diszpozíció: a hipotézisben megfogalmazott feltételek beállta esetén a jogi norma címzettjének milyen magatartást kell / lehet / szabad tanúsítania<sup>13</sup>), valamint
- a jogkövetkezmény: a diszpozícióban leírt magatartás meg nem valósulása esetén várható (gyakran negatív) következmények, szankciók, azaz a joghatás taglalása (Szilágyi 2006: 253-260).

A fenti elemek szintén szerepelhetnek azonos mondaton belül, vagy több mondatra elosztva, valamint a hipotézis és a diszpozíció különválasztásának szükségességét esetenként vitatják is (pl. Jakab 2001). Olyan eset is előfordulhat, amikor valamelyik (akár több) szerkezeti elem csak implicit formában van jelen szövegszinten, pl.:

- 2) „Aki mást megöl, az büntetett követ el, és öt évtől tizenöt évig terjedő szabadságvesztéssel büntetendő” (2012. évi C. törvény a Büntető Törvénykönyvről - Btk. 160. § (1))

Ebben az esetben a norma diszpozíciója nagyjából a „tilos mást megölni” mondattal foglalható össze, ez azonban ebben a formában nem jelenik meg a szövegben, mindösszesen a negatív jogkövetkezményből vezethető vissza a szó köznapi és nem formális logikai értelmében (vö. Jakab 2010).

A fenti példák csupán néhány, jogi szaktudás nélkül is illusztrálható jelenséget mutattak be, amelyeket a jogi normák szövegszerű megjelenése kapcsán feltétlenül szükséges tekintetbe venni. Ezen felül azonban a legtöbb esetben számos további általános elv és szabály figyelembevételére is szükség van. Mivel a jelen dolgozatnak nem tárgya a szorosabban vett jogértelmezés tételes ismertetése, így a továbbiakban csak egy szempontrendszer aleteit mutatom be annak érdekében, hogy rávilágítsak a normavilágosságot kizárólag a bírák perspektívájából vizsgáló AB gyakorlat hátterére.

---

<sup>13</sup> Ez utóbbinak is köszönhetően (valamint a jog általában vett normatív – előíró jellege és nyelvhasználata miatt) a jogi szakszövegek gyakran képezik olyan kutatások tárgyát is, amelyek a logikából ismert, és a deontikus modalitáshoz szorosan kapcsolódó kötelezőség / megengedethez / tilalom hármásának szövegszerű megjelenéseit vizsgálják, és (többek között) arra a kérdésre keresik a választ, hogy a formális logikai szabályok mennyire transzparenssé képződnek le a jogi domén szövegeiben, egyáltalán megfeleltethetők-e egymással a jogi érvényesség és a logikai érvényesség fogalmai? Ennek eszközeként egyebek mellett készültek kutatások a logikai konnektívumok szövegszintű megjelenési formáiról és a deontikus logika szerepéről a jogi doménben (pl. Markovics: 2017, 2018) vagy éppen a deontikus modalitást kifejező természetesnyelvi eszközökről (Kaczmarek: 2010).

### 2.3.2 Jogértelmezési módszerek

Az absztrakt normatív tartalom alkalmazása a konkrét esetre több különféle értelmezési stratégia alkalmazása által válik lehetségessé. Az értelmezés során tekintetbe vehető módszereket a jogfilozófia és jogbölcsélet berkein belül hagyományosan a következők szerint szokás tagolni (Tóth 2016, Szilágyi 2006 alapján):

- Nyelvtani értelmezés<sup>14</sup> (interpretatio grammatica): a jogi norma nyelvi megfogalmazása és a konkrét tényállás közötti kapcsolat vizsgálatára utal, vagyis a normában szereplő szavak, kifejezések, fogalmak, illetve mondatrészek elemzését jelenti nyelvtani szabályok alapján.
- Logikai értelmezés (interpretatio logica): az írott norma fogalmaiból kiindulva vizsgálja ezeknek és a tényállás elemeinek egymáshoz való logikai viszonyát, vagyis ez a módszer a logikai értelmezés a formális logika szabályainak alkalmazását jelenti.
- Rendszertani értelmezés (interpretatio systematica): a jogszabály valódi tartalmára más jogtételekkel, jogintézményekkel, a jogággal, jogrendszerrel fennálló összefüggéseinek vizsgálta révén következtet.
- Történeti értelmezés (interpretatio historica): a jogalkalmazó a történeti értelmezés során azt vizsgálja, hogy az adott norma annak konstruálásakor milyen fennálló probléma megoldására szolgált, vagyis a cél ez esetben a jogalkotó szándékának rekonstruálása.

A fenti felsorolás közel sem teljes, hiszen az egyes módszerek számos további részre bonthatók; a rendszertani értelmezésen belül például megkülönböztetünk joglogikai, analógia útján történő stb. eseteket. A joglogikai értelmezés esetén (hasonlóan a formális logikai rendszerekhez) speciális, csak a jogra jellemző következtetési szabályok alkalmazása történik. Egy ilyen nevesített következtetési szabály például az *Argumentum a contrario* (avagy az ellenkezőből következtetés elve). Erre egy egyszerű példa lehet az az általános

---

<sup>14</sup> Megjegyzendő, hogy a „nyelvtani” értelmezés terminusa vitatott. Blutman (2007: 4) mellett érvel, hogy pontosabb lenne a „nyelvi értelmezés” terminus alkalmazása, hiszen, mint írja: „...a szemantika a nyelv és a nyelv által jelölt jelenség közötti viszonyra (jelentés), a pragmatika a nyelv és a nyelvhasználó közötti viszonyra (a nyelvhasználat körülményei), míg a szintaktika (az általános nyelvészetben: szintaxis) a nyelv egységeinek egymás közötti formális szerveződési szabályaira, és viszonyára (nyelvtan) utal. Egy jogszöveg nyelvi jelentését e három vonatkozás keresztpontjában találhatjuk meg.” Véleménye szerint a „nyelvtani értelmezés” kifejezés mindösszesen a szintaktikai sajátágok tekintetbe vételére utal, ugyanakkor a komplex értelmezési folyamat során vizsgált sajátágok nem korlátozódnak pusztán csak erre az egy szintre, abban legalább ilyen jelentős szerep jut (jogi) szemantikai tartalom, és esetleg a használatának körülményeinek (pragmatikai oldal) értékelése is.

bírói gyakorlat, mely szerint, ha egy jogszabály kifejezetten nem zárja ki a fellebbezés lehetőségét, akkor abból az következik, hogy a bírósági határozat ellen fellebbezésnek van helye.

A jogalkalmazás során nem minden esetben szükséges valamennyi értelmezési stratégia alkalmazása, illetve ezek használatának gyakorisága is eltérhet jogáganként és akár személyenként is. Ficsor (2018: 39) megjegyzi, hogy a büntetőjog esetében például a nyelvtani értelmezés elsődlegességére kell számítanunk, ez azonban nem jelenti automatikusan azt, hogy polgári jog esetében is hasonló lenne a helyzet. A megfelelő értelmezési stratégia kiválasztása sok esetben kellő gyakorlatot igényel, illetve érdemes szem előtt tartani azt a tényt is, hogy az egyes helyzetekben lehetnek már korábban kialakult („bevett”) értelmezési gyakorlatok, stratégiák is, amelyek a joghoz nem értő laikusok számára nem, vagy csak nehezen hozzáférhetők.

Az értelmezést esetenként tovább nehezítheti a jogi fogalmak absztrakt, rugalmas jellege. Az ilyen fogalmak alkalmazására főként azért van szükség, hogy a jog képes legyen alkalmazkodni a változó társadalmi, gazdasági körülményekhez. Ez a fajta rugalmasság azonban nem ellentétes irányban hat a normavilágossággal (amely a pontos meghatározottságot várja el), hanem azt kiegészítve működik.

Mindemellett Minya – Vinnai (2018: 14) megemlíti a jogi szakszókincs négy jellemző osztályát is a következők szerint:

- 1.) jogi jelentésű, de közismert szavak pl.: *öröklés, hagyaték, tanú,*
- 2.) „köznyelvi hangzású”, de valójában a laikusok számára homályos referenciával rendelkező szavak pl. *jogképesség, jogi személy, elbirtoklás,*
- 3.) köznyelvi eredetű, de annál megszorítottabb értelemben használt szavak pl.: *dolog, jószág, test,*
- 4.) valamint a köznyelvvvel diszjunkt halmazt alkotó szakszókincs pl.: *alperes, zálogjog.*

Megjegyzni továbbá, hogy ez a fajta tagolódás sok esetben nem pusztán a laikusokat, de a jogászokat is komoly értelmezési feladat elé állíthatja.

A joghoz nem értők esetében a legkomolyabb nehézséget talán a 3.) és 2.) csoport jelentheti. A 3.) esetében, ha az értelmező nincs tisztában a szó speciális használati esetének jellemzőivel, akkor létrejön ugyan egyfajta „megértés”, de valójában az adott szó kontextusfüggő referenciáját ez csupán elfedi, nem pedig feloldja. A 2.) csoport esetében

inkább egyfajta kvázi-megértés állhat elő, kialakul valamilyen homályos kép az adott kifejezéssel kapcsolatban, de a konkrét (szaknyelvi ismereteket igénylő) definíció nélkül ez mégis elégtelen marad.

### **2.3.3 Ki „jogosult” a jogot értelmezni?**

A jogi norma felépítésének (elméleti szempontú és szövegszerű megjelenés tekintetében vett) összetettsége, valamint az értelmezés folyamatának rövid áttekintése már következtetni enged rá, miért tekinti az AB a jogalkalmazót elsődlegesen és kizárólagosan a norma címzettjének. E tekintetben azonban a legfontosabb maga a „jogértelmezés” gyakorlati megvalósulása.

Valós élethelyzetekben ugyanis az egyes jogszabályok jelentését a jogértelmezés módszereinek alkalmazásával a bírák feladata meghatározni, ebben az esetben pedig nincs jelentősége (például egy büntetőperben), ha a vádlott máshogyan értelmezné az adott szöveget; a bírói értelmezés lesz az abszolút érvényű (Focsor 2018: 47). Egy másik fontos érv lehet, hogy míg a joghoz értők körében az értelmezés egy rendkívül jól definiált folyamat, mely tartalmazza például a fentebb (részben) már tárgyalt stratégiákat, addig a laikusok esetében ilyen konzisztens és koherens rendszerről aligha beszélhetünk. Mindemellett, még ha feltételezzük is ezen technikák ismeretét, akkor is megoldandó probléma maradna a dogmatika, és a jogi terminusok fogalmi hálójára jellemző, gyakran rendkívül bonyolult összefüggések kezelésének szükségessége.

Mindezek alapján, ha a jogalkotó arra törekedne, hogy a norma minden lehetséges címzettjének egyaránt érthetően fogalmazzon, egy pontosan annyira heterogén elvárásmatrixnak kellene megfelelnie, mint amennyire heterogén bármely társadalom bármilyen korban (pl. iskolázottságra, életkorra vagy egyéni nyelvváltozatra tekintettel). Mondhatjuk, hogy ez praktikus megoldhatatlan feladat lenne a szaknyelviség által támasztott követelmények figyelembevételével egyidejűleg. Feltehetőleg éppen emiatt a jogalkotó egy hasonló absztrakcióval él, mint például a nyelvtudomány, amikor az egyes nyelvi készségeket az „ideális nyelvhasználó” -ra tekintettel, hozzá relatívan értelmezi (a formális szemantika esetében például lásd Gyuris et al. 2008: 7) azzal a különbséggel, hogy a jog esetében az ideális (szak)nyelvhasználó egy konkrét társadalmi csoportot, vagyis a joghoz értőket jelenti.

Erre a párhuzamra Gellér (2012: 65-64) részben utal is, meglátása szerint ugyanis egy jogszabályt akkor tekinthetünk egyértelműnek, amennyiben azt:

- egy átlagos intelligenciájú,
- józan gondolkodású
- laikus polgár

is képes megérteni. Ez a meglátás, habár elvi síkon sokkal inkább rokon a jelen dolgozat motivációjával, a gyakorlatban (önmagában legalábbis) mégis fenntarthatatlan a társadalom (előképzettségbeli, szövegértési képességeket figyelembe vevő stb.) széttagoltsága miatt.

Utolsó érvként a jogalkalmazó perspektívájának privilegizálása mellett Ficsor (2018: 48) azt a problémát említi, mely szerint a jogi szövegek határozottsága / határozatlansága nem pusztán nyelvi természetű, hanem kiemelt szerepe van a jelenségben a jog argumentatív karakterének is. Ez voltaképpen arra a körülményre utal, mely szerint jogi környezetben egy állítás bizonyossága nem független a meggyőződéstől. Ehhez Szabó (2010: 83) a következő magyarázatot rendel: „Az az állítás, hogy >>ezt a pert megnyerjük<<, nem fogalmazható meg ugyanolyan bizonyossággal, mint hogy >>ennek a körnek a kerülete 3,7 cm<<. Az előbbi vitatható, az utóbbiban legfeljebb tévedni lehet. Ez azonban csak azt jelenti, hogy a jog(tudomány) nem egzakt tudomány, mert a bizonyosság a jogon belül nem független a meggyőződéstől. Az álláspontot vagy vélekedést megalapozó jogi tudásra bizonyossági következtetést nem, de meggyőző érvelést lehet alapozni.” Az ilyen és ehhez hasonló esetek feloldása, intralingvális átfordítása szintén a jogalkalmazó feladatának áll elő.

## **2.4 Szaknyelviség**

A jogértelmezés kérdése explicit módon elvezet egy további problémához is, amely a jogi nyelv szaknyelvi státuszával kapcsolatos, és amelynek lényege, hogy a szaknyelvek felé általában nem jelentkezik elvárásként a közérthető megfogalmazás. Ennek oka abban keresendő, hogy azok céljuk szerint minden más felett a szakemberek egymás közötti kommunikációjának eszközeiként szolgálnak.

### 2.4.1 Nyelvváltozatok

Bussmann (1998: 1260) nyelvváltozat alatt egy adott nyelv olyan ágát nevezi meg, amely egységes és specifikus nyelvi rendszert képez, és amelyet nyelvészetén kívüli kritériumok szerint elkülönített beszélők csoportja használ.

A nyelvek egyes változatait számos szempont szerint osztályozhatjuk, helyezhetjük koordináta rendszerbe, vagy vethetjük össze. A preskriptív nyelvészeti irány (pl. nyelvtervezés, nyelvművelés) képviselői közül Balázs Géza a belső nyelvváltozatokat (csoportnyelveket) vertikális és horizontális tengelyek mentén helyezte el annak függvényében, hogy azok a társadalom földrajzi és szociológiai szempontú tagolódása szerint miképpen oszthatók fel. Ebben a rendszerben a szaknyelvek az ún. társadalmi nyelvváltozatok között foglalnak helyet az argó (~szleng), az életkori nyelvváltozatok és a hobbinyelvek között (ez utóbbihoz értendők például a „sportnyelvek” is, vö. Sipos 2000).

A deskriptív szemléletben (mivel ennek célja a normateremtés helyett a minél inkább adekvát leírás) sokkal kevésbé az egyes nyelvváltozatok közötti relációk, illetve alá-fölérendeltségi viszonyok a hangsúlyosak; e szemléletben a cél jellemzően az egyes nyelvváltozatok jellegzetességeinek megragadása, a használati esetek, (pragmatikai) kontextus stb. felmérése.

Az alkalmazott nyelvészet berkein belül az egyes nyelvváltozatok, nyelvjárások kutatásával behatóan a szociolingvisztika tudománya foglalkozik. Tárgyát tekintve a szociolingvisztika a nyelvhasználat és a társadalom összefüggéseit empirikus módszerekkel, szisztematikusan vizsgáló tudományterületként írható le (Chambers 2015: 1), ezzel összefüggésben pedig általánosan a szaknyelvek csoportja is az érdeklődési körébe tartozik. A téma kiterjedtségét (egyben pedig a szaknyelvek esetében a szakértői értelmezés / interpretáció szükségességét) jól jelzi, hogy a szaknyelvi regisztereket ezen felül vizsgálja még például a fordítástudomány<sup>15</sup>, vagy speciálisan a jogi szaknyelvet maga a jogtudomány is.

---

<sup>15</sup> Jellemző problémakör a területen például az egyes fogalmak ekvivalenciájának (és egyáltalán az ekvivalencia mibenlétének) kérdése, azaz annak kutatása, hogy a fordítás során előálló szöveg milyen szempontok szerint tekinthető azonosnak az eredetivel. Heltai (2012: 11) például összefoglaló jelleggel denotatív, konnotatív és pragmatikai ekvivalenciát említ. Az előbbieket ebben a sorrendben a szavak lexikai jelentésének azonosságát / hasonlóságát, a kollokációs jelentések hasonlóságát, illetve a kommunikációs szándék egybeesését jelenti. Ez különösen fontos kérdés például olyan esetekben, amikor egy adott jogi fogalom a lefordítandó nyelven létezik, azonban az országban, amelynek célnyelvére a fordítás történik, nincs pontos megfelelője. Tekintettel arra, hogy szemben például a természettudományos terminológiákkal, a jogi fogalmak országokként eltér(het)nek, sok esetben jelentkezik az imént említett jelenség, amit a terminológiai ekvivalensek hiányaként írhatunk le. Ilyen esetben a sikeres fordításhoz elengedhetetlen lehet akár a kapcsolódó jogi hagyományok összeegyeztetése is, ahogyan arra Hettinger (2020) is rámutat. Egy ilyen példa

Az eddigieken felül ismert még olyan osztályozás is, amelyben a nyelvi jelenségek tekintetében azonos változatot használó beszélők nyelvváltozatai három főbb szempont alapján különülnek el, ezek; az etnikai, területi és társadalmi nyelvváltozatok. Utóbbi esetében használatos a szociolektus terminus is, valamint érdekesség, hogy ebben a felosztásban a szaknyelvek is társadalmi nyelvváltozatnak minősülnek (Kálmán – Trón 2007: 34).

#### **2.4.2 A szaknyelviség által támasztott kritériumok**

Szabó (1997: 170) a szaknyelvek jellemzőjeként a köznyelvtől eltérő terminológiát említi, kiemelve, hogy azok grammatikai sajátosságaikban mindenben megegyeznek az előbbivel. A jogi szaknyelv azonban e tekintetben is speciálisnak nevezhető, tekintettel arra, hogy a nyilvánvaló terminológiai különbségek mellett gyakran merül fel vele kapcsolatban a köznyelvnél jelentősen hosszabb és összetettebb mondatok, a személytelen szerkezetek, vagy akár a (túlzott) nominalizáció (igétlen mondatok) problematikája, amely tényezők együttesen tovább nehezíthetik a jogi szakértelemmel nem rendelkező laikusok szövegértését (Vincze 2018: 9).

Vinnai (2011: 98-100) a jogi szaknyelv és a köznyelv összehasonlítása kapcsán a különbségek ismertetését a morfoszintaktikai jegyekkel kezdi, kiemelve, hogy e tekintetben a köznyelvhez képest a jogi szaknyelvben:

- gyakoribbak például az „-ás”, „-és”, „-ság”, „-ség” képzővel ellátott (igéből képzett) főnevek,
- lexikai / frazeológiai szinten jelen vannak:
  - köznyelvi szavak köznyelvi és speciális jogi értelemben is,
  - tekintettel arra, hogy a jog egyre több más tudományterülettel érintkezésben a mindennapok egyre kiterjedtebb részét szabályozza, így benne megjelenik más tudományterületek szakszókészlete,
  - speciális jogi terminusok,
- szintaktikai szinten: személytelen, E/3. személyű, nominalizált és funkciógés szerkezetek (pl.: „intézkedik” helyett „intézkedést fogatosít”).

---

lehet a magyar Kft. (korlátolt felelősségű társaság) angol megfelelőjének keresése, amely egy az egyhez mérten nem feleltethető meg például egyetlen angolszász jogi (cégformát leíró) fogalomnak sem (Varga 2013).



Az eltérések legfőbb oka alapvetően az a fajta pontosságra törekvés, amely minden szaknyelvet alapjaiban áthat, és amelynek oka a jelentés lehető legpontosabb, leginkább egyértelmű közlésének igénye, amely számos ponton tetten érhető.

Szaknyelvi jellegéből adódóan a jogi nyelv is törekszik a köznyelvben gyakran előforduló többértelműségek, homályosságok kiküszöbölésére valamennyi nyelvi szinten. Éppen ezért sok olyan fordulatot, (a jogi szövegekben) hagyományos fogalmazásmódot és terminust alkalmaz, amelyek speciális szemantikája, alkalmazási szabályai a nem szakmabeli számára elsősre nem nyilvánvalók. A köznyelvben (a kontextus egyértelműsítő hatásának köszönhetően csak esetlegesen zavaró hatású) lexikai szemantikai többértelműségek (pl.: „ég” – ige/főnév), szintaktikai többértelműségek („visító gyerekeket és anyukákat láttam a parkban”) és hatóköri többértelműségek („ebben a teremben mindenki legalább két nyelvet beszél”)<sup>16</sup> is gyakran előfordulnak. A helyzetet tovább nehezíti, hogy olyan esetek is elképzelhetőek, amikor valamely, a köznyelvben szinonim szóalakok a jogi szaknyelvben egymástól teljesen eltérő jelentést kapnak, például a „kártérítés” és „kártalanítás” szópár esetében (Minya – Vinnai: 2018: 15), de előfordulnak például a köznyelvi jelentésnél sokkal határozottabban definiált, ezért pedig sokkal szűkebb jelentésű köznapi szavak is, pl.: „vagyon”, „hagyaték” (Tóth 2016: 49). Ilyen értelemben tehát a szaknyelv több esetben nem új (így a köznyelvi regisztertől könnyen megkülönböztethető) terminust vezet be, hanem meglevő lexikai egységeket ruház fel a pontosság érdekében speciális jelentéstartalommal.

A jogi szaknyelv néhány további tulajdonsága már tárgyából, a jogból vezethető le. Vinnai (2017: 70-71) megemlíti például, hogy a köznapi nyelvhasználattal ellentétben a jogi környezetben alkalmazott fogalmak döntő többségének csak szemantikai értelemben vett denotációja van, jelölete azonban nincs<sup>17</sup>. Egy másik sajátossága a jogi doménnek, hogy alkalmazott fogalmai normatívak, azaz nem leírják a valóság valamely állapotát, hanem

---

<sup>16</sup> A kiemelt mondat többértelműségét például a természetes nyelvi kvantorok (pl. *minden*, illetve *van olyan*, *hogy*) alulspecifikáltsága okozza. A mondat kétféle jelentése közül (mindenki ugyanazt a két nyelvet beszéli, vagy akár mindenki különböző kettőt) a kontextus ismerete nélkül a helyes interpretáció kiválasztása nem is lehetséges minden kétséget kizáróan. Az ehhez hasonló jelenségekkel behatóbban foglalkozik a nyelvtudomány több ága is, ezek közül a formális szemantika a jelenségek matematikai pontosságú magyarázatára is kísérletet tesz. (Bővebben lásd például: Gyuris – Maleczki – Varasdi 2008: 187 – 201.)

<sup>17</sup> A (formális) szemantikai megközelítésében denotáción (más szóval intenzión) a mondat „jelentését” értjük, azaz valamely lexikai egység elvont, a beszéden kívül értelmezhető részét, ennek valós világbeli megjelenése a jelölet (extenzió), pl. a *kutya* szó jelentése mindazon jegyek összessége lehet, amellyel a kutyát mint fajt jellemezni szokás, ellenben a jelölete az a konkrét valós világbeli példány, amelyre a szó használatával utalni szeretnénk. A fenti distinkciót Frege (1892) -től szokás eredeztetni, de később átvette és továbbfejlesztette például a mondatjelentésre (mondatintenzió) vonatkozóan Carnap (1975) is. Hasonló felosztás figyelhető meg a leggyakrabban Ogden – Richards nevével fémjelzett szemantikai háromszög esetén is, amely a *nyelvi jel – jelentés – jelölet* felosztást tekinti irányadónak (Ogden – Richards 1923).

előírnak: egy ideális állapotot vázolnak fel, és ennek megvalósítása lesz a tulajdonképpeni céljuk.

### **2.4.3 Szaknyelviség és közérthetőség kollíziója**

Mindezek alapján talán még nyilvánvalóbb, miért indokolható az a hozzáállás az AB részéről, hogy a jogi szaknyelv értelmezésének szükségességét elsődlegesen a bírák hatáskörébe utalja, és a normák lehetséges címzettjei közül az ő perspektívájukat priorizálja.

Mindazonáltal ez a jogalkalmazásban jelentősen alulreprezentálttá teszi azt a tényt, hogy a norma címzettjeinek többsége (annak ellenére, hogy a normák értelmezése sok esetben kétségkívül szakembert nem nélkülözhető feladat) mégiscsak a laikusok közül kerül ki.

Fontos továbbá kiemelni, hogy a joggal, mint társadalmi alrendszerrel kapcsolatosan él a vélelem, mely szerint a jogszabályokat mindenki ismeri, ezen felül pedig alapelv, hogy a jogismeret hiánya nem mentesít a felelősség alól, néhány speciális eset kivételével (Tóth 2016: 48). Ez azonban mind morálisan, mind pedig gyakorlati szempontból rendkívül aggályossá teszi, hogy a lehetséges címzettek közül, például büntetőjog esetében a nyilvánvaló többség, azaz a jogi eljárások alanyainak nézőpontja látszólag (legalábbis rendszerszinten) egyáltalán nem érvényesül a kodifikációs folyamat során.

### **2.4.4 Normavilágosság és/vagy közérthetőség?**

Összefoglalóan elmondható, hogy a köznapi értelemben vett világos megfogalmazás, illetve a jogszabályok esetén felmerülő határozottság problémája tehát, habár egyaránt az értelmezés egyértelműségét célzó fogalmaknak tekinthetők, mégis nagyon különböző keretrendszerben maradnak csak értelmezhetők. Ugyanezért kapcsolatuk nem írható fel egyetlen relációként sem.

A normavilágosság valójában a jogalkotó válasza arra a kérdésre, hogy minden szükségtelen alulspecifikáltság kiküszöbölése sikeres volt-e egy jogszabály esetében. Ennek hátterét a jogértelmezés során használatos elvek, valamint a jog absztrakt normatív jellegét tárgyaló, és a jogi nyelv speciális nyelvi jellemzőinek hátterét megvilágító fejezetek voltak hivatottak illusztrálni. Közérthetőnek ellenben köznapi értelemben (ahogy arra a fejezet bevezetőjében is történt utalás) akkor tarthatunk egy szöveget, ha azt mindenki gyorsan és

könnyen képes megérteni. A fogalmak megalkotása mögötti szándék tehát párhuzamosnak tekinthető abban az értelemben, hogy céljuk mindenki számára lehetőséget teremteni egy szöveg azonos értelmezéséhez eljutni. Működésükhöz azonban eltérő kontextus szükséges; a normavilágosság a jogalkalmazó, míg a közérthetőség a laikus szempontjából megvalósítandó célkitűzés.

A normavilágosság szempontjából (tekintettel arra, hogy azt a jogalkotó csak a szakmabeliek szempontjából értelmezi) az értelmezés képességének hiánya nem merülhet fel problémaként. Mivel a megfelelő képzést mindenkinek el kellett végeznie, aki ez alapján jogosult a jogszabályok végső értelmezésére, a kompetencia hiányát nem szükséges aktív tényezőként kezelni. Ebből a szempontból tehát a normavilágosság sokkal inkább az értelmező „elmozdulását” várja el a normaszöveg irányába.

Ezzel szemben a közérthetőség, még pontos definíció nélkül is intuitív módon a szöveg változását kívánja meg annak érdekében, hogy a befogadó számára az (akár egy eredeti változathoz képest) előnyöket tudjon biztosítani.

A kettő közti híd megteremtéséhez még fontos tisztázni olyan alapvető fogalmakat, mint például, hogy mit is érthetünk *érthetőségen* vagy szűkebb értelemben véve *közérthetőségen*. Ennek megfelelően a dolgozat következő fejezete főként erre kísérel meg választ adni.

### 3. Megközelítési módok a *közérthetőség* fogalmával kapcsolatban

A közérthető fogalmazás mibenléte kapcsán viszonylag gyorsan szembeötlik, hogy annak létezik egy mára nemzetközileg is elfogadott meghatározása. Az International Plain Language Federation honlapján is olvasható meghatározás magyar fordításban a (szintén a közérthetőség témájával foglalkozó) vilagosbeszed.hu honlapon olvasható. A megfogalmazás szerint; „akkor nevezünk egy szöveget közérthető megfogalmazásúnak, ha a célközönség:

- meg tudja találni, amire szüksége van;
- megérti, amit talált; és
- fel tudja használni az információt a saját igényeinek kielégítésére.”<sup>18</sup>

Az International Plain Language Federation olyan ernyőszervezet, amelyek a nemzetközi közérthető nyelvhasználatot foglalkozó tömörüléseket, szakértői csoportokat foglalja egységben, mint például a Centre for Plain Language<sup>19</sup>, a Clarity-international<sup>20</sup>, amely saját folyóirattal és konferencia sorozattal is rendelkezik a témában, vagy éppen a Plain Language Association International<sup>21</sup>. Munkájuk eredményeképpen 2023. júniusától a közérthető megfogalmazásnak immár a Nemzetközi Szabványügyi Szervezet (ISO) által kibocsátott sztenderdje<sup>22</sup> is létezik, amely ajánlások szintjén tartalmazza a legfontosabb közérthetőségi ajánlásokat. Ennek ismertetése a 3.6 fejezet tárgya.

Amit egyelőre érdemes ebben a rövid meghatározásban észrevenni, az az, hogy a fenti meghatározás erősen pragmatikai központú. A szakértők számára helyes iránymutatást ad arra vonatkozóan, hogy milyen általános elveket kövessenek egy szöveg fogalmazása, szerkesztése során, azonban (legalábbis a jelen formájában) implementációs célokra nem alkalmazható.

Hasonló a helyzet a magyar nyelven, a kozerthetofogalmazas.hu oldalon elérhető segédlettel is, amely mind témacsoportok szerint (a dokumentum szerkesztése, tagolás,

---

<sup>18</sup> Erdetiben: “A communication is in plain language if its wording, structure, and design are so clear that the intended readers can easily find what they need, understand what they find, and use that information.” (<https://www.iplfederation.org/plain-language/> Elérés: 2024. 01. 30.)

<sup>19</sup> <https://centerforplainlanguage.org/> (Elérés: 2024. 01. 30.)

<sup>20</sup> <https://www.clarity-international.net/about/aboutus.html> (Elérés: 2024. 01. 30.)

<sup>21</sup> <https://plainlanguagenetwork.org/> (Elérés: 2024. 01. 30.)

<sup>22</sup> <https://www.iso.org/standard/78907.html> (Elérés: 2024. 01. 30.)

fogalmazásmódja stb.) és nyelvi szintek szerint ("Szavak", „Mondatok”, „Bekezdések”) is támpontot ad a szöveg könnyen érthetővé tétele kapcsán.

Az oldal nagy előnye, hogy maga is közérthetően fogalmazza meg azokat a szövegszerkesztési elveket, amelyeket követni érdemes. Az egyes instrukciók értelmezését példák segítik, azok megértéséhez pedig semmilyen nyelvészeti szaktudás nem szükséges. A funkciógék kerülését például a „Húzz ki mindent, ami nem bír jelentéssel” pont alatt említi az oldal, majd példaként az alábbi hozza: „Előadást tart” Azaz előad. Az ilyet terpeszkedő kifejezésnek hívják. Ilyenkor az eredeti, kifejező igét (előad) elrejtöd egy főnévben, amit aztán kiegészítesz egy semmitmondó igével (tart). Külön súlyosbítja, ha a cselekvő szerkezetből eközben szenvedő lesz (megrendezésre kerül → megrendez).” Az ilyen és hasonló megfogalmazások összeolvasva könnyen követhető ajánlásrendszeré állnak össze, ismét csak az emberi szövegértésre és kreativitásra alapozva.

A fent ismertetett ajánlások, meghatározásoknak ez a kreatív volta amiatt is fontos, hiszen a jelen értekezés tárgya elsődlegesen egy gyakorlatban alkalmazható *szoftver* felépítése.

Ez esetben, ha „kézzel írt” szabályokban gondolkodunk, mint amilyenek például a MS Word helyesírásellenőrző moduljában is szerepelnek, akkor konkrét nyelvi jelenségeket kell azonosítanunk, amelyeket aztán preferáltként, vagy kerülendőként jelölünk meg. Ilyen lehet például a funkciógés szerkezetek esetében egy tételes felsorolás az érintett kifejezésekkel, és a javasolt helyettesítésekkel. Gépi tanulási megoldás esetében pedig konkrét tanítóadatokra van szükségünk, amelyek megmutatják a modell számára, milyen szövegbeli jellegzetességek elfogadottak közérthetőségi szempontból, és melyek nem.

Az eltérés háttere természetesen az, hogy az elfogadott meghatározás embereknek, sok esetben szakértőknek szól, akik képesek azt az adott helyzetnek megfelelően, gyakran saját tapasztalatokra alapozva értelmezni és alkalmazni. A szoftveres rendszerek működése ennél sokkal inkább mechanikus alapokon nyugszik.

Éppen ezért nem kerülhetjük meg azt, hogy a közérthető nyelvhasználattal foglalkozó jogtudományi és nyelvtudományi irányzatok áttekintése során megkeressük azokat a konkrét jellemzőket, amelyek befolyásolják egy szöveg érthetőségét. Ehhez elsőként érdemes tisztázni a tágabb értelemben a megértést befolyásoló tényezőket, majd pedig számba venni azon kutatási irányokat, amelyek így vagy úgy vizsgálták már a kérdéskört.

### 3.1 Pszicholingvisztika

A „megértés” fogalmával összefüggésben, annak mibenlétét elsőként a pszicholingvisztikai perspektívából érdemes vizsgálni. Tekintettel arra, hogy ez az az irányzat (szakág) a nyelvtudományon belül, amely egyszerre igyekszik számolni a nyelvfeldolgozás és produkció pszichológiai és neurobiológiai tényezőivel, valamint dedikált célja a nyelv (szöveg, szavak, beszéd stb.) megértésének vizsgálata, így ideális kiindulási alapot jelenthet általánosan a megértés, konkrétan a megértést nehezítő tényezők vizsgálatához.

#### 3.1.1 A megértés folyamata és tényezői

Pléh – Lukács (2014: 251-252) a megértés fogalmának lehetséges értelmezései kapcsán a folyamat több szintjét különbözteti meg. Nyelvi szinttől függetlenül különbséget tesz például mondatmegértés és értelmezés, azaz a szó szerinti (understanding) és az átfogó értelmezés (comprehension) között. Ahogyan fogalmaz: „Ez a terminológiai többszörösség azt sugallja, hogy a megértésnek van olyan oldala, amely közvetlenebbül a nyelvi anyaggal, és van olyan oldala, amely egész tudásrendszerünkkel, ismeretvilágunkkal és az adott bonyolultabb szöveg teljességének összefüggéseivel, például társas kontextusával is foglalkozik. Vagyis már e köznapi szóhasználat szerint is a megértésnek van egy elemzőbb, analitikusabb, és egy egészelesebb, holisztikusabb szakasza” (Pléh – Lukács 2014: 251).

| <b>Moduláris megközelítések</b>   | <b>Interakciós megközelítések</b>  |
|---|--|
| Jellemzően lineáris feldolgozást feltételeznek (az egyes nyelvi szintek kiértékelése néhány kivételtől eltekintve szigorúan egymás után zajlik).                                | A domináns tényező a megértési összetevők közötti interakció.  |
| Feldolgozás domináns iránya: alulról felfelé (bottom-up), azaz a hangok → szavak → mondatok → szöveg irány érvényesül, a fordított irány csak utólagos vagy marginális szerepű. | Feldolgozás domináns iránya: fentről lefelé (top-down), de hangsúlyosan mindkét irányban zajlik információáramlás. |
| A (kontextuális- / világtudás csak másodlagos szerepet játszik a megértésben, csak a feldolgozás utolsó fázisában lép be az értelmezési folyamatba.                             | A kontextuális ismeretek és a már megértett részek folyamatosan közrejátszanak a további feldolgozásban.           |

1. táblázat: A megértés fogalma kapcsán felmerülő elméletek egy lehetséges szembeállítás.

A megértés kapcsán a pszicholingvisztika esetében középpontban (ahogyan sok más nyelvészeti ágban is) egyértelműen a mondatjelentés áll. Ennek kapcsán Pléh – Lukács (2014: 251-252) a moduláris és az interakciós elméletek közötti fontosabb különbségeket emeli ki az egyes nyelvi szintek feldolgozására vonatkozóan. Ezek fontosabb tulajdonságait az 1. táblázat szemlélteti.

A moduláris felfogás alap gondolata tehát az, hogy a mondatmegértés során a feldolgozás minden lépése önállóan megy végbe, a kommunikáció az egyes modulok között limitált. Ennek értelmében például a szintaktikai feldolgozás során nem számítanak a szemantikai és kontextuális információk. Ezzel szemben az interakciós modellek szerint a feldolgozás valamennyi állomása egyszerre kezdődik meg, ezért az egyes állomások eredményei szimultán képesek befolyásolni egymás kimenetét, éppen ezért tehát az egyes reprezentációs szintek is bármely irányból módosíthatják egymást (Bowers – Davis 2004).

Pléh – Lukács (2014: 251 – 254) a „megértés” folyamatát (eltekintve a fenti elméleti keretek közötti választástól) általánosítva három fő szakaszra osztja fel, amelyek a bemenet, feldolgozás és a kimenet. Ezek közül a bemenet fogalmát nyelvi mozzanatokból álló jegyhalmazként interpretálja. Ilyen jegyek lehetnek például a szegmentális nyelvi információk és szupraszegmentális jegyek (intonációs minta, hangsúlyozás), de hasonlóan ide sorolható írott szövegek esetében az íráskép és a központosítás használata. A feldolgozás szempontjából központi elemként (hasonlóan például a formális szemantikához és a szintaxishoz) a mondat egységét teszi meg, tehát a megértés folyamatának legfontosabb elemeként is a mondat konstrukcióját azonosítja.

A fenti modellben a bemenet legfontosabb jegyeiként:

- a szórendet (pl. *diskurzus-konfigurációs*- vs. *argumentum-konfigurációs nyelvek*)<sup>23</sup>,
- a(z *egy-egy konstituenseken belüli szavakra vetített*) távolságot (pl.: A barna fiú, akit te is ismersz, de: A barna, általad is ismert fiú),
- az alaktant (pl. *ragok jelenléte*),
- a szupraszegmentális jegyek közül az intonációt (pl. *kijelentő / kérdő hangsúly*) és a nyomatékot,
- szemantikai információkat (pl. *szemantikai jegyek; +/- élő stb.*)
- a mondaton belül értelmezett- és a világról szóló ismereteket,

---

<sup>23</sup> Kiemelés: a szerzőtől.

- valamint a mondat igazságát (pl. a három páros szám)

említi meg (Pléh – Lukács 2014: 256), valamint érintőlegesen kitér a szemantikai és pragmatikai tényezők szerepére, kiemelve a mondatjelentést módosító hatásukat.

### 3.1.2 A megértést nehezítő pszicholingvisztikai tényezők

A hazai pszicholingvisztikai szakirodalom a mondatmegértést nehezítő tényezők kapcsán többnyire a mellérendelés – alárendelés szembeállításán keresztül tárgyalja a legfontosabb, megértést befolyásoló tényezőket (Pléh – Lukács 2014: 275-278, Kas – Lukács 2012, Pléh 2013: 46-51 alapján):

- mellérendelések esetében:
  - az időviszonyokat kódoló tagmondatok esetében a feldolgozás könnyebb, ha a tagmondatok sorrendje követi az eseményekét,
  - az „és” kapcsolt mellérendelések feldolgozása egyszerűbb a szembeállításán alapulóknál,
  - ha a tagmondatok predikátumai között van lehetséges ok-okozati összefüggés, azt hajlamosak vagyunk automatikusan odaérteni (pl. az „eltörtem a lábam, és nem mentem vizsgázni” -mondat esetében ahhoz az „eltörtem a lábam *(ezért)* nem mentem vizsgázni” plauzibilis értelmezést rendeljük),
  - egyetlen tagmondatot vizsgálva az annak elején elhelyezkedő szavak lesznek az értelmezés számára legjobban hozzáférhetők.
- alárendelések esetében nehezíti a megértést:
  - ha az ige elválik a hozzá kapcsolódó vonzatoktól,
  - a főmondat megszakítottá válik („a macska, ami nyávog, megharapott” > „A nyávogó macska megharapott”),
  - a főige előtt sok főnévi csoport halmozódik fel,
  - az ismétlődő szerepviszonyt tartalmazó mondatokat, ahol alanyból lesz alany, tárgyból tárgy, sokkal könnyebb megérteni,
  - eltérés a szokásos (SOV – Szubjektum, Objektum, Verbum) szórendtől (pl. „kávét főz Peti” > „Peti kávét főz”).

Pléh (1998) ezek mellett egy másik kategóriarendszerben kiemeli még a következőket is:

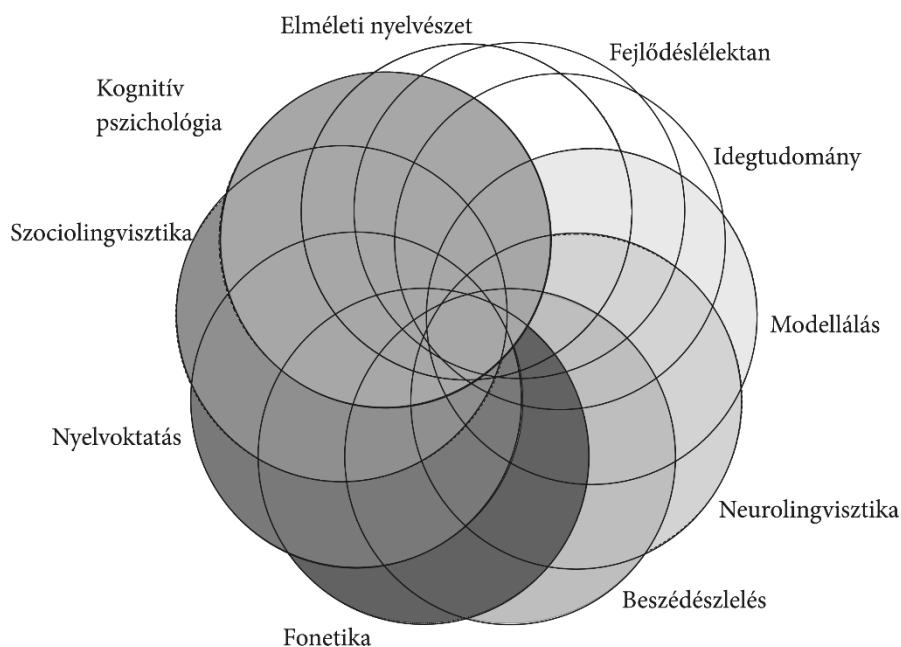


- mondathossz, szóhossz: növekedésükkel a szükséges kognitív befektetés a feldolgozás során növekszik,
- ígétlen mondatok: a mondatbeli viszonyok leképezését nehezíti a központi elem, vagyis az ige hiánya,
- fordított vagy nem kanonikus szórend.

Mindezekon túl Gibson – Pearlmutter (1998) kiemeli még a mondat szintaktikai szerkezete gyakoriságának fontosságát is a megértés kapcsán; általánosan szólva minél szokatlanabb egy szerkezet, a feldolgozása is annál nehezebb lesz.

Ha a közérthetőség kapcsán olyan tényezőket keresünk, amelyek e tekintetben pozitívan hathatnak a szövegekre, akkor a pszicholingvisztikai szakirodalomban található és az előbbieken listázott ellenpéldák jó kiindulási alapot adhatnak a fontosabb faktorok meghatározásához.

A megállapított jellemzőkészlet amiatt is különösen fontos, hiszen a mai értelemben vett pszicholingvisztika jelentős mértékben támaszkodik kísérletes módszerekre, illetve érintkezése egyéb társtudományokkal (3. ábra), például a neurolingvisztikával, valamint annak vizsgálati gyakorlatával is fokozatosan felértékelődik.



3. ábra: A pszicholingvisztika mai értelemben vett, integrált felfogása és kapcsolata a többi szakterülettel (az ábra forrása: Pléh – Lukács 2014: 84, 1.30. ábra)

A vizsgálati módszerek tekintetében ezzel párhuzamosan a már régebb óta alkalmazott mérési technikái mellett, mint amilyen a szemmozgás követés (eye-tracking) vagy a mondatok értelmezéséhez szükséges olvasási idő mérése, jelenleg olyan újabb technikák is szerepet kapnak, mint például a kiváltópotenciál mérés vagy a funkcionális agyi képalkotó módszerek (pl. fMRI, EEG). Az ezekkel nyerhető adatok a korábinál sokkal pontosabb képet festhetnek a nyelv értelmezéséért felelős agyi területekről és a különböző helyzetekben az értelmezéshez szükséges befektetett kognitív erőfeszítésről is (Pléh – Lukács 2014: 82-86, Byrd – Mintz: 2010).

Ilyen módon a megértéssel kapcsolatosan korábban csak áttételesen rendelkezésre álló következtetések igazolása egyre inkább egzakt és empirikus irányba terelődik, ami jelentős alappal szolgálhat minden olyan kutatásnak, amely (akár közvetve) a megértéssel kapcsolatos kérdéseket tesz meg vizsgálata tárgyának.

### **3.1.3 Pszicholingvisztika a tárgyalóteremben**

A pszicholingvisztika eredményeinek, hipotéziseinek és (elsődlegesen kísérleti) módszereinek integrálása a jogi nyelv nyelvészeti kutatásába az 1970-es évekre datálható. Ennek egyik legkorábbi és talán legtöbbet idézett példája Charrow & Charrow (1979) átfogó vizsgálata. A szerzőpáros arra tett kísérletet, hogy a szövegösszefüggések kognitív feldolgozási folyamatai során keletkező nehézségeket azonosítsa és olyan általánosabb lexikai, szintaktikai, szemantikai jelenségeknek feleltesse meg, amelyek ezen feldolgozási nehézségek előfordulását a leginkább magyarázzák és amelyek kiiktatásával a jogi szövegek megértése egyszerűbbé válik.

Vizsgálatuk során esküdteknek szóló tájékoztató szövegeket<sup>24</sup> vetettek alá részletes elemzésnek. Ezek olyan (szándékoltan) joghoz nem értőknek szóló instrukciók, amelyek a bírósági tárgyalás során a döntéshozatalban nyújtanak segítséget, illetve jogi iránymutatást.

A vizsgálat során az instrukciók szövegeit szegmentumokra bontották, azaz olyan grammatikai vagy szemantikai alapon szervezett konstituensekre, amelyek az előzetes

---

<sup>24</sup>BAJI: Book of Approved Jury Instructions in: California Jury Instructions - Civil - Book of Approved Jury Instructions (5th ed. 1969)

tanulmányok szerint egy egységként dolgozhatók fel.<sup>25</sup> A vizsgálat két fő részből tevődött össze, az első fázisban az eredeti szövegezésű instrukciókat alkalmazták. A felkért esküdtek feladata az volt, hogy kétszeri meghallgatás után saját szavaikkal minél több részletet pontosan visszaidézzenek (parafrazáljanak) a hallott szövegrészletből.

Az adatok kiértékelése során a fő kérdések között szerepelt az instrukciók sorrendjének hatása a megértésre, a mondatosság szerepe és a résztvevők demográfiai változóinak esetleges összefüggése az általuk elért eredményekkel.

Az egyik legfontosabb kérdés a felidézett szövegek konceptuális komplexitásához, összetettségéhez kötődött. A parafrázis-tesztek lefolytatása előtt a szerzők jogi szakembereket (ügyvédek, bírókat) kértek meg az adott instrukcióban szereplő jogi/normatív tartalom értékelésére. Az értékelést egy 11 fokú skálán kellett elvégezniük, ahol az 1 a legegyszerűbb, a 11 pedig a leginkább absztrakt, nehezen megérthető jogi koncepciók jelenlétét jelentette. A jogi szakértők előzetesen azt a felhívást kapták, hogy az összetettség értékelésénél minden nyelvtani tényezőt hagyjanak figyelmen kívül és kizárólag a jogi gondolat bonyolultságára legyenek figyelemmel.

Ennek motivációjaként a szerzőpáros azt az akkoriban jogászok körében bevett nézőpontot idézte fel, amely szerint a jogi tematikájú szövegekben nem a mondat szerkesztés okozza a megértési nehézséget, hanem az átadni kívánt jogi fogalom (legal concept), emiatt pedig a szövegek lexikai, szintaktikai egyszerűsítése nem hathat ki a megértés fokára<sup>26</sup>.

Az első eredmények felvételét követően a szöveget újrafogalmazták egy közérthetőbb változatra, ügyelve arra, hogy azok jelentése (tehát a hordozott normatív tartalom) ne módosuljon.

A 2. táblázat a főbb eredményeket szemlélteti. Mivel a kutatásban mindössze 35 esküdt vett részt, így az adatok nem reprezentatívak, arra azonban utalnak, hogy nyelvészeti szempontok szerint végzett (nem normatív tartalmi) átalakítások is előnyösen befolyásolhatják a jogi szövegek megértését.

---

<sup>25</sup> A felbontás során a fő szempont a fellelhető minimális predikátum-argumentum szerkezetek egy szegmentumként való kezelése volt, felsorolások esetében pedig minden egyes tag önálló szegmentumnak számított.

<sup>26</sup> A kísérlet végső eredményei ezzel éppen ellentétes eredményt hoztak; azon instrukciók esetében, amelyeket a jogászok konceptuálisan összetettként értékelték szignifikáns javulás volt tapasztalható az átfogalmazás után. A szerzők ezzel kapcsolatban felvetik, miszerint kérdéses, mennyire választották valójában külön az értékelést végző szakemberek a nyelvi nehezítő faktorokat és a normatív tartalom összetettségét (Charrow & Charrow 1979: 1334 - 35).

| Felvetett probléma  | Megoldási kísérlet  | Eredményezett-e szignifikáns javulást? <sup>27</sup>  |
|---|---|---|
| Lexikai szint   |   |   |
| <ul style="list-style-type: none"> <li>Vonatkozó névmások, kopulák elhagyása</li> <li>Nominalizáció</li> </ul>  | <ul style="list-style-type: none"> <li>Visszaillesztés</li> <li>Igésítés (lehetőség szerint)</li> </ul>   | <p>Nem mért.</p> <p>Igen (0.05)</p>                   |
| Szintaktikai szint  |   |   |
| <ul style="list-style-type: none"> <li>Passzív mondszerkesztés</li> <li>Többszörös tagadások, homályos prepozíciós kifejezések</li> <li>Hátravetett, közbevetett tagmondatok</li> </ul> | <ul style="list-style-type: none"> <li>Cselekvő igék használata (lehetőség szerint)</li> <li>Helyettesítés affirmatív megfogalmazással, megfelelő prepozíciós kifejezéssel</li> <li>Áthelyezés</li> </ul> | <p>Igen (0.01)<sup>28</sup></p> <p>Nem</p> <p>Nem</p> |
| Teljes szövegre vonatkozó általános elvek   |   |   |
| <ul style="list-style-type: none"> <li>Soktagú listák</li> <li>-</li> </ul>   | <ul style="list-style-type: none"> <li>1-2 elemű listákkal való kiváltás</li> <li>Szövegszervezési változtatások a kulcsfogalmak kiemelésével, vagy bővebb kontextus biztosításával</li> </ul>            | <p>Nem</p> <p>-</p>                                   |

2. táblázat: Pszicholingvisztikai szempontú módosítások és mért hatásai (Charrow & Charrow 1979 alapján)

Habár a kontinentális jogrendszerek keretein belüli vizsgálatokra kevesebb példa áll rendelkezésre, azok következtetései nagyban hasonlóak. Neumann (2009) például arra tett kísérletet, hogy német bírósági döntések és az azokat ismertető újságcikkek értehetőségét vesse össze. Az eredmények kiértékeléséhez hasonlóképpen pszicholingvisztikai módszereket (a szövegek feldolgozásához szükséges idő mérése, a megértést ellenőrző kérdések stb.) használt, egyik legfontosabb következtetése szerint pedig a megértés elősegítéséhez nem elengedhetetlenül szükséges a jogi szövegek nagymértékű átalakítása, elegendő lehet néhány kisebb módosítás is, amelyeket nyelvészetileg motivált pontokon végeznek el

<sup>27</sup> Zárójelben az eredeti kísérletben mért szignifikanciaszintek

<sup>28</sup> Alárendelt tagmondatok esetében. A főmondatokat tekintve az eredmények nem voltak szignifikánsak.

### 3.2 Egyszerű lexikai alapú módszerek

Egy másik (történetileg talán az első) megközelítésben, amely kizárólag írott szövegek tulajdonságainak vizsgálatára épít, az olvashatóság (readability, legibility)<sup>29</sup> egy olyan fogalom, amely főként egyszerű jellemzők vizsgálatával hivatott jellemezni egy szöveg érthetőségét, vagyis azt, hogy a szöveg a szerző szándéka szerint legyen képes elérni és befolyásolni a célközönségét (Tekfi 1987: 262).

Habár a témával foglalkozó kutatások egészen az 1920-as évekre vezethetők vissza (Mesmer 2008), az olvashatóság definícióját ebben a keretben hagyományosan mégis a Dale – Chall (1948) által megadott meghatározáshoz szokás viszonyítani. E szerint a legtágabb értelemben az olvashatóság egy adott nyomtatott anyaggal kapcsolatos összes olyan elem (beleértve a kölcsönhatásokat is) összessége, amely befolyásolja az olvasók egy csoportjának sikerességét. A siker definíciója ez esetben azt jelenti, hogy

- mennyire értik meg,
- mennyire olvassák optimális sebességgel, illetve
- mennyire találják érdekesnek

az olvasók az adott szöveget.

Ebben a megközelítésben az olvashatóságot befolyásoló jegyek főleg tipográfiai jellegűek és a szöveg elrendezését érintik, esetleg egyszerű módszerekkel közvetlenül a szövegből származtathatók. Fontos megemlíteni, hogy az olvashatóság (readability) a mai értelemben vett közérthetőség egyfajta előképe, a szöveg e tulajdonságának megítélése azonban a terminust használó szerzők által (mai szemmel) rendkívül mechanikus eljárások alkalmazása mentén történt.

#### 3.2.1 Az érthetőség mérése képletekkel: olvashatósági formulák

Annak érdekében, hogy az olvashatóság fogalmát mérhetővé tegyék, az irányzat képviselői e célra elsőként olvashatósági formulákat, illetve olvashatósági pontszámokat<sup>30</sup> alkalmaztak.

---

<sup>29</sup> Tekfi (1987) értelmezésében a legibility a szövegnek elsődlegesen a nyomdatechnikai eljárások alkalmazása során kialakuló könnyű olvashatóságára utal (választott betűtípus és térközök, a szöveg átlátható elrendezése és tagolása stb.) ezzel szemben a readability a szöveg maga által hordozott jegyeket jelenti (pl. hosszú mondatok, a köznyelvben kevésbé ismert kifejezések és szavak jelenléte ronthatja a szöveg ezen tulajdonság szerinti megítélését).

<sup>30</sup>Eredetileg: readability formulas and readability scores.

Ezeknek az 1920-as évektől kezdve több száz többé-kevésbé különböző válfaja vált ismertté, és alkalmazásuk az 1960-as / 1970-es évekig rendkívül népszerű volt többek között az újságírásban, a kutatásban, az egészségügyben, a jogban, a biztosítási üzletágban és egyes ipari vállalatok körében is. Reneszánszukat az 1950-es években élték, amikor is az amerikai katonaság kidolgozta saját értékelési rendszerét a műszaki kiképzési anyagokhoz, ezáltal pedig talán az első kormányzati szervezet volt a tengerentúlon, amely célzott standardot alkalmazott hivatalos szövegek érthetőségének javítására (Dubay 2004: 2).

Az olvashatósági formulák jellemzően mindössze néhány paramétert vesznek figyelembe és ezek alapján egy képlet szerint számított számértékkel kísérlük meg jellemezni a bemeneti szöveg olvashatóságát. Az egyik legismertebb, a Flesch-Kincaid egyszerű olvashatósági formula<sup>31</sup> például főként a mondatok átlagos szószámból és a szöveg szavainak átlagos (karakterben számított) hosszából állít elő egy 0 és 100 közötti pontszámot. A kor szerzői szerint az így számított pontszám egyenesen arányos a szöveg olvasási nehézségével.<sup>32</sup>

A formulák legnagyobb előnye, hogy könnyen hozzáférhető adatokból állítanak elő egyszerűen értelmezhető mérőszámot, illetve ma már sok esetben egyetlen gombnyomásra elérhetőek online felületeken is<sup>33</sup> az értékelendő szöveg egyszerű beillesztése vagy feltöltése után. Éppen ezért alkalmazói körük nem korlátozódik kizárólag statisztikai vagy nyelvészeti ismeretekkel rendelkezőkre, a szöveggel kapcsolatos visszajelzést bárki szinte azonnal elérheti. Népszerűségüket jól jelzi például, hogy egy implementált olvashatósági mérés mindmáig megtalálható a világszerte egyik legnépszerűbb szövegszerkesztője, a Microsoft Word<sup>34</sup> felületén is, amennyiben a beviteli nyelv angolra van állítva (Farral: 2012).

---

<sup>31</sup>Flesch-Kincaid Reading Easy test:  $206.835 - 1.015 \times (\text{szavak összesített száma}/\text{mondatok száma}) - 84.6 \times (\text{szótagok összesített száma}/\text{szavak összesített száma})$ . A végeredmény minden esetben egy 0 és 100 közötti szám, ahol a konstansok biztosítják, hogy az eredmény ne eshessen ezen a tartományon kívülre, a többi érték pedig a dokumentumtól függ – ezek állítják be az érték nagyságrendjét a 100 fokozatú skálán.

<sup>32</sup> Az olvashatósági tesztek validitását számos kritika érte, sok esetben például túlzott egyszerűségük miatt, pl.: Maxwell (1978), Connaster (1999). Helyes használatuk szintén gyakran vita tárgya. Fry (2004) például határozottan felhívja a figyelmet arra a jelenségre, mely szerint az olvashatósági mutatók limitált paraméter alapján állítanak predikciókat, így ha kizárólag ezek alapján strukturálunk újra egy szöveget, az adott esetben hatástalannak bizonyulhat, vagy a kiindulásnál is komplexebb szövegtestet eredményezhet, miközben a mutató számszerűen javul.

<sup>33</sup> pl.: <https://www.webfx.com/tools/read-able/> (A link ellenőrizve: 2021.07.29.)

<sup>34</sup> Az értékelést a szövegszerkesztő a már említett Flesch-Kincaid egyszerű olvashatósági formula, valamint a Flesch-Kincaid Grade Test alapján végzi, mely utóbbi a szöveg egyszerű megértéséhez szükséges iskolai végzettség szintjét hivatott megbecsülni. A teszt eredménye az USA-beli iskolarendszer keretei között értendő.

### 3.2.2 A megközelítés kritikái

Kritikusaik szerint az olvashatósági formulák mindösszesen néhány felszínes szerkezeti jegyet vesznek tekintetbe, és ebből próbálnak meg következtetéseket levonni azokról az összetett kognitív folyamatokról, amelyek egy szöveg értelmezése közben zajlanak (Anderson – Hart-Davis 2010: 283).

Részben éppen erre reagálva az olvashatósági formulák lassan kiterjesztették a számításba vett nyelvi jellemzőkészletet szintaktikai és lexikális szemantikai jegyekre is.

Klare (1976) több mint harminc, a témában elérhető tanulmányt hasonlított össze, amelyek az olvashatósági formulák hatásait vizsgálták. Konklúzióként megkísérelte azon lexikai és szintaktikai jegyek felsorolását, amelyek véleménye szerint egyes esetekben a pontszámok folyamatos követésével újraírt szövegek érthetőségének javulásáért feleltek. Összegzése alapján a legfontosabb faktorok, amelyek kihatottak a szövegek érthetőségére, a következők voltak:

| Lexikai jegyek   | Szintaktikai jegyek  |
|--|--|
| <ul style="list-style-type: none"><li>• tartalmas / funkciósavak aránya</li><li>• a tartalmas szavak közismertsége, hossza és gyakorisága</li><li>• konkrét / absztrakt vonatkozás</li><li>• asszociatív jelleg</li><li>• cselekvő vagy nominalizált igei konstrukció(k)</li></ul> | <ul style="list-style-type: none"><li>• mondathossz (főként tagmondatoké)</li><li>• aktív / passzív mondatszerkesztés</li><li>• afirmatív / negatív mondat</li><li>• beágyazottság jelenléte</li><li>• beágyazottság alacsony / magas foka (többszörös összetételek)</li></ul> |

3. táblázat: A megértést befolyásoló tényezők Klare (1976: 148) szerint.

A fenti felsorolás élesen elüt az olvashatósági formulák számításánál korábban használt változóktól, főként azokon a pontokon, amelyek már (implicit módon) szemantikai szempontokat is feltételeznek.<sup>35</sup>

<sup>35</sup> Ilyen például a lexikai szinthez sorolt absztrakt vonatkozású, vagy éppen asszociatívként megjelölt lexikális elemek, vagy a szintaktikai szinten a tagadó mondatszerkezetekkel kapcsolatos megfigyelés, mely szerint azok értelmezése nehezebb kognitív feladat (különösen, ha egymásba ágyazott tagadásokról van szó, mint például az „*a gépek meg nem építetlensége*” szerkezet esetében, ahol önálló tagadószó, és szuffixum is kifejez negációt).

### **3.3 Plain Language**

Egy harmadik irányvonal a közérthetőséggel kapcsolatos kérdések vizsgálatára a Plain Language Movement (a továbbiakban: PLM), amely az angolszász országokból indult útjára nagyjából az 1970-es években

A PLM olyan alulról szerveződő irányzat angol nyelvterületen (főként az Egyesült Államokban és Angliában), amelynek fő célkitűzése a kommunikáció hatékonyságának javítása. Ezt főként a hivatali ügyintézés érintő területeken kísérli meg, pl. biztosítási ügylet, törvényhozás, bírósági eljárások, és általánosságban véve minden olyan szaknyelvi közeg (főként, de nem kizárólag állami szervek kommunikációs csatornái) esetén, ahol a szaknyelvi szöveg címzettjei nem kizárólag az adott szakterületen járatos hozzáértők, hanem többségében laikusok (Anderson 1988: 232).

Ebben a keretrendszerben ugyan a közérthető nyelvezet fő célja a kommunikáció hatékonyságának növelése / fenntartása, azonban az irányzat képviselői alapelveként tekintenek arra is, hogy „mindenkinek joga van megérteni azt, ami gyakorlati élete során fontos számára” (Heltai – Nagy 1997: 401). Ennek felfejtéséhez azonban, hogy ebből az egyszerű gondolatból miként lett a tengerentúlon a hivatali kommunikációt szabályzó törvény, egészen a múlt század elejéig kell visszatekinteni.

#### **3.3.1 Történeti előzmények**

A közérthető kommunikáció iránti igény gyökereit az Egyesült Államokban azokban a társadalmi-gazdasági egyenlőtlenségekben találhatjuk meg, amelyek a múlt század első évtizedeiben a fogyasztók és az egyes termékeket előállító nagyvállalatok között fennálltak. Az ilyen egyenlőtlenségekkel szemben fokozatosan egyre inkább intoleránsá váló fogyasztói réteg aktivista mozgalmakat indított, amelyek elsőként a korrekt tájékoztatás és érdekvédelem területén nyilvánultak meg. Nem sokkal később fokozatosan adaptálták azt az elvet is, miszerint az „egyenlőségnek” ez az akkoriban még új dimenziója kiterjeszhető a nyelvhasználatra is. Konkrétan arra a nyelvhasználatra, amely a nagyvállalatok, valamint az állam kommunikációját jellemzi a szélesebb tömegek irányába. Ennek az elmozdulásnak az alapját a fogyasztói mozgalmak megjelenése képezte.



A legáltalánosabb megfogalmazás szerint „a fogyasztói mozgalom olyan gyűjtőfogalom, amelynek referenciáját mindazon tevékenységek adják, amelyek célja a vásárló/fogyasztó elsőbbségének érvényesítése egy gazdaságban” (Finch: 1985: 23)<sup>36</sup>.

Noha az USA-ban a fogyasztói öntudat dokumentált megnyilvánulásait gyakran az 1920-as évekig is szokás visszavezetni<sup>37</sup>, az szélesebb társadalmi bázisra csak az 1930-as években tett szert. A 20-as évek erősen teljesítő gazdasága után a 30-as években elhúzódó Nagy Gazdasági Válság és az azt követő lassú talpra állás általánosan meggyengítette a bizalmat a nagy piaci szereplőkben. Ebben az évtizedben sorra jelentek meg az olyan, akár több százezer példányban elkelt könyvek (pl.: Kallett – Schlink: 1933), amelyek a kor piaci gyakorlatát illették éles kritikával<sup>38</sup>. A növekvő elégedetlenségre válaszul Roosevelttel képviselőket biztosított a fogyasztóknak a New Deal Programokban, ezzel kvázi hivatalosan is elismerve az ilyen kezdeményezések létjogosultságát (Finch: 1985: 27). A következő időszakban sorra alakultak azok az immár hivatalos, nonprofit szervezetek (pl.: Consumers Union, 1936), amelyek dedikált célja információk és tanácsadás nyújtása a vásárlókat érintő kérdésekben, mint például termékbiztonsággal, az egészségügyi ellátással, a pénzügyi szolgáltatásokkal és az élelmiszer-előállításokkal kapcsolatos problémák.

Az üzleti szektor ekkoriban leginkább fenyegetésként tekintett a fogyasztói mozgalomra, mint olyan törekvésre, amely saját érdekeit sérti. Annak ellenére, hogy Roosevelttel egy a kongresszushoz intézett üzenetében (Message to Congress) már 1933-ban sürgette a fogyasztóvédelmi szabályok szigorítását, az első valódi eredmény végül csak 5 évvel később, 1938-ban született meg egy közfelháborodást kiváltó tragédia hatására.

A Sulfanilamide nevű vegyület egy tablettá és por formájában ekkor már régóta forgalomban levő, bakteriális fertőzések kezelésére alkalmazott gyógyszer volt. 1937-ben egy értékesítője piaci igényt jelzett a szerre folyékony oldat formájában is. Ennek megfelelően a gyártó előállította a megfelelő formájú készítményt, azonban a forgalomba hozatal előtt annak

---

<sup>36</sup> A szerző fordítása

<sup>37</sup> Chase – Schlink (1927) például azon marketing gyakorlatok ellen emel szót, amelyek a fogyasztók számára megnehezítették a termékek valós értékének felmérését. Fő célja a vásárlói döntéshozatal felmérése és annak megállapítása volt, hogy egyes termékeknek a valóságban a vásárlók számára betöltött funkciója mennyiben felel meg a gyártó előzetes állításainak. A könyv különös jelentősége főként abban állt, miszerint kellően széles körben vált ismertté ahhoz, hogy a fogyasztói érdekvédelem szükségességét képes legyen beemlíteni a köztudatba.

<sup>38</sup> A könyv legfontosabb állítása az volt, miszerint az akkoriban értékesített (főként élelmiszer- és gyógyszeripari termékek) forgalmazása nemegyszer annak pontos ismerete nélkül történik, hogy a mai értelemben vett adalékanyagok milyen hatással vannak hosszabb távon az azt elfogyasztókra. Emellett olyan éles kijelentéseket is tartalmaz, miszerint a gyártók gyakran tudatában vannak olyan mellékhatásoknak is, amelyeket a profitmaximalizálás érdekében eltitkolnak a szélesebb nyilvánosság elől.

toxicitását semmilyen formában nem vizsgálták (mindössze az elegy ízét, megjelenését és illatát), az ilyen vizsgálatokra ugyanis semmilyen korabeli törvény nem kötelezte a gyártót. A korábban hasznos szerről hamar kiderült, hogy új formájában rendkívül mérgező, mire azonban az erről szóló adatok beérkeztek a gyártóhoz, és az visszahívta a készítményt, valamint elindította a fogyasztók széleskörű tájékoztatását, a szer már közel 100 emberéletet követelt (Balletine 1981).

Az eset nyilvánosságra kerülése után a felháborodás olyan méreteket öltött, amely a kormányzatot is lépéskényszerbe hozta, ezt követően több, a fogyasztói érdekérvényesítő képességet kedvezően érintő törvény is született<sup>39</sup>.

A fogyasztói mozgalmak talán legjelentősebb, következő hulláma már az 1960-as évekre tehető. J.F. Kennedy 1960-as elnöki kampánya során ígéretet tett arra, hogy megválasztása esetén egy fogyasztói tanácsadói irodát fog létrehozni Washingtonban annak biztosítására, hogy a fogyasztók hangját a Kongresszusban is meghallják. Később, 1962-ben a Kongresszushoz intézett, a polgárok igényeiről szóló első beszédében körvonalazta a fogyasztói jogok nyilatkozatát.

A későbbi Plain Language Movement szempontjából legfontosabb kijelentései szerint a fogyasztóknak joguk van a biztonsághoz, a tájékoztatáshoz, a meghallgatáshoz ( ~érdekérvényesítéshez) és a választáshoz (Finch 1985: 31).

Az utolsó, egyben talán legfontosabb eredmények gyűjtőpontjában egy akkor 31 éves ügyvéd, Ralph Nader 1965-ben megjelent könyve (Nader 1965) állt, amelyben a szerző élesen bírálja a gépjárműgyártókat, amiért azok számos esetben a dizájnt és a formatervezést részesítik előnyben az utasok biztonsága helyett. Talán a korban a közúti balesetek során tapasztalható növekvő halálozási arány miatt, de Nader könyve hamar bestseller lett, és nagyban hozzájárult az első komolyabb gépkocsibiztonsággal kapcsolatos rendelkezésének<sup>40</sup> megszületéséhez. A nagy médianyilvánosság is hozzásegítette Nadert, hogy az évtized végére a fogyasztói mozgalmak egyik legnagyobb hatású alakja váljon belőle.

---

<sup>39</sup> pl.: Wheeler-Lea Act of 1938, Wool Products Labelling Act of 1939

<sup>40</sup> National Traffic and Motor Vehicle Safety Act of 1966

A nyilvánosság bevonásával, valamint oktatási és kutatási szervezetek<sup>41</sup> alapításával számos további, a fogyasztókat védő törvény<sup>42</sup> elfogadását segítette elő. Ezek a törvények már nem pusztán reagálást jelentettek néhány elszigetelt eseményre, hanem egyfajta szemléletváltásról adtak tanúbizonyságot, amely során az átlagember érdekei sokkal jelentősebb fókuszot kaptak a korábban egyáltalán nem szabályozott, vagy éppen az eredetileg a nagyipar érdekeit prioritizáló területeken.

### **3.3.2 A Plain Language Movement térnyerése**

Kennedy 1962-es beszéde, majd Nader tevékenysége nyomán reflektorfénybe került az a gondolat, mely szerint az állampolgároknak tudatában kell lenniük, és meg is kell érteniük az őket illető jogokat és kötelességeket. Mindez megadta az alapot, hogy az egyéni érdekérvényesítő képesség kiterjesztése ne csak a fogyasztókat védő törvények elfogadtatásában ölthessen testet, de a hivatalos dokumentumok szövegezésére is hatással lehessen. Ezek fogalmazása során ekkoriban semmilyen értehetőségi kitétel vagy szabályozás nem létezett; ez volt tehát a hiány, amely a PLM megszületéséhez elvezetett.

A fogyasztói mozgalmak előbb-utóbb beszivárogtak olyan területekre is, amelyek nyelvhasználatukban korábban nem foglalkoztak a könnyű olvashatósággal. Ilyenek voltak például a bankok ügyfélszolgálati kommunikációs csatornái, vagy éppen az állami hivatalok tájékoztató anyagai.

A kezdeti időkben a közérthetőségért folytatott aktivista tevékenység individuálisan zajlott, és olyan karizmatikus egyéniségek tevékenysége köré összpontosult, mint Chrissie Maher angol ügyvédnő, aki 1971-től megalapította például az Egyesült Királyság első közösségi újságját, a Tuebrook Bugle-t. Ennek lapjain gyakran adott hangot abbéli véleményének, mely szerint a (kormányzati és magánkézben levő) szervezeteknek közérthető angolt kell használniuk a hozzájuk fordulókkal kapcsolatos kommunikációban. 1979-ben többedmagával megalakította a Plain English Campaign irodát, amely a mai napig elsősorban a közérthető fogalmazás ellenőrzésével és oktatásával foglalkozik. Az ilyen, és

---

<sup>41</sup> pl.: Center for the Study of Responsive Law, Public Interest Research Group, Center for the Study of Responsive Law

<sup>42</sup> pl.: Hazardous Substances Labelling Act of 1960, Kefauver-Harris Amendment (1962), Fair Packaging and Labelling Act of 1966, National Traffic and Motor Vehicle Safety Act of 1966.

ehhez hasonló irodák<sup>43</sup>, szervezetek létrejöttével a PLM tehát fokozatosan egyfajta intézményesülés irányába kezdett elmozdulni.

A tengerentúlon eközben az első olyan intézkedésként, amely során egy magánvállalat elkötelezte magát kommunikációs anyagainak közérthetőbbé tétele mellett, 1975. január 1-ére szokás datálni (Felsenfeld & Sieger 1981, Asprey 2003), amikor az akkori First National City Bank (mai nevén Citybank) saját kezdeményezésére hatályba léptette új, közérthetőségi követelményeket is tekintetbe vevő fogyasztási hitelekéről szóló tájékoztatóját. A korabeli piaci szereplők között ez a lépés megütközést keltett, hiszen teljesen ellentétes volt az addigi bevett gyakorlattal.

Alig néhány évvel később a frissen szerveződő PLM a figyelem középpontjába került, amikor Jimmy Carter 1978-ban elnöki rendeletet adott ki a szövetségi szabályozások világosabbá, kevésbé nehézkesé és költséghatékonyabbá tétele érdekében (Dorney 1988: 49). Ugyanerre az időszakra tehető számos olyan törvény<sup>44</sup> beiktatása, amelyek ugyan név szerint nem hivatkoznak a PLM-re, azonban szellemiségükben már világosan tükrözik annak központi gondolatait (Willerton: 2015 6-7). Később ezt számos szövetségi tagállam is követte helyi rendelkezések beiktatásával.

Az ezt követő időszakban sorra alakultak mind az USA-ban, mind pedig Európa-szerte<sup>45</sup> az olyan szervezetek, amelyek a hivatali kommunikáció világosabbá tételéért küzdöttek, ezzel pedig a PLM immár nem egyéni, hanem társadalmi méretekben folytathatta tovább tevékenységét.

### **3.3.3 Mozgalomból kutatási terület**

A PLM kezdetei kapcsán már felmerült, hogy a hivatalos szövegek egyszerűsítése irányában sok olyan szereplő is komoly erőfeszítéseket tett, akik kapcsolatban voltak a joggal, közöttük több esetben akár jogvégezettek is helyet kaptak. Éppen ezért érdekes néhány fontosabb munkán keresztül végigkövetni, hogy az angolszász jogtudomány művelői közül kik és

---

<sup>43</sup> Center for Plain Language: tagjai főként kormányzati tisztségviselők, akadémikusok és az üzleti szféra képviselői, akik nonprofit szervezetként dolgoznak a PLM előre mozdításáért, Legal Writing Institute: működési területük dedikáltan a jogi nyelv, Burton Foundation: szintén a jogi nyelvvel foglalkozó szervezet, tagjai főként az akadémiai szféra képviselői stb.

<sup>44</sup> Magnuson-Moss Warranty-Federal Trade Commission Act of 1973, Consumer Leasing Act of 1976, Electronic Fund Transfer Act of 1978

<sup>45</sup> pl.: Plain Language Europe

milyen formában igyekeztek hozzájárulni a folyamat előmozdításához akár módszertani javaslatokkal, akár kritikai megjegyzésekkel.

Azon munkák közül, amelyek jogtudományi – jogelméleti szempontból közelítik meg a jogi szövegek nyelvhasználatát, Kimble (2012: 47-54) a következőket említi a legnagyobb hatásúként:

- David Mellinkoff, a kaliforniai jogi egyetem (UCLA) korábban ügyvédként is praktizáló oktatója volt. Az ő nevéhez köthető az első olyan összefoglaló írás (Mellinkoff 1963), amely az angolszász jogi terminológiák eredete, illetve a szakma szociolektusa kialakulásának történetén keresztül, néhány általa jellemzőnek ítélt grammatikai jegyet is tekintetbe véve igyekezett átfogó képet adni a jogi nyelv(változat) természetéről. Fő megállapításai javarészt értéksemlegesen igyekeztek felhívni a figyelmet a jogi dokumentumok gondos megszővegezésének fontosságára<sup>46</sup>.
- Richard C. Wydick szintén a kaliforniai egyetem oktatója volt és már a PLM növekvő népszerűségnek időszakában írta meg azon könyvét (Wydick 1979), amelyben olyan konkrét javaslatokat fogalmaz meg, mint például „felesleges” szavak/kifejezések elhagyása<sup>47</sup>, a (túlzott) nominalizáció kerülése vagy például olyan szerkesztések kerülése, mint az alany, állítmány, tárgy hármásának túlzott eltávolítása egymástól a mondaton belül.
- A michigani ügyvédi kamara (State Bar of Michigan) saját szaklapjában elsőként indított önálló rovatot a PLM-mel kapcsolatos hírek, információk közlésére (Willerton 2015, 7).

---

<sup>46</sup> A szerző munkájában főként az olyan lexikai elemek, anakronizmusok túlhasználata ellen szólal fel, amelyek álláspontja szerint az érdemi tartalomhoz nem járulnak hozzá, ellenben a megértést hátrányosan érintik, vagy éppen félreértésre adhatnak okot (pl.: a *saith*, mint a say E/3 alakú, régies formája jogszabályok szövegeiben, szemben a standardnak számító *said* alakkal).

<sup>47</sup> A nyelv lexémáit két fundamentális csoportra bontja fel, tartalmazó szavakra (working words) és funkciósavakra (glue words). Véleménye szerint ez utóbbiak száma minden esetben kevesebb kell, hogy legyen egy mondaton belül, mint a tartalmazó szavak összesített száma, mivel ezek hordozzák a mondat valódi, közvetíteni kívánt jelentését, a funkciósavak pedig csak a grammatikai jólformáltságért kell, hogy felelősek legyenek. Megjegyzi, hogy a(z angol) jogi szövegekben a funkciósavak arány kiugróan magas, ez pedig a megértést feleslegesen nehezítő tényező. A terminus ugyanakkor nem esik egybe a magyar grammatikában definiált funkciószó kategóriával, mivel a szerző ide sorolja például a tipikusan idiomatikus és szakmaspecifikus állandósult szókapcsolatokat is, mint például:

(1) „It was said during the course of the meeting that...”

Az ilyen esetekben például a kiemelt kifejezés leváltására egyszerű prepozíciót javasol:

(2) „It was said in the meeting that...”

- Már a 90-es években Michèle Asprey ausztrál jogász összefoglaló munkát (Asprey 1991) jelentetett meg, melyben konkrét esetjogi, törvényhozási példákon keresztül mutatja be a közérthető megfogalmazás gyakorlati megvalósításának lehetőségeit számos jogi szakágban.
- 2001-ben Bryan A. Garner gyakorlatokkal színesített könyvet adott ki gyakorló jogászok számára a témában (Garner 2001).

A fenti felsorolás közel sem lenne teljes, ha mindazon írásokat akarnánk számba venni, amelyek az elmúlt közel hatvan évben foglalkoztak a PLM és a jogi szaknyelv integrálhatóságával. Azonban már ennyi példából is jól kivehető, hogy az angol nyelvterületen a jogi szakma képviselői egészen a kezdetektől aktívan részt vettek a jogi szaknyelvnek a PLM támasztotta új irányvonalhoz igazításában. A PLM fontosabb történeti állomásait, valamint az egyes korszakokhoz hozzárendelhető vezérelveket az I. függelék is szemlélteti.

### **3.3.4 Federal Plain Language Guidelines**

Az amerikai kormányzat 2011-ben kiadott egy szövetségi közérthetőségi iránymutatást is Federal Plain Language Guidelines néven (a továbbiakban: FPLG)<sup>48</sup>. Ebben részletesen kifejtik, hogy álláspontjuk szerint milyen elvárásoknak kell megfelelnie egy közérthető szövegnek.

A dokumentum elvitathatatlan előnye, hogy a kapcsolódó nyelvészeti kutatások hivatkozása (pl.: Garner 2001, Murawski 1999: 73-75., Charrow & Erhardt & Charrow 2007 stb.) által direkt kapcsolatot teremt a kormányzati kommunikáció és a tudományos szféra eredményei között. Ilyen formán egyben ez is tekinthető a legpontosabb, konkrét nyelvi példákat is tartalmazó dokumentumnak arra vonatkozóan, hogy a PLM képviselői milyen nyelvi jellemzőket tartanak követendőnek vagy éppen kerülendőnek. Ezeket bővebben a 4. táblázat mutatja be.

A nyelvi különbségek miatt nem minden, az FPLG lapjain tárgyalt eset releváns, vagy akár szemléltethető megfelelő magyar példával. Ugyanakkor a javaslatok többsége (a szükséges átalakítás után) jelentőséggel bírhat a magyar nyelvű jogi szövegek vizsgálata során is.

---

<sup>48</sup> Federal Plain Language Guidelines, US Government, March 2011, online: <https://www.plainlanguage.gov/media/FederalPLGuidelines.pdf> (Elérés: 2021.08.27.)

| Lexikai szint   | Mondatszint   |
|---|---|
| <p><u>Igék</u> esetében:</p> <ul style="list-style-type: none"> <li>• cselekvő igék használata passzív szerkesztés helyett</li> <li>• egyszerűbb jelentésű igék használata<sup>49</sup></li> <li>• nominalizáció kerülése</li> <li>• modalitások jelölése a megfelelő segédigék kitételével</li> <li>• összevonások megengedése<sup>50</sup></li> </ul> <p><u>Főnevekre, névmásokra</u> vonatkozóan:</p> <ul style="list-style-type: none"> <li>• névmások explicit kitétele; személytelen szerkezetek kerülése<sup>51</sup></li> <li>• rövidítések kiváltása</li> </ul> <p>További javaslatok:</p> <ul style="list-style-type: none"> <li>• funkcióigék kerülése</li> <li>• köznyelvben is használatos kifejezések jogi kontextusban történő újradefiniálásának kerülése<sup>52</sup></li> <li>• definiált kifejezések konzisztens használata</li> <li>• technikai, jogi és idegen zsargon mellőzése</li> <li>• archaizmusok kerülése</li> </ul> | <ul style="list-style-type: none"> <li>• átlátható tagolás, lehetőség szerint több rövidebb mondat alkalmazása egyetlen többszörösen összetett konstrukció helyett</li> <li>• a főbb mondatrészek (A, Á, T) egymás szűk környezetében tartása, ha lehetséges<sup>53</sup></li> <li>• többszörös tagadás és „kivételek alóli kivételek” kerülése</li> <li>• a legfontosabb normatív tartalomnak a mondaton belüli előre mozgatása</li> </ul> |
| <p><b>Nagyobb kontextusra (~bekezdés) vonatkozó ajánlások</b></p>   | <p><b>A teljes szövegre vonatkozó általános megjegyzések</b></p>  |
| <ul style="list-style-type: none"> <li>• tételmondat alkalmazása</li> <li>• átvezető szavak, kötőelemek használata<sup>54</sup></li> <li>• rövid bekezdések; témánkénti szövegtagolás</li> </ul>  | <ul style="list-style-type: none"> <li>• szemléltetés példák alkalmazásának szükségessége</li> <li>• többtagú felsorolások lehetőség szerint listákba szedése, táblázatok, illusztrációk alkalmazása</li> <li>• kereszthivatkozások minimalizálása</li> </ul>   |

4. táblázat: Összefoglalás a PLM irányzatban keletkezett közérthetőségi javaslatokról a FPLG alapján

<sup>49</sup> Például az *eleget tesz az X-szerinti feltételeknek* alak helyett az *X alapján alkalmazható* típusú szerkezetek előtérbe helyezése.

<sup>50</sup> Pl.: *don't* használata a *do not* helyett a természetesebb, a beszélt nyelvhez közelebbi összhatás elérése érdekében.

<sup>51</sup> Főként olyan esetekben, ahol a cselekvés elvárt végrehajtója nem kormányzati szerv, hanem a hozzá forduló ügyfél, pl.: „Az adóvisszatérítéssel kapcsolatos dokumentumok másolatait benyújtandók.” de: „Az adóvisszatérítéssel kapcsolatos dokumentumok másolatait Önnek be kell nyújtania.”

<sup>52</sup> vö. 2.3.2 fejezet, a jogi szakszókincs négy jellemző osztálya.

<sup>53</sup> A felsorolás több ponton összecseng a korábban a pszicholingvisztika kapcsán kifejtett, ott a megértést nehezítő tényezőként felsorolt nyelvi jegyekkel (3.1.2 fejezet).

<sup>54</sup> Mint például: „ezért”, „vagyis”, „ezzel ellentétben” stb.

### 3.4 A „Jog és nyelv” irányzat

Az FPLG esetében tehát már láthattuk, hogy a jogi és nyelvész szakemberek együttműködése a hivatali éraban korántsem példa nélküli. Tisztán az egyes tudományterületek szemszögéből megközelítve a kérdést kitűnik, hogy a jogi nyelv nyelvészeti szempontú tanulmányozásának számos válfaja ismert, amelyek eltérő neveken több specifikus közegben vizsgálják a jogi témákhoz kötődő nyelvhasználatot. Ezek egyik legnagyobb hatású vonala a „jog és nyelv” néven ismertté vált interdiszciplináris irányzat.

Az Egyesült Államokban a 70-es években kezdődtek meg azon kutatások, amelyek a jog- és nyelvtudomány egyes területeinek metszetében igyekeztek empirikus módszerekkel vizsgálni a jogi környezetben kialakuló speciális beszédhelyzeteket. A jogtudományon belül kiforrott „jog és társadalom” irányzat<sup>55</sup> egyfajta specializálódása lett később a „jog és nyelv” (law and language) kutatási irány. Ennek korai szakaszában a vizsgálódások fő csapásiránya az egyenlő bánásmód követelményének gyakorlati érvényesülése volt, célja pedig, hogy feltárja azokat a nyelvi-nyelvhasználati okokat és szokásokat, amelyek miatt ez az elv gyakran nem tud maradéktalanul érvényre jutni. Ilyen értelemben tehát a fő kérdés a jogban megjelenő hatalmi viszonyok (nyelvi) megjelenése, gyakorlása, vagy éppen az ez által elkövetett visszaélések feltárása volt (Vinnai 2017: 87-90).

A 70-es évek végén és a 80-as évek első felében indultak meg ténylegesen azon vizsgálódások, amelyek immár a jogi nyelv nyelvészeti releváns jegyeivel foglalkoztak. A főbb kutatási területeket ekkoriban három nagyobb csoportra lehetett felosztani (Levi – Walker: 1990):

- a.) a tárgyalótermi nyelvhasználat (pl. az esküdteknek adott bírói utasítások érthetősége<sup>56</sup>)
- b.) a nyelv, mint a jog tárgya (szólásszabadsággal kapcsolatos kérdések, szóbeli bűncselekmények stb.)
- c.) írott jogi szövegek vizsgálata.

---

<sup>55</sup> A „jog és társadalom” vonulatába tartozó írások rendszerint a társadalmi fejlődés, jogi normák és jogi gyakorlatok, valamint a jog és a jogi intézmények szerepe és funkciói közötti metszéspontot kutatják. Leggyakrabban multidiszciplinárisak, éppen ezért gyakran törekszenek szociológiai, pszichológiai, antropológiai, nyelvészeti, etnográfiai, kriminológiai, közgazdaságtani, vagy épp politikatudományi álláspontok ötvözésére.

<sup>56</sup> Erre jó példa Charrow – Charrow (1979) már említett elemzése, melynek során a cél a szövegösszefüggések kognitív feldolgozási folyamatai során keletkező nehézségek azonosítása, továbbá azon lexikai, szintaktikai, szemantikai jelenségek azonosítása volt, amelyek kihatnak a feldolgozás nehézségére (vö. 3.1.3 fejezet).



Ez utóbbi nagy csoportba tartozott például a szerződések, formanyomtatványok, valamint a jogszabályok nyelvezetének elemzése is (Vinnai: 2017: 94), éppen ezért ez a terület állhatott legközelebb ahhoz, hogy a PLM számára gyakorlati támpontokat nyújtson.

Mindhárom esetben kiemelt figyelmet kaptak olyan, a jogszociológiában is élénk érdeklődésre számot tartó kérdések, mint a jog és a jogi hatalom természetének megítélése, vagy akár olyan (inkább a szociolingvisztikához közel álló) problémák, mint az amerikai jogi eljárások során a résztvevők többsége által alkalmazott, a férfiakra jellemző beszédstílus előnyben részesítése. A korszak sokszínűségét jól jellemzi, hogy olyan kutatásra is akadt példa, amely a jogi egyetemek elsőéves hallgatóit bevonva igyekezett választ találni a jogvégzettek és ügyfelek között megfigyelhető kommunikációs szakadékra azáltal, hogy részletesen feltárta a jogi nyelv elsajátítása során jelentkező nyelvhasználati változásokat a diákok esetében (Mertz: 2007)<sup>57</sup>.

A 90-es évektől kezdve a korábbinál gyakrabban kerültek a figyelem középpontjába a jogi nyelvhasználat nyelvi eszközei. Ebből a szempontól a jog és nyelv kutatások egyik legjelentősebb, egyben történetileg legújabb válfaja az igazságügyi nyelvészet (forensic linguistics, a témáról lásd pl. Coulthard – Johnson 2013, vagy magyarul Kurtán – Tamás – Vinnai 2020). Az igazságügyi nyelvész szakértők közreműködhetnek például olyan eljárásokban, ahol megnyilatkozások szerzőjének azonosítása vagy egy másik dokumentum szerzőjével való megfeleltetése a cél. Erre jó példa a Vincze (et.al. 2021) által ismertetett kutatás, melyben a szerzők ezen folyamat (részleges) automatizálásra tesznek kísérletet. Hasonlóan fontos kérdéskör annak megállapítása is, hogy egy jogi kifejezéseket is tartalmazó szöveg mennyire érthető a laikusok számára (Vinnai 2017: 112), ezért pedig a PLM által keresztülvinni kívánt elvek akár tekinthetők az igazságügyi nyelvészet egyfajta gyakorlati alkalmazásának is.

Egy másik megközelítés, amelynek alkalmazásával közelebb kerülhetünk a jogi szövegek nyelvi jellemzőinek feltáráshoz, egyúttal az érthetőségüket megnehezítő (esetleges)

---

<sup>57</sup> Az elemzés egyebek mellett megállapította, hogy ennek a fajta kommunikációs szakadéknak a háttérében egy egészen egyszerű, de szisztematikus jelenség áll. A joghallgatók ugyanis a jogesetek elemzése során egy olyan speciális látásmódot fejlesztenek ki, amelyben a lényegi vonások kiemelése válik prioritássá, az egyes tényezők kiválnak a társadalmi kontextusból, és esettípusokká válnak a bennük szereplőkkel együtt (Vinnai 2017: 102).

Ez a fajta alkalmazkodás nem meglepő, tekintetbe véve a jogi normák absztrakt jellegét, vagyis, hogy a bennük megfogalmazott elvek és szabályok a valós eseményeknél sokkal általánosabban vannak jelen, ezen általános elvek „átfordítása” a konkrét jogesetre pedig a jogalkalmazó feladata. Ezáltal azonban a jogi kontextusban elhangzó megnyilatkozás vagy mondat is elsődlegesen szaknyelvi jellegűvé válik, amely háttérbe szorít(hat)ja a laikus szempontjait.

tényezők megismeréséhez, az a fenti felsorolás harmadik pontjaként említett írott jogi szövegek vizsgálata. Ahhoz, hogy a témában megalapozott következtetéseket vonhassunk le, segítséget nyújthat a nyelvtudománynak egy viszonylag új iránya, a korpusznyelvészet, amely pontosan a nagy szöveges adatbázisok elemzésével foglalkozik.

A digitalizációnak köszönhetően ma már nem ritka a több százezer, vagy millió szóból álló jogi tematikájú korpuszok összeállítása sem, ez pedig lehetővé teszi, hogy a jog nyelvezetét a 2000-es évek előtt még elképzelhetetlen módszerekkel is elemezhessük.

### **3.5 Korpusznyelvészet**

Egy általánosabb megközelítés szerint a *korpusz* ténylegesen előforduló írott, vagy lejegyzett beszélt nyelvi adatok gyűjteménye. Ezek esetében a szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegeket tartalmaz, és nemcsak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket is (bekezdés, mondat)<sup>58</sup>.

Az erre épülő nyelvtudományi ág, a korpusznyelvészet Lindquist – Levin (2018: 47) meghatározásában „...egy módszertan, amely számos kapcsolódó módszert foglal magában, amelyeket számos különböző elméleti irányzatú tudós használhat. Másrészt nem tagadható, hogy a korpusznyelvészet gyakran társul egy bizonyos nyelvszemlélettel is. E szemlélet középpontjában az áll, hogy a nyelv szabályai a használaton alapulnak, és hogy változások akkor következnek be, amikor a beszélők a nyelvet használják az egymással való kommunikációra. Az érvelés szerint, ha egy adott nyelv, például az angol nyelv működése érdekel, akkor érdemes a nyelvhasználatot tanulmányozni. Ennek egyik hatékony módja a korpusz-módszertan alkalmazása”.

#### **3.5.1 Az „adatvezérelt fordulat”**

A fenti kijelentések mögött az a fajta szemléletváltás érthető tetten, amely a nyelvtudományban az elmúlt évtizedben kezdett kiteljesedni. A XX. század nagy részében a nyelvészeti kutatásokban az a szemléletmód tükröződött, amely szerint a grammatikai kompetencia fontosabb jelenségeit nem lehet közvetlenül nyelvi adatokból kinyerni, sem pedig

---

<sup>58</sup> Forrás: <https://www.mnsz.hu>

azokból induktív módszerekkel levezetni. A nyelvi kompetencia a chomskyánus álláspont szerint (pl. Chomsky 1957, 1980) velünk született és ezáltal nagyrészt genetikailag meghatározott készség (innátizmus). Ennek a következtetésnek főként olyan megfigyelések adták az alapját, mint a nyelvelsajátítás életkorhoz kötött volta, az a tény, hogy a gyerekek ebben a korban hiányos / töredékes nyelvi adattal találkoznak, mégis képesek az anyanyelvüket helyesen megtanulni, illetve, hogy véges hallott jólformált mondat után is képesek végtelen számú grammatikailag helyes mondat megformálására.

Ez a szemléletmód magával vonta azt is, hogy az érdeklődés tárgya az absztrakt nyelvi képesség, illetve a generatív (univerzális) grammatika volt. Ugyanezért domináns adatforrásként az introspekció, a fő adattípus pedig az ebben gyökerező grammatikalitási ítéletet volt. E mögött az a vélekedés húzódott meg, miszerint a nyelvet generáló szabályrendszer minden anyanyelvi beszélőnek rendelkezésére áll, ezért pedig az általuk önmegfigyeléssel elérhető adatok is eleve helyesek kell, hogy legyenek.

A korpusznyelvészeti megközelítés ezzel szemben a szövegekben eleve meglévő minták felismerésével próbál a nyelvre vonatkozó megállapításokat tenni. Azáltal, hogy a század végére a digitális adathordozók fejlődésének köszönhetően fokozatosan lehetőség nyílt nagy mennyiségű adat (esetünkben szöveg) gyorsan és könnyen hozzáférhető módon való tárolására, valamint hatékony feldolgozására, az „adat” szerepe számos tudományterületen (így a nyelvtudományban is) jelentősen felértékelődött (Yasuhiko et al. 2016).

Az ún. első generációs korpuszok óta, mint amilyen például az angol nyelvű szövegeket tartalmazó LOB (Lancaster-Oslo/Bergen) korpusz (Johansson 1978), számos nagyságrendi ugrás következett a tartalmazott szövegszavak számát illetően, így mára nem szokatlan a több(száz)milliárdos szó-, vagy tokenszám sem<sup>59</sup>.

Az általános célú korpuszok mellett, mint például angolra a már beszélnyelvi leiratokat is tartalmazó Bank of English (Moon 2009) vagy magyarra az MNSZ2 (Oravecz et al. 2015) és a Gigaword Korpusz (Oravecz et al. 2014) mellett megjelentek a tematikus célú, azaz valamilyen speciális kutatási célt szolgáló korpuszok is. Ilyen tervet valósított volna meg például a névelem-felismerést támogatni kívánó, de végül befejezetlenül maradt HunNER

---

<sup>59</sup> A nyelvtudományban, de különösen a számítógépes nyelvészetben inkább a token kifejezés használatos. A „szó” köznapi jelentésétől főleg az különbözteti meg, hogy token lehet például írásjel is ("a kutya": 2 token, 2 szó, de „ma, reggel”: 3 token, 2 szó). Újabb alkalmazásokban, például Nagy Nyelvmodellek (Large Language Models) esetében ráadásul egy-egy szó is több tokenre bomlik. Ennek többnyire a modellek belső működésében keresendő okai vannak.

projekt (Simon et al. 2006) de a nemrégiben elkészült NerKor<sup>60</sup> is. A gépi tanítási feladatok (Machine Learning, a továbbiakban: ML) kapcsán kiemelendő például a gyakran morfológiai és szintaktikai elemzők tanítására és tesztelésére használt Szeged Dependency Treebank (Vincze et al. 2010), vagy éppen a számítógépes pragmatikai kutatások támogatására készült HuComTech Corpus (Hunyadi et al. 2011).

### 3.5.2 Miskolc Jogi Korpusz

Az összeállított korpuszok tematikájának diverzifikálódása mára eléggé kiforrottnak tekinthető, így léteznek olyan specializált gyűjtemények is, amelyek a jogi szaknyelv egzakt, kvantitatív és kvalitatív vizsgálatát egyaránt lehetővé teszik.

Egy nemrégiben elkészült hazai példa erre az OTKA-112172 kutatás során nyelvész, jogász és informatikus szakértők együttműködésével létrejött Miskolc Jogi Korpusz (a továbbiakban: MJK, Vincze 2018), amely 2,4 millió szövegszót tartalmaz, és az összeállításánál a fő cél a magyar jogi nyelv jellemző szövegeinek/ szövegtípusainak minél pontosabb reprezentálása volt<sup>61</sup>.

A korpusz vizsgálata során a kérdésfeltevés a jelen dolgozat eddigi megközelítéséhez képest valamelyest fordított volt; a szerzők arra kerestek választ, hogy a hazánkban a jogi szaknyelvvvel kapcsolatosan élő előítéletek annak bonyolultságáról és nehéz érthetőségéről mennyiben megalapozottak (Vincze 2018: 9). A válaszhoz a hat szövegosztály szerint szervezett (kódexjog, ítéletek, magyarázó szövegek, fórumok, átiratok és jogszabályok)<sup>62</sup> korpusz beható elemzése statisztikai és korpusznyelvészeti módszerekkel történt.

Összehasonlítási alapként a kutatás a Szeged Dependency Treebank-et (Vincze et al. 2010) alkalmazta, amely a köznyelvet volt hivatott reprezentálni<sup>63</sup>. A szófaji gyakorisági adatok

---

<sup>60</sup> <https://github.com/nytud/NYTK-NerKor>

<sup>61</sup> Hasonlóan a jogi doménből épül fel például a British Law Report Corpus (BLRC, Rizzo – Pérez: 2012), illetve a Corpus of German Law (Vogel et al. 2018) is.

<sup>62</sup> I.) Kódexjog; 1952. évi III. törvény a polgári perrendtartásról; 1998. évi XIX. törvény a büntetőeljárásról; 2004. évi CXL. törvény a közigazgatási hatósági eljárás és szolgáltatás általános szabályairól; 2012. évi C. törvény a Büntető Törvénykönyvről; 2013. évi V. törvény a Polgári Törvénykönyvről **teljes szövege**, II.) bírósági és törvényszéki ítéletek szövegei, III.) metanyelvi / magyarázó szövegek; jogszabályok miniszteri indoklásai és jogi egyetemeken használt tankönyvek szövegei, IV.) jogi fórumok szövegei, V.) bírósági tárgyalásokon és rendőrségi kihallgatásokon készített átiratok (beszélt nyelvi szövegtest), továbbá VI.) Jogszabályok: egyéb jogszabályok **szövegrészletei**.

<sup>63</sup> A vizsgálat során az összehasonlításhoz a SzT eredeti 6 részkorpuszából (fikció, 14-16 év közötti tanulók fogalmazásai, újságcikkek, informatikai- és jogi szövegek, valamint üzleti rövidhírek) a jogi doménbe tartozó részkorpusz eltávolításra került a jobb összevethetőség érdekében.

igazolták például azt az előfeltevést, mely szerint a jogi szövegekben a főnevek aránya jelentősen magasabb, mint a köznyelvben. A MJK esetében ez az arány összesítetten 26,71%-ot<sup>64</sup>, míg a Szeged Dependencia Treebank (a továbbiakban SzT) esetében mintegy 19,29%-ot tett ki. Az igék számarányának alakulása ezzel ellentétesen alakult; a MJK esetében összesen 7,93%, a SzT-et tekintve pedig 11,88% -os arány volt megfigyelhető. Ez a két adat összességében tehát a jogi szövegek erősen nominalizált természetére utal. Az összesített szófaji adatok alapján, és a Kendall-együtthatót alapul véve a köznyelvi szövegekhez a jogi fórumok és az ítéletek álltak a legközelebb (Vincze 2018: 12-15).

| Jellemző                                     | Részkorpusz               | A SzK vonatkozó értékéhez viszonyított arány |
|--|---------------------------|--|
| Mondathossz (átlagos szószám szerint)        | <b>I., II., III.</b>      | <b>MJK &gt; SzK</b>                          |
|  | IV., V., VI.              | MJK < SzK                                    |
| Mondathossz (átlagos karakterszám szerint)   | IV., V., VI.              | MJK < SzK                                    |
|  | <b>I., II., III.</b>      | <b>MJK &gt; SzK</b>                          |
| Átlagos szóhossz                             | IV., V.                   | MJK < SzK                                    |
|  | <b>I., II., III., VI.</b> | <b>MJK &gt; SzK</b>                          |
| Igétlen mondatok                             | ∇                         | MJK > SzK                                    |
| Funkciósavak                                 | II., IV., V., VI.         | MJK < SzK                                    |
|  | <b>I., III.</b>           | <b>MJK &gt; SzK</b>                          |
| Retrokondicionális <sup>65</sup> szerkezetek | <b>I., II., III., IV.</b> | <b>MJK &gt; SzK</b>                          |
|  | VI., V.                   | MJK < SzK                                    |
| Funkcióigék: <i>valamire kerül</i>           | ∇                         | MJK > SzK                                    |

5. táblázat: Pszicholingvisztikailag releváns jegyek (jelmagyarázat: I.: kódexjog, II.: ítéletek, III.: metanyelv, IV.: fórumok, V.: átiratok, VI.: jogszabályok, ∇ : minden részkorpuszra érvényes, **kiemelés**: ahol az adott jellemző gyakorisága nagyobb) forrás: Vincze (2018: 31-33).

A függőségi nyelvtani szintaktikai elemzés megmutatta továbbá, hogy a kódexjogi, valamint jogszabályi szövegek viszonylag sok közbevetést tartalmaztak, emellett több alá- és mellérendelő szerkezetet, ami az összetett mondatok magasabb arányára is enged

<sup>64</sup> Részkorpuszok szerinti felosztásban: Fórumok – 20,84%, Ítéletek – 28,56%, Kódexjog – 30,19%, Jogszabályok – 28,82%, Metanyelvi szövegek – 28,70%, Átiratok: 16,08%.

<sup>65</sup> Fordított kondicionális szerkezet, amely esetében a következmény megelőzi a feltételt; a *ha* → *akkor* fordítottja, pl.: „Elmegyek vizsgázni, de csak (**akkor**,) **ha** előtte rendesen felkészültem.”

következtetni. Hasonlóképpen Kendall-együtthatóval jellemezve, a köznyelvtől vett legnagyobb távolságot a szintaktikai viszonyok alapján a jogszabályszövegek és a kódexjogi szövegek mutatták (Vincze 2018: 16-20).

A 3.1.2 fejezetben már részletezett pszicholingvisztikai tényezők közül az elemzés során vizsgáltakra tekintettel az 5. táblázat szemlélteti a Miskolc Jogi Korpusz (a továbbiakban: MJK) egyes részkorpuszainak tulajdonságait.

Az egyes (rész)korpuszok továbbá a Jaccard-távolságot figyelembe véve az alábbiak szerint rendezhetők sorba:

Köznyelv < Átiratok < Fórumok < Jogszabályok < Kódexjog < Ítéletek < Metanyelv

4. ábra: a MJK részkorpuszainak távolsága a SzK szövegeitől (forrás: Vincze 2018: 33, 20. ábra).

A fenti jellemzők, valamint a szövegtípusok (részkorpuszok) Jaccard-távolság szerinti sorbaállításából kirajzolódik a kép, mely szerint a jogi domén szövegei nyelvileg sem egységesek, azonban az előzetesen feltételezett nyelvi egyszerűsítésre is módot adó jellemzők valóban strukturálják azokat.

A kapott eredmények alapján a jogi szövegeket a referenciakorpusztól (SzK) elválasztó leginkább hangsúlyos tulajdonságok;

- a funkciógék fokozott jelenléte,
- az ígétlen mondatok kiugró gyakorisága,
- retrokondicionális szerkezetek többszöri előfordulása,
- illetve a megnövekedett átlagos szó-, valamint mondathossz.

Fontos azonban megjegyezni, hogy a fenti felsorolás nem érvényes egységesen minden vizsgált részkorpusz esetében, azaz a jellemzők gyakorisága a választott szövegtípus szerint is variábilis.

### 3.6 Plain Language: ISO-standard

Ahogy Vinnai (2023) is írja, a PLM sikerei közül a legutóbbit a közérthető fogalmazás nemzetközi sztenderdjének kidolgozása és közzététele jelenti. A szabvány<sup>66</sup> 2023 júniusában jelent meg a Nemzetközi Szabványügyi Bizottságnál (International Organization for Standardization).

Az elkészült dokumentum a 3. fejezetben ismertetett meghatározás alapján strukturálva tesz javaslatokat a szöveg fogalmazójának. Visszaidézve, a legfontosabb kritériumok ebben a keretben azt várják el a szövegtől, hogy az olvasó számára biztosítsa a szükséges információ **megtalálását**, annak **megértését**, illetve, hogy a megtalált információ olyan formában álljon rendelkezésre, amely az olvasó számára könnyen **felhasználható**.

A sztenderd is ennek megfelelően szerveződik. A konkrét útmutatót tartalmazó rész (5. Guidelines) négy fontosabb egységre bomlik, amelyek egyenként

- a közérthetően megírandó dokumentumba kerülők információk relevanciáját,
- az egyes információk könnyű fellelhetőségét,
- azok megértésének elősegítését, illetve
- felhasználhatóságának növelését tárgyalják.

Ezek közül az első fejezet főként a szöveg célközönségének azonosítására helyezi a hangsúlyt, kiemelve például, hogy fontos a várható olvasók átlagos olvasottságának, kulturális háttérének és nyelvi képességeinek felmérése, céljainak meghatározása, és a kontextus ismerete is, amelyben az adott szöveget olvasni fogják.

A második fejezet főként a szöveg tagolására, a vizuális és egyéb szemléltető elemekre helyezi a hangsúlyt. Kiemeli például, hogy a legfontosabb információt a bevezető gondolatok közé érdemes elhelyezni, és hogy folyamatok leírásában a legcélszerűbb a kronologikus sorrendet követni, valamint, hogy ha egy információra várhatóan csak keveseknek lesz szüksége, azt érdemes később elhelyezni, mint azt, amely mindenkit érint.

A harmadik fejezet talán a dolgozat szempontjából a legfontosabb, hiszen ez foglalkozik a szöveg szoros értelemben vett érthetőségével. Az itt felsorolt nyelvi jellemzők természetesen követik azokat a főbb szempontokat, amelyek a 3.3.4 fejezetben már előkerültek. Az alábbi

---

<sup>66</sup> ISO/DIS 24495-1 Plain language

felsorolás néhány példát hoz, szembeállítva azoknak az FPLG-ben már megismert ekvivalenseivel.

- **Lexikai szint:**

- ISO: rövidítések használata csakis indokolt esetben, például, ha az olvasóközönség jobban ismeri a rövidített formát, vagy ha a kiírt teljes alak különösen hosszú (ISO 24495-1 2023: 7).
- FPLG: rövidítések kerülése, vagy ha szükséges, helyettesítés a teljes alakkal.
- ISO: szakkifejezések használata csakis akkor, ha az olvasók értik és preferálják azokat, vagy ha azok megtanulása az olvasó céljai érdekében szükséges (uo.).
- FPLG: technikai, jogi és idegen zsargon mellőzése, ha mód van rá.

- **Mondatszint:**

- ISO: olyan mondatszerkezetek alkalmazása, amelyek az olvasó számára ismerősek, pl. a szokásos (angol) szórend megtartása az alany, ige, tárgy tekintetében (uo.).
- FPLG: a főbb mondatrészek (alany, állítmány és tárgy) egymás szűk környezetében tartása.
- ISO: redundáns szavak, clichék és homályos módosítószavak elhagyása.
- FPLG: funkciógék kerülése.

Bár a fenti felsorolás közel sem teljes, az ennyiből is kivehető, hogy a legtöbb esetben, amikor a sztenderd javaslatai lefordíthatók konkrét nyelvi jelenségekkel kapcsolatos ajánlásokra, azok egybeesnek a PLM, és konkrétan az FPLG ajánlásaival.

Mindezekon felül persze a dokumentum számos olyan esetet ismertet, amelyek konkrétan a szövegalkotónak szólnak, ezeket összeolvasva pedig konzisztens és részletes útmutatót kapunk. A standard egésze 96 svájci frankért vásárolható meg, egyedi licenz alapján másolása és terjesztése tilos. Éppen ezért a jelen helyen mindössze példálózó jelleggel néhány kiragadott részletet van lehetőség közölni a dokumentum tartalmának illusztrálása céljából.



## **4. A közérthetőségi törekvések eddigi eredményei**

Az eddigiekben röviden összefoglaltam a jogi nyelvhasználat és a közérthetőség fogalmának kapcsolatáról a szakirodalomban fellelhető kutatási irányokat, a főbb történeti állomásokat, valamint azt, hogy egy-egy adott megközelítés milyen (preferált vagy kerülendő) nyelvi jellemzők megragadása által igyekszik meghatározni a „közérthető megfogalmazás” mibenlétét.

A jelen fejezetben az ezen irányzatok által elért eredmények ismertetésére térek ki, különös tekintettel azon kormányzati programokra, jogszabályokra, esetleg civil kezdeményezésekre, amelyek a közérthetőség kérdése nyomán születtek. Ez különösen releváns, hiszen a jogi / hivatalos szövegek érthetőségének elősegítése, habár csak nyelvészeti és jogtudományos kutatások metszetében valósulhat meg sikerrel, alapvetően mégis gyakorlati probléma. Ilyen formán tehát ha a tudományos eredmények nem képződnek le valamilyen formában magukban a jogszabályokban, és a téma nem kap kellő (akár kormányzati, akár jogalkalmazói) figyelmet, akkor a kidolgozott megoldási javaslatok, eredmények érvényre jutása sem biztosítható.

A magyarországi eredmények értékeléséhez és egy közérthetőséget támogató szoftveres megoldás lehetőségeinek felméréséhez ezért jó alappal szolgálhat, amennyiben a fontosabb, közérthető nyelvhasználattal összefüggő eredményeket több szinten is áttekintjük. A jelen fejezet ehhez a nemzetközi trendek, irányvonalak és eredmények ismertetésével járul hozzá.

### **4.1 Egyesült Államok**

A közérthetőséget támogató szabályozások terén az USA vitán felül élen jár. A PLM kezdete óta az egymást követő amerikai kormányzatok reakciója a jogi dokumentumok közérthető megszövegezése iránt konstansnak mutakozó igényre igen változatos képet mutat.

Az első kormányzati intézkedés, amelyben már jelentős hatása volt a PLM kapcsán kialakult, a hivatali nyelvezetét illető közhangulatnak Richard Nixon 1972-es állásfoglalása. Ez lényegében azt tartalmazta, hogy a Federal Register (az USA kormányának hivatalos napilapja) a továbbiakban a laikusok számára is érthető nyelvezeten kell, hogy íródjon. 1978-ban a szövetségi kommunikációs bizottság (Federal Communications Commission)

rövid kérdés-válasz párok formájában adott ki szabályozást a CB rádiókra vonatkozóan<sup>67</sup>. A szövegben emellett kiemelt figyelmet fordítottak a személyes névmások és az aktív mondatszerkesztés használatára, egyéb stilisztikai jegyei pedig ezt teszik az első Plain English-ben írt jogszabállyá<sup>68</sup>.

Szintén 1978-ban Jimmy Carter elnöki rendeletben<sup>69</sup> kötelezte a kormányzati ügynökségeket arra, hogy meglévő rendeleteiket vizsgálják felül aszerint, hogy azok változatlan formában megtarthatók, módosítandók vagy hatályon kívül helyezendők, bevezetve ezzel a visszamenőleges szabályozási felülvizsgálatot (retrospective regulatory review). A rendelet előírja a kormányügynökségek számára, hogy:

- egyszerű és (költség)hatékony rendeleteket dolgozzanak ki,
- a javasolt szabályokat azonosítsák az általuk elvárt jelentős társadalmi hatásokkal,
- vegyék tekintetbe a gazdasági tartalmú szabályok gazdasági hatásait,
- és vizsgálják felül a hatályos kormányzati ügynökségi szabályozásokat, hogy azok megfeleljenek az iránymutatás céljainak.

Egy következő elnöki rendeletben<sup>70</sup> Carter előírta, hogy az ügynökségek csak a szükséges űrlapokat alkalmazhatják, azokat kötelesek a lehető legegyszerűbbé és legrövidebbé tenni, valamint, hogy előirányozzák a papírmunka feldolgozásához szükséges költségeket éves szinten<sup>71</sup>.

A következő időszakban, Ronald Reagan elnöksége alatt időszakos visszalépés történt a PLM térnyerésében. Az új elnök visszavonta elődje nyelvi egyszerűsítést célzó rendeleteit<sup>72</sup>, annak ellenére, hogy kinevezett kereskedelmi minisztere, Malcolm Baldrige továbbra is szorgalmazta a közérthető nyelvhasználatot a kereskedelmi és ipari gazdasági szegmensben (Bowen et al.: 1991).

---

<sup>67</sup> Federal Register, April 4, 1978, 13976-13988.

<sup>68</sup> A törvényszöveg összegző része (Summary) szó szerinti idézésben (kiemelés a szerzőtől): „The FCC is adopting all new rules for the Citizens Band (CB) Radio Service. The rules are simpler than they used to be and are written in „plain English”. We expect voluntary compliance with the CB rules to increase as a result of our action.”

<sup>69</sup> E.O. 12044

<sup>70</sup> E.O. 12174

<sup>71</sup> Ugyanennek a hatékonysági törekvésnek a nem sokkal későbbi állomása volt a bürokrácia csökkentését célzó Paperwork Reduction Act (The Paperwork Reduction Act of 1980) is.

<sup>72</sup> Bár a Paperwork Reduction Actet továbbra is hatályban tartotta. Az intézkedés fő indoka az volt, hogy véleménye szerint a hangsúlyt a törvényi szabályozások költséghatékonyságára kell helyezni, szemben a szabályozások megfogalmazási módjának fejlesztésével.

Az ezt követő időszak inkább stagnálást jelentett a PLM hivatalos elismertségében, annak ellenére, hogy számos, összességében előremutató esemény történt az amerikai törvényhozásban.

Ilyen volt például Bill Clinton tevékenysége, aki hivatali ideje alatt ismét felélesztette a közérthető nyelvhasználatot, mint kiemelt kormányzati célt. A közérthető megfogalmazásról szóló memorandumában a kormányzati tisztségviselők számára írja elő az ehhez szükséges elvi kereteket, két későbbi elnöki rendelete<sup>73</sup> pedig ismét előírta ezt a kormányzati ügynökségeknek is<sup>74</sup>. Aelnökétől, Al Gore-tól származik az a híres idézet is, miszerint „a közérthető megfogalmazás állampolgári jog”, amellyel a hivatali kommunikációs csatornák nyelvhasználatának egyszerűsítésére utalt (Deiterich et al.: 2006).

Nagyjából az ezredforduló utáni évektől kezdve a PLM egyre inkább szervesen integrálódott a legkülönbözőbb kormányzati szervek kommunikációs politikájába, elég csak a Szövetségi Repülési Igazgatóság (Federal Aviation Administration), a Nemzeti Egészségügyi Intézmények (National Institutes of Health) vagy a Mezőgazdasági Minisztérium (Department of Agriculture) PLM-et támogató programjaira gondolni.

Talán a legnagyobb hatású mozzanat a 2010. október 13-án Barack Obama elnök által aláírt törvény a közérthető megfogalmazásról (Plain Writing Act of 2010<sup>75</sup>), amelyet három későbbi elnöki rendelet is megerősített<sup>76</sup>.

A Plain Writing Act az egyes kormányzati ügynökségeket minden általuk kiadott, vagy felülvizsgált dokumentumban közérthető megfogalmazás használatára, munkavállalóik ilyen irányú képzésére, a törvény betartása érdekében pedig adminisztratív folyamatok megalkotására kötelezi, ezen felül előírja, hogy az érintett szervezetek honlapján külön szekció kialakítására történjen meg, amely fórumot nyit a rendelkezéssel kapcsolatban beérkező megkeresések kezelésére, végül, de nem utolsósorban pedig előírja a szervezeteknek egy, a témával foglalkozó kapcsolattartó kijelölését is.

A jogelmélet és jogszociológia, valamint a nyelvtudományi munkák egyfajta szintézisét testesíti meg az amerikai kormányzat által 2011 márciusában kiadott szövetségi közérthetőségi iránymutatás, a 3.3.4 fejezetben már tárgyalt FPLG.

---

<sup>73</sup> E.O. 12866, E.O. 12988

<sup>74</sup> „All information provided to the public by the agency shall be in plain, understandable language.” E.O. 12866, sec. 6

<sup>75</sup> H.R. 946; Pub.L. 111–274

<sup>76</sup> E.O. 13563

A már említetteknek megfelelően a dokumentum részletes tanácsokat tartalmaz szövegszerkezési elvekre vonatkozóan, a szövegeknek a várható olvasóközönség szempontjait szem előtt tartó kontextualizációja kapcsán, lexikális szinten szófaji kategóriák szerint lebontva taglalja a legtipikusabb finomítandó szerkezeteket és kifejezéseket, ezen felül külön fejezetet szentel a webes szövegek hatékony összeállításának, és tárgyalja az egyes bekezdések, mondatok optimálisként javasolt szórendjét, tagolását és hosszát is.

Habár jellegéből adódóan idézett szerzői<sup>77</sup> többségükben a jog területéről érkeztek, valamennyien figyelembe vesznek nyelvészeti kutatásokat, eredményeket is munkáikban. Ennek köszönhetően az elkészült útmutatóban mindkét szaktudományos álláspont képviselteti magát.

## 4.2 Európai Unió

Az Európai Unió jelenleg 27 tagállammal és 24 hivatalos nyelvvel rendelkezik. Az egyes intézmények összesítve több mint 4000 fős fordítói és 800 fős tolmács gárdával rendelkeznek. Ebben az inherensen többnyelvű környezetben az elmúlt években az angol vette át az általánosan elfogadott *lingua franca*, azaz közvetítőnyelv szerepét a korábban jellemző francia helyett. Éppen ezért ezt használják az uniós szerveknél dolgozó politikusok, az adminisztratív és civil szervezetek tagjai a legtöbb esetben saját anyanyelvük helyett is (Olivares-Beltrán – Morell 2017).

Az évek során az angolnak egy minden mástól elkülönülő nyelvváltozata is kialakult a Brüsszel környéki területen, ezt gyakran euro-angolként (Euro-English vagy Eurish) aposztrofálják. Az itt használt nyelvváltozatot főként lexikai tekintetben nagyban befolyásolják egyéb nyelvek is, mint például a már említett francia nyelv (Forche 2012).

Az ebben a régióban születő dokumentumok jelentős része ezen felül természetesen jogi természetű; ez a két tényező együttesen pedig sok esetben kihívás elé állítja a nyelvezetet nem ismerő olvasót, ha értelmezni szeretné a keletkezett szövegeket. Ez még akkor is igaz, ha egyébként az illető az angol valamely standard változatát anyanyelvi szinten beszéli.<sup>78</sup> A tényezőknek ez az együttállása (a nem standard angol, a jogi és egyéb zsargonok, és az ehhez

---

<sup>77</sup> pl.: Veda R. Charrow, Robert P. Charrow, Bryan Garner, Joseph Kimble

<sup>78</sup> Ennek oka főként az egyes lexikai elemek, vagy idiomatikus szerkezetek itt létrejött új jelentései, vagy a meglévő jelentések módosulása. Jellemző példa az eltérő szóhasználatra a *forsees* használata a *provide* ige helyett, a *file* megnevezés helyett a *dossier* alkalmazása vagy éppen, ha döntéshozatalról van szó, akkor a *made* ige helyett az *adopted* gyakori mondatba illesztése.

adódó, az uniós intézményekhez kötődő terminológia) jelentősen megnehezíti nem csak a PLM elveinek alkalmazását, de a hatékony kommunikációra is negatív hatással bírhat akár az egyes intézmények között, akár az intézmények és az állampolgárok között.

Egy ilyen fajta akadályoztatás, akadályozottság esetén, amennyiben az EU intézményei és a hozzájuk forduló uniós polgárok kapcsolatát tekintjük, a praktikusság és hatékonyság elvi sérülésén túl jogi összeférhetlenség is felmerül. Az Európai Unió Alapjogi Chartája 41. cikke kimondja ugyanis, hogy;

- 3) „Mindenkinek joga van ahhoz, hogy ügyeit az Unió intézményei, szervei és hivatalai részrehajlás nélkül, tisztességes módon és ésszerű határidőn belül intézzék.”

A fentiekből, illetve a tisztességes eljáráshoz való jogból az eljárás nyelvi dimenziójának figyelembevételével impliciten következik az EU intézményeitől a joghoz nem értők irányában megfogalmazott kommunikáció világos és hatékony volta is.

#### **4.2.1 Uniós szint**

Az Egyesült Államokban egészen a PLM kezdetektől kiemelt jelentőséget kapott a nyelvi egyszerűsítésnek a fogyasztóvédelemre gyakorolt pozitív hatása. Ennek a pozitív hatásnak a kulcsa természetesen az, hogy a közérthető kommunikáció segítségével a tájékoztatás általános minősége jelentősen javítható. Kedvező körülmény, hogy ez a témakör az Unió esetében is központi szerepet foglal el, elég csak (ismételten) az Alapjogi Charta néhány vonatkozó cikkére gondolni. Ilyen például a 36. cikk, amely az általános gazdasági érdekű szolgáltatásokhoz való hozzáférésről rendelkezik, továbbá a 38. cikk is, amely egyetlen lényegre törő állításban fogalmazza meg, hogy „Az Unió politikáiban biztosítani kell a fogyasztók védelmének magas szintjét”.

Az Európai Bizottság első olyan kiadványa, amely tekintetbe véve a fentieket, javaslatokat tett a kötelékében dolgozó fordítók és szerzők számára az 1982-ben megjelent, majd azóta is folyamatosan frissített English Style Guide (legfrissebb változatban: European Commission 2020). Bevezető fejezetében a kiadvány több jelentős munkát is megemlít a PLM nemzetközileg is elfogadott ajánlásaira vonatkozóan (Cutts 1999, Williams 1995). Emellett azonban arra is felhívja a figyelmet, hogy a jogi vagy bürokratikus nyelvezetnek mind az egyes dokumentumok előkészítő folyamataiban, mind pedig a törvényhozásban

megvan a maga szerepe és létjogosultsága. Grasso (2018) ezt annak jeleként értelmezi, hogy a szöveg paradox módon egyszerre kötelezi el magát az egyszerű nyelvhasználat (Plain Language) alkalmazása mellett, és deklarálja, hogy a szakzsargonnak van létjogosultsága bizonyos környezetben. Azonban tekintetbe véve az egyes dokumentumok jellemző címzettjeinek körét (vö. 1.2 fejezet) egy olyan, szintén plauzibilis olvasat is következik, amely szerint a dokumentum sokkal inkább az egyes intézmények belső kommunikációjában tartja elsődlegesnek a már bevett eurolektust, míg a „nagyközönség” felé címzett dokumentumokban teret enged a nyelvi egyszerűsítés egyes formáinak.

Az első nagyobb volumenű megállapodás, amely tartalmában már érintette az egyes intézmények kommunikációs nyelvezetét is, az 1999-ben hatályba lépett „Intézményközi megállapodás a jogalkotás minőségének javításáról”, amely az Európai Parlament, az Európai Unió Tanácsa és az Európai Bizottság között jött létre és módosított változata mind a mai napig hatályban van.

A dokumentum I. szakasz (1) bekezdése szövegezésében a következőket deklarálja:<sup>79</sup>

- 4) „A három intézmény egyetért abban, hogy hatáskörük gyakorlása során ... tiszteletben tartják az uniós jog alapelveit, így a demokratikus legitimitáció, a szubszidiaritás és az arányosság, valamint a jogbiztonság elvét. Emellett megállapodnak arról, hogy az uniós jogszabályok szövegezésekor ösztönzik az egyszerűséget, az egyértelműséget és a következetességet, valamint ösztönzik a jogalkotási folyamat lehető legnagyobb mértékű átláthatóságát.”

A dokumentum következő bekezdése ennél is konkrétabban fogalmaz, amikor is kijelenti, hogy:

- 5) „A három intézmény egyetért abban, hogy az uniós jogszabályoknak érthetőnek és világosnak kell lenniük, biztosítaniuk kell, hogy a polgárok, közigazgatások és vállalkozások könnyen megérthessék jogaikat és kötelezettségeiket...”

A világos és érthető megfogalmazás elvárása olyan újabb, nagyhatású rendeletekben is felmerül, mint például a 2016-ban hatályba lépett Általános Adatvédelmi Rendelet, ismertebb nevén GDPR<sup>80</sup>. A dokumentum fő érdekessége, hogy a hangsúlyt elődlegesen a szabályozás által érintett ipari szereplőkre vonatkoztatva hangsúlyozza a közérthető

---

<sup>79</sup> Kiemelések az 5) – 6) példában a jelen szerzőtől.

<sup>80</sup> (EU) 2016/679

megfogalmazás fontosságát, ezzel mintegy kiterjesztve a korábban nagyrészt csak önreflexív módon követendőnek jelölt elvet. Jó példa erre a preambulum 39. és 58. bekezdése, de a rendelet főszövegében<sup>81</sup> is az alábbi olvasható:

- 6) „Az adatkezelő megfelelő intézkedéseket hoz annak érdekében, hogy az érintett részére a személyes adatok kezelésére vonatkozó ... valamennyi információt és ... minden egyes tájékoztatást tömör, átlátható, érthető és könnyen hozzáférhető formában, világosan és közérthetően megfogalmazva nyújtsa ...”

Ennek a dokumentumnak érdekessége, hogy semmilyen meghatározást nem tartalmaz (akár értelmező rendelkezések formájában), amely támpontot nyújtana a világos és közérthető nyelvezet definitív jegyeinek megállapításához.

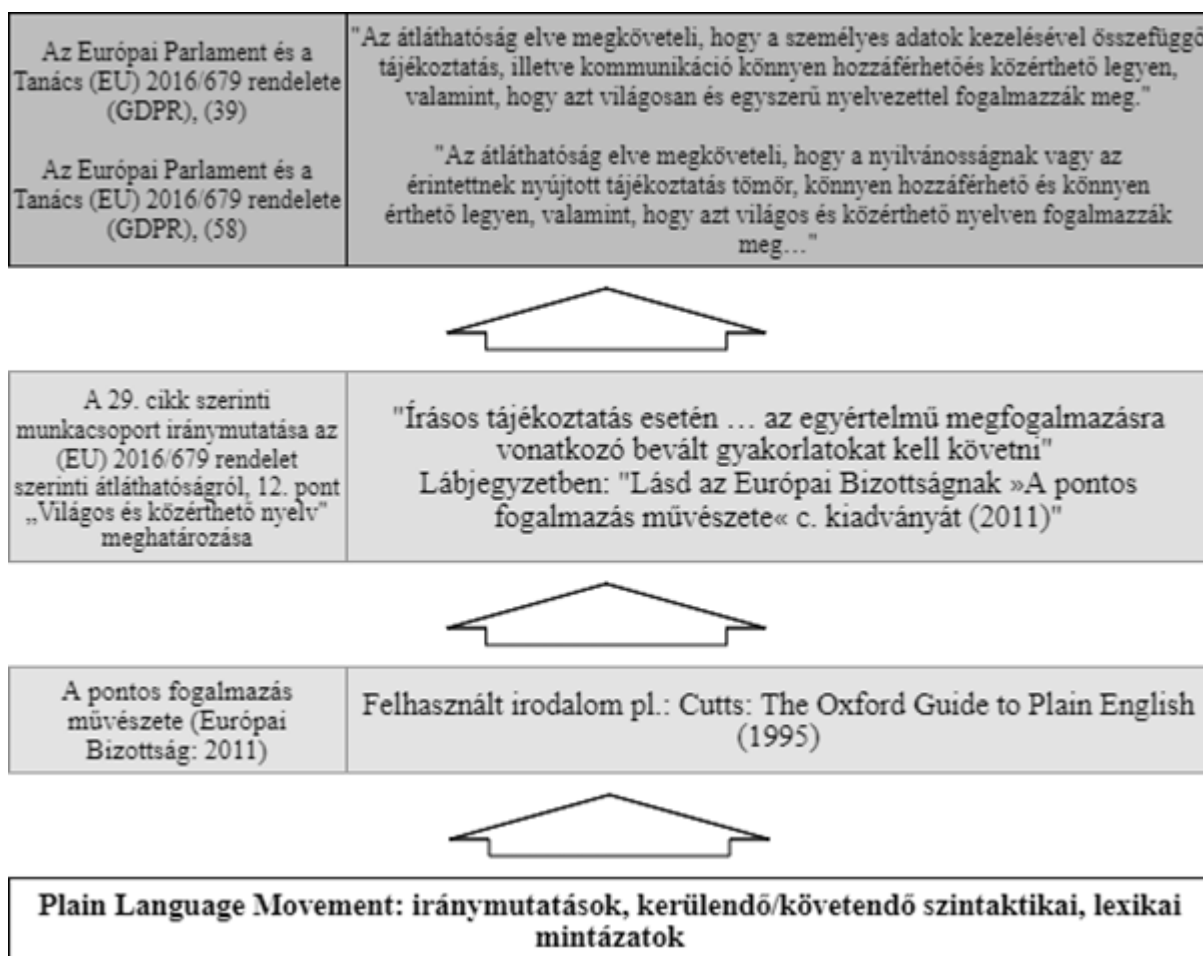
Ezt a hiányt ugyanakkor a 29. cikk szerinti munkacsoport iránymutatása az (EU) 2016/679 rendelet szerinti átláthatóságról segít pótolni. A GDPR-ban szereplő „világosan és közérthetően” fordulat értelmezése kapcsán egyértelműen visszaüt az Európai Bizottság korábbi kiadványára (European Commission 2011), amely pedig angol változatában egyebek mellett Cutts (1995) közérthető angolról szóló munkáját nevezi meg forrásként. A dokumentumok egymásra hivatkozását az 5. ábra szemlélteti részletesen.

A Bizottság közérthetőséggel foglalkozó kiadványa magyarul „A pontos fogalmazás művészete” címen jelent meg, és referenciáit (az angol kiadás forrásainak tételes felsorolásán túl) olyan magyar munkákkal bővíti ki, amelyek segítenek az irányelvek nyelvspecifikus alkalmazhatóságát is megteremteni (pl.: B. Kovács 1999). Ennek köszönhetően a magyar változat nem pusztán az eredeti tükörfordítása lesz, de olyan lokalizációs törekvésekre is utal, amelyek visszatükrözik a „közérthetőség” képlékeny fogalmának inherensen nyelvspecifikus jellegét is.

Mindezek mellett az Európai Bizottság Fordítási Főigazgatósága 1998-ban indította útjára indította a „Fight the Fog” kampányt, melynek elsődleges célja a szövegszerkesztők és fordítók képzése volt az egyszerű nyelvhasználat elsajátítására (Wagner 2005: 224). A kampány során a célnyelvet kizárólagosan az angol jelentette. Ez annak már korábban említett, az Unió szervek közötti lingua franca státuszából is következik.

---

<sup>81</sup> III. fejezet, 1.szakasz 12.cikk (1)



5. ábra: Közérthető nyelvhasználatról foglalkozó források egymásra hivatkozási hierarchiája.

Szintén az Európai Bizottság kezdeményezésére 2010-ben elindult a *Clear Writing* kampány, amely informális elődjénél sokkal kiterjedtebben kívánta megvalósítani a nyelvi egyszerűsítés fontosságának propagálását.<sup>82</sup> A 2004-es EU bővítés után nyilvánvalóvá vált, hogy kiemelt szükség van a közérthető nyelvezet fontosságának hangsúlyozására, eszközök biztosítására, amelyekkel az elérhetővé válik, képzések tartására a témában, továbbá annak az elképzelésnek az előtérbe helyezésére, hogy az egyes uniós dokumentumokat azok véglegesítése és közzététele előtt egy anyanyelvi beszélőnek is meg kell vizsgálnia (Wagner

<sup>82</sup> A kampány elindításában komoly szerepet játszhatott, hogy az Európai Unió akkori soros elnökségét betöltő Svédország kiemelt célként kezelte az EU-s intézmények átláthatóságát, továbbá a második „Barrosobizottság” gazdasági céljainak (az EU szabályozási környezetének egyszerűsítése és javítása a kisvállalkozások adminisztratív költségeinek visszaszorítása érdekében) eléréséhez is fontos volt a hírneve szerint túlbürokratizált intézményrendszer közelebb hozása az európai uniós polgárokhoz.



2010: 4). A kampányhoz kapcsolódó kiadvány (European Commission, 2016) azóta is frissül; benne a PLM-mel összhangban levő gyakorlati fogalmazási tanácsok szerepelnek az unió 24 hivatalos nyelvén.

A kampány 10 éves fennállásának alkalmából 2019-ben az Európai Bizottság konferenciát is rendezett, amelynek fő üzenete (az elért eredmények összefoglalásán túl) a közérthető nyelvezet fontosságának hangsúlyozása volt, valamint a lehetséges fejlődés útjainak keresése az elkövetkező évtizedre vonatkozóan.<sup>83</sup>

#### **4.2.2 Tagállami szint**

A jelen fejezetben néhány fontosabb tagállami, a törvénykezést érintő,<sup>84</sup> valamint a civil szervezetek által jegyzett legfontosabb eredményeket mutatom be országok szerinti bontásban.

Általánosságban elmondható, hogy az uniós tagállamok közül a közérthető hivatali nyelvezet kialakításában a skandináv államok járnak az élen. Esetükben a PLM mindkét jellemző motivációja a legtöbb esetben egyszerre van jelen, vagyis többnyire azonos fontosságúnak tekintik a könnyen érthető kommunikáció nyújtotta fogyasztóvédelmi előnyöket, illetve az egyszerű érthetőséget demokratikus jogként is értelmezik például az állami szervek kommunikációjával szemben megfogalmazott elvárások kapcsán.

##### **4.2.2.1 Svédország**

Bandle (et al. 2005) a hivatali nyelvezet egyszerűsítésére irányuló erőfeszítések sikerességének megalapozójaként a Svédországban régóta működő nyelvtervezési, nyelvművelési folyamatokat említi, amelyek (más országokhoz hasonlóan) kulcsszerepet játszottak az egységes nyelvtani rendszer, helyesírás és a standard nyelvváltozat kijelölésében.

---

<sup>83</sup> Mivel az uniós dokumentumok jellemzően átesnek valamilyen fordítási folyamaton, így a fentiekén túl az Európai Bizottság számos programot tart fenn a fordítási munkák egységes minőségellenőrzésére, azoknak közérthető nyelvezettel történő megvalósítására (pl.: Programme for Quality Management in Translation – 22 Actions, Quality Management in Translation). A jelen tanulmánynak azonban nem célja fordítástudományi kérdéseket vizsgálni, így ezek bővebb ismertetésére nem kerül sor.

<sup>84</sup> Amennyiben van rá példa az adott ország kontextusában.

A 20. század második felére a figyelem a politikai és adminisztratív szövegek felé fordult. Ennek háttérében egyrészt az állt, hogy az állami adminisztratív kapacitás a jóléti állam megerősödésével párhuzamosan folyamatosan növekedett, a kormányzati szervek személyi állománya ennek megfelelően jelentősen gyarapodott, a korábbi írásos kommunikációs hagyományok pedig folyamatosan destabilizálódtak, másrészt pedig ezzel párhuzamosan egyre több állampolgár került kapcsolatba a hivatalokkal és azok dokumentumaival (Bandle et al. 2005: 1977).

Ezzel párhuzamosan a hivatali nyelvezet lehetséges revíziójának ügyében a kormányzat is támogatólag lépett fel, melynek köszönhetően a nyelvtudomány képviselőinek hozzájárulása a jogalkotáshoz (de legalábbis a jogi szövegalkotáshoz) mára évtizedes múltat tekint vissza. A legfontosabb állomások között említhetjük (Asprey 2003: 87):

- az első nyelvész szakértő bevonását, amely már a 70-es évek közepén megtörtént a kormányzati kabinetiroda tevékenységébe, melynek feladata ekkoriban a törvények és rendeletek szövegének modernizálása volt, valamint
- 1980 óta egy nyelvész szakértői csoport működését, amely a parlamenthez kerülés előtt felülvizsgálja az írásos anyagokat, továbbá képzéseket tart és nyelvi útmutatókat készít.

Ez a szemléletmód hamar továbbterjedt a minisztériumok tevékenységén, amelyet jól példáz, hogy:

- 1993 óta a kormány megbízásából közérthetőségi szakértői csoport (*Plain Swedish Group*) működik<sup>85</sup> három bíró, két nyelvész, három politikatudós és egy titkársági tag közreműködésével,
- 2003-as adatok szerint is az összes svéd kormányzati szerv több mint fele részt vett valamilyen közérthetőséget támogató projektben.

Nord (2018: 9) a PLM törekvéseinek kettős megítélésre hívja fel a figyelmet, hangsúlyozva, hogy azok fő célpontját a kormányzati szervek tevékenysége és kommunikációja adja (szemben a magánvállalatokkal, ahogyan az például a PLM intézményesülése előtt az USA esetében megfigyelhető volt, vö. 3.3.1 fejezet). Ennek okaként azt az elvárásbeli különbséget

---

<sup>85</sup> A csoport később a Jogi- és nyelvészeti tervezet-felülvizsgáló osztály szervezeti egységében dolgozott tovább, a svéd igazságügyi minisztérium fennhatósága alatt. A különböző nyelvtervezési kérdések jelenleg az Institute for Language and Folklore hatáskörébe tartoznak: <http://www.isof.se/om-oss/verksamhet/about-the-institute.html>

nevezi meg, ahogyan a svéd lakosság a hivatalokhoz és a piaci szereplőkhöz viszonyul. Indoklása szerint, míg a magánvállalatok irányában a korrekt tájékoztatás, a megfelelő minőségű ügyfélkapcsolati csatornák kialakítása – a jó kommunikációt is beleértve – „mindösszesen” egyfajta elvárásként mutatkozik meg, amelynek való megfelelés egyszersmind az ilyen szereplők üzleti érdeke is, addig az állami szervektől érkező dokumentumok esetében a polgárok demokratikus jogaik részeként tekintenek arra, hogy azok számukra is könnyen megérthetőek legyenek.

Az érvelésben tehát ugyanaz a logika érhető tetten, amely eredetileg az amerikai mozgalmat is életre hívta; a közérthető hivatalos szövegek nemcsak a jogbiztonságot segítik megerősíteni, de növelik az adminisztratív folyamatokba vetett bizalmat, valamint a demokratikus érdekérvényesítő képességet is.

Ezt jól alátámasztja Landqvist – Spetz (2020) érvelése, akik szerint a svéd nyelvtervezést kezdetektől fogva áthatja az alapvető demokratikus funkció. Ez a gyakorlatban azt jelenti, hogy a nyelvet/nyelvhasználatot az állampolgárok alapvető eszközének tekintik ahhoz, hogy jogaikat a társadalomban gyakorolni tudják.

2009-ben az országban hatályba lépett a nyelvhasználati törvény,<sup>86</sup> amely a 11. § -ban kimondja, hogy:

- 7) „Az állami hatósági szervek nyelve (~nyelvhasználata) ápoltnak, egyszerű és érthetőnek kell, hogy legyen.”<sup>87</sup>

Ezzel Svédország egy, az USA-hoz hasonló fejlődési utat járt be, legalábbis ami a közérthetőségi törekvések törvényi szintre emelkedését illeti.

#### **4.2.2.2 Norvégia**

Hasonlóan Svédországhoz, Norvégiában is hatályban van olyan jogszabály, amely a hivatali nyelvhasználatot érinti. Az ország közigazgatási eljárásról szóló törvénye (hivatalos angol fordításban: Administrative Procedure Act) 9. §-ában a következő rendelkezés áll:

---

<sup>86</sup> The Language Act (2009:600)

<sup>87</sup> A törvény hivatalos angol fordításában: „The language of public authorities must be cultivated, simple and comprehensible.”

- 8) A hatóságoknak egyértelmű, könnyen érthető és lényegre törő nyelvezetet kell használniuk<sup>88</sup>.

Ennek előzménye, hogy a tudományos közösségben már egészen a 80-as évektől aktuális kérdésnek számított a közérthető nyelvezet hivatali használata (Sarangi – Slembrouck 2014), 2008 után pedig már a kormányzati kommunikációs politika is rendszerszinten foglalkozott vele. A kapcsolódó a legfontosabb állomások, amelyek a közérthető nyelvhasználatot vizsgálták más-más szempontból, az alábbiak szerint alakultak (Johannessen et al. 2017 alapján):

- **nyelvpolitika** (2008): nyelvi oktatás fejlesztése, bürokratikus nyelvezet egyszerűsítése, ahol lehetséges, fenntartva, hogy előfordulhatnak speciális kontextusok, ahol ez nem kivitelezhető;
- **kommunikációpolitika** (2009): a fő hívószavak a koherencia és a kormányzati kommunikáció nyitottsága volt;
- **e-kormányzat politika** (2012): iránymutatások készültek az adminisztrációs nyelv fejlesztésére;
- **digitális ügyrend** (2016): felhasználóbarát szemlélet kialakítása a digitális ügyintézési csatornákon.

Ez utóbbi szempont, vagyis a „felhasználó”, a hivatalos csatornával interakcióba kerülő középpontba helyezése jól mutatja az eltolódást a PLM két irányzata között. Az első főként objektív alapokon képzelel el a közérthetőség megvalósulását és iránymutatások készítésével, problémásnak ítélt és preferált nyelvi formák, lexikai elemek stb. meghatározásán keresztül kívánja elősegíteni ennek megvalósulását. A másik irányvonal ezzel szemben főként a pragmatikai tényezőket, a szöveg címzettjének feltételezhető kontextuális tudását veszi figyelembe, éppen ezért nem hisz az első változat által propagált uniformizált megoldások hatásosságában (Nord 2018: 10).

---

<sup>88</sup> A törvény hivatalos angol fordításában: „An authority shall use language that is clear, easy to understand and to the point.”

### **4.2.2.3 Dánia**

Dánia esetében, összehasonlításban a két korábban elemzett skandináv országgal, a kormányzati szférától erősen távolodó tendencia mutatkozik meg a PLM-kezdemenyezések intézményi háttérét figyelembe véve.

Habár a hivatali nyelv felülvizsgálatának igényét Norvégia és Svédország esetében sem a kormányzati szervek fogalmazták meg elsőként, az ezzel összefüggő elvárások hamar bekerültek a hivatali köztudatba is, végül pedig az erre reagáló szabályozások törvényi erőre is tudtak emelkedni.

A dán bíróságok szervezete (Danmarks Domstole), valamint Koppenhága önkormányzata indított a témával összefüggő kampányokat a 2010-es évek elején, ugyanakkor a szövegek korpusznyelvészeti elemzése arra mutat rá, hogy az érintett hivatalok dokumentumainak szövegezése nem változott szignifikánsan ezek hatására (Kjærgaard 2015: 131). Civil szervezetek (mint például a Dán Fogyasztói Tanács – Forbrugerrådet Tænk) szintén gyakran tűzik napirendre a kérdést, azonban elvárásaik célpontja inkább a magánszektor tájékoztatói anyagainak felülvizsgálata.

### **4.2.2.4 Finnország**

A finn megközelítés központjába az a gondolat került, mely szerint a PLM-mel kapcsolatos munkát állami szervezeteknek kell koordinálnia annak érdekében, hogy a munka a lehető leghatékonyabban mehessen végbe. Piehl – Sommardahl (2015) több lehetséges megközelítést említ arra vonatkozólag, hogy az egyes országok milyen módon emelik be a közérthető fogalmazás megvalósítását saját államszervezetük működésébe:

- egy lehetséges módszer a kormányzati kommunikáció nyelvezetének törvényi szabályozása; ezt az utat választotta például az USA, Svédország, valamint Finnország,
- főleg az angolszász országokban a kapcsolódó „lektorálási”, egységesítési és egyéb nyelvi ellenőrző szolgáltatásokat a kormányok jellemzően külső szervezetektől veszik igénybe,
- ezzel szemben például a skandináv államokban vagy Észtországban a közérthető kommunikációhoz kapcsolódó nyelvtervezési munkát kormányzati szervezetek végzik.

A finn példa esetében több ilyen szervezet is létezik, amelyek közül az Institute for the Languages of Finland nyelvtervezési intézet részt vett többek között az 1982-ben elfogadott finn nyelvtörvény<sup>89</sup> előkészítésében is, amely egészen 2003-ig hatályban is maradt.

A hivatali nyelvhasználatot külön szabályzó 2003-as törvényben<sup>90</sup> a Finn Igazságügyi Minisztérium a következők szerint fogalmaz a törvény 9. cikk 1. bekezdésében:

- 9) „A hatóságnak megfelelő, világos és *(könnyen)* érthető nyelvezetet kell alkalmaznia”.<sup>91</sup>

A fentiekén túl jelentős szerep jut még az olyan programoknak is, amelyek kormányzati és civil szervezetek, valamint egyetemek együttműködésével például kérdőívvezéssel keresik a választ arra, hogy az átlagember hogyan éli meg a hivatali nyelvhasználattal való találkozást, milyen előkészítés szükséges a hivatali dolgozók mindennapjaiban egy közérthetőségi törekvés bevezetéséhez, vagy éppen arra, hogy miképpen lehet a nyelvhasználattal kapcsolatos tudatosságot a mindennapi hivatali eljárások részévé tenni (Piehl – Sommardahl 2016: 78).

A legfontosabb következtetés, amely mindezeket a kezdeményezéseket összeköti az a központi gondolat, amely szerint (ahogyan az már a svéd példa kapcsán is felmerült), az olvasóbarát tájékoztató anyagok és egyéb hivatali szövegek erősítik a közhivatalokba vetett bizalmat, továbbá, hogy a közigazgatás csakis akkor lehet legitim, ha a rá bízott ügyeket hatékonyan és könnyen intézi, amelynek fontos aspektusa a nyelvhasználat is (Piehl – Sommardahl 2016: 81).

#### 4.2.2.5 Németország

Az eddig ismertetett országoktól eltérően a német nyelvterületen zajló, és Plain Language-hez kötődő eredményekről angol nyelven rendkívül limitált információ érhető el az elért eredmények jelentős volumene ellenére is.

Pottmann (2020) betekintést enged a mozgalom németországi történetének néhány fontosabb állomásába, ideértve a közreműködő szervezeteket, amelyek a német hivatali

---

<sup>89</sup> 1922 Language Act (Språklag), kiegészítve 1982 -ben. 2003-tól: Language Act No. 423/2003.

<sup>90</sup> Administrative Procedure Act (434/2003)

<sup>91</sup> *Dötltel szedve*: a jelen dolgozat szerzőjének kiegészítése. A törvény hivatalos angol fordításában: „An authority shall use appropriate, clear and comprehensible language.”

nyelv közérthetőbbé tételében szerepet játszanak, valamint a legfontosabb eredményeket, melyeket e kezdeményezések mindmáig felmutattak.

Talán a legfontosabb különbség az angolszász rendszerhez képest az egyszerűsített megfogalmazás célcsoportját érinti. Míg előbbi esetében a közérthetőségi elvárások fő kiindulási pontja, egyben az ilyen módon elkészített szövegek célcsoportja a prototipikus értelemben vett „átlagember”, addig a német megközelítés alapvetően azokra fókuszál, akik valamilyen szervi vagy kognitív adottság hiánya miatt, esetleg eltérő anyanyelvükből adódóan nem beszélnek megfelelő szinten a német nyelvet, ezért esetükben a hivatali szövegek megértése extra korlátokba ütközik. Ennek megfelelően az a három csoport, akikre véleményük szerint kiemelt figyelmet kell fordítani;

- a fogyatékosokkal élők,
- a (funkcionális) analfabéták,
- illetőleg a németet, mint idegennyelvet tanulók (Pottmann, 2020: 81-82).

A nyelvterületen a közérthető megfogalmazás fontosságának jó indikátora lehet, hogy a Német Szövetségi Köztársaságban 2002-ben elfogadott „A fogyatékos személyek esélyegyenlőségéről szóló törvény” (Behindertengleichstellungsgesetz<sup>92</sup>) például előírja, hogy a szövetségi kormányzat által fenntartott internetes információs szolgáltatásoknak (nyelvileg is) akadálymentesnek kell lennie annak érdekében, hogy a fogyatékkal élők korlátozások nélkül férhessenek hozzá a szükséges információkhoz.

Jól látszik tehát, hogy az olvasóbarátabb fogalmazásmód propagálása mögötti elvi alapok jelentősen eltérnek az angol, valamint az eddig tárgyalt európai országokban szokásosaktól. Létezik azonban a német szakirodalomban is egy, a Plain Language angolszász értelmezéséhez konvergáló fogalom, illetve ajánlásrendszer. E kettő közül előbbit a Leichte Sprache, míg az angolszász megközelítéssel rokon nyelvi egyszerűsítési kezdeményezést az Einfache Sprache terminusa ragadja meg<sup>93</sup>. A kettő közötti eltérést, valamint a kapcsolódó angol terminológiát a 6. táblázat szemlélteti.

---

<sup>92</sup> [https://www.un.org/development/desa/disabilities/wp-content/uploads/sites/15/2019/11/Germany\\_Equality-for-Persons-with-Disabilities-Act-BGG.pdf](https://www.un.org/development/desa/disabilities/wp-content/uploads/sites/15/2019/11/Germany_Equality-for-Persons-with-Disabilities-Act-BGG.pdf)

<sup>93</sup> Tekintettel arra, hogy a dolgozat tárgya főként inkább az Einfache Sprache „irányzathoz” áll közel, így a Leichte Sprache alkalmazásával kapcsolatos intézkedések bővebb összefoglalását bővebben lásd Radünzel (2017) szerint.

| Német terminus          | Angolszász megfelelő | Jelentés   | Szinonim (angol) megnevezések és kapcsolódó fogalmak                          |
|-------------------------|----------------------|--|---|
| <i>Leichte Sprache</i>  | <i>Easy English</i>  | <u>Könnyen érthetőség</u> : olyan írói stílus, amely segít azoknak az embereknek, akiknek nehézséget okoz az adott nyelv olvasása és megértése. Egyszerűbb, és alacsonyabb szintű olvasási készségeket követel meg, mint a Plain Language. Az angol esetében nevezik még Easy Read-nek is. Könnyen felismerhető például a teljes szöveget jellemző rövid mondatokról, és a sok képi illusztrációról. | Easy Read, Easy Write, Easy Info, Easy Access, easy-to-read, aphasia friendly |
| <i>Einfache Sprache</i> | <i>Plain English</i> | <u>Közérthetőség</u> : olyan „közvetlen” megfogalmazási mód, amely azoknak szól, akik nyelvi készségei átlagosan fejlettek, viszont gyorsan és hatékonyan szeretnék a nekik címzett információt megérteni. Nyelvileg közelebb áll a standard nyelvváltozathoz.   | Plain language, plain writing, layman terms, layperson terms                  |

6. táblázat: A szaknyelvi regiszter szempontjából laikus, valamint az adott nyelvet nehezen beszélő emberek nézőpontját tekintetbe vevő nyelvhasználati irányzatok a németben és az angolban<sup>94</sup>.

A fenti megosztásban az Einfache Sprache „stílushoz” kapcsolódóan Pottmann (2020: 83) a következőket emeli ki követendőként;

1. „egyszerű és követhető mondatszerkezet, gondolati ugrások nélkül,
2. 10-11 szónál nem hosszabb egyszerű mondatok, vagy 15 szónál nem hosszabb összetett mondatok, amennyiben mellékmondat is beágyazásra kerül – ilyenből lehetőleg csak egy szerepeljen a mondatban,
3. aktív (~cselekvő) mondatszerkezet,
4. a szókinccs megválasztásakor törekedni kell az előbeszédhez való hasonlóságra,
5. a metaforákat, idiomatikus kollokációkat és elvont kifejezéseket fel kell váltani konkrét, szó szerinti jelentéssel bíró kifejezésekkel,

<sup>94</sup> vö. [https://centreforinclusivedesign.org.au/wp-content/uploads/2020/04/Easy-English-vs-Plain-English\\_accessible.pdf](https://centreforinclusivedesign.org.au/wp-content/uploads/2020/04/Easy-English-vs-Plain-English_accessible.pdf)



6. az idegen szavakat és a szakkifejezéseket anyanyelvi vagy egyszerűbb szavakkal kell helyettesíteni, vagy ha ez nem megoldható, akkor a szakkifejezések magyarázata jelen kell, hogy legyen (például lábjegyzet formájában),
7. ragaszkodni kell a helyesírási szabályok betartásához”.

Egy más megközelítésben a német Szövetségi Munkaügyi és Szociális Minisztérium kiadványában (Bundesministerium für Arbeit und Soziales – BMAS, 2014: 22-33<sup>95</sup>) a preferált nyelvi jellemzők helyett a diszpreferáltak kerülnek előtérbe. Az FPLG-hez hasonlóan a dokumentum konkrét példákkal is illusztrálja egy „helytelenül” és egy „helyesen” megfogalmazott mondat lehetséges eltéréseit. Az ebben szereplő főbb ajánlások a Leichte Sprache vonatkozásában egyszerű szabályokba szervezve jelennek meg a dokumentumban. Ezek közül néhányat a következő felsorolás ismertet.

- i. Egyszerű szavak használata → az egyszerűség főleg az idegen szavak mellőzését, és ezen felül a morfológiai egyszerűséget jelenti a szemantikai egyszerűség helyett, egyúttal a „formális nyelvhasználat” kerülését (Pottmann, 2020: 84),
- ii. Olyan kifejezések használata, amelyek éppen a megfelelő mennyiségű információt hordozzák, azaz a kontextus, és a befogadó közönség kontextuális tudását figyelembe véve túlságosan tág jelentésű szavak mellőzése. Ide tartozik annak figyelembevétele is, hogy fontos információk ne vesszenek el a szövegből.
- iii. Rövidítések mellőzése → főleg a terminus technicusok esetében.
- iv. Nominalizáció kerülése, ahol az igésítés megoldható<sup>96</sup>.

Amint a fenti felsorolásokból is kitűnik, a Plain Language-hez rokonítható Einfache Sprache esetében a javaslatok sokkal inkább elvont maximák leképezésének tekinthetők (jól összeesengenek a PLM, a pszicholingvisztika, valamint a Jog és Nyelv irányzathoz már ismert ajánlásokkal), ezzel szemben a Leichte Sprache-hez köthető elvárások főként konkrétabb nyelvi jelenségeket neveznek csak meg kerülendőként.

---

<sup>95</sup> <https://www.bmas.de/DE/Service/Publikationen/Broschueren/a752-leichte-sprache-ratgeber.html;jsessionid=EC26A3B117EFAABC1DA541BED2874867.delivery2-replication#:~:text=Leichte%20Sprache%20hilft%20vielen%20Menschen,von%20Texten%20in%20Leichter%20Sprache.>

<sup>96</sup> A fentiek mellett még számos nyelvspecifikus javaslat is olvasható a szövegben, mint például a németben megjelenő kétféle kötőmód (Konjunktiv I., II.) alakok helyettesítése határozószavakkal.

### 4.3 Magyarország

Habár a közérthetőségi elvárást tekintetbe vevő (vagy éppen azt az állami szervek, piaci szereplők irányába elvárássá nyilvánító) törvényi szabályozás a magyar jogban nem számít kifejezetten elterjedt gyakorlatnak (vö. 2. fejezet), ennek ellenére más szempontból a kérdést hazánkban is növekvő figyelem övezi.

A 2.2.3 fejezetben már említettem, hogy a hazai gyakorlatban a normavilágosság (vagyis a jogszabályok pontos értelmezéséhez fűződő elvárás) főként csak a jogalkalmazó szemszögéből értelmeződik az AB gyakorlatában. Ezt a képet azonban valamelyest árnyalja, hogy a hazai jogi közösség (köztük állami szervek is) többször, több irányból is igyekeztek már a közérthető fogalmazás lehetőségeit megvizsgálni.

#### 4.3.1 A Kúria joggyakorlat-elemző munkacsoportjának véleménye

Bár néhány jogszabály már korábban is deklarálta a közérthető nyelvhasználat fontosságát<sup>97</sup>, a Kúria elnöke által 2013. január 17-én a bíróságok szervezetéről és igazságszolgáltatásról szóló 2011. évi CLXI. törvény (Bsz.) 29. § (1) bekezdése alapján a „Határozatszerkesztés” tárgykör vizsgálatára létrehozott joggyakorlat-elemző munkacsoport összefoglaló véleményének<sup>98</sup> publikálása (Orosz 2014) mégis jelentős ugrásnak tekinthető ezekhez képest.

Ennek keretében ugyanis sor került a legfelsőbb bírói fórum polgári és közigazgatási ügyszaka gyakorlatának vizsgálatára dedikáltan a szövegezés és a dokumentumszerkesztés közérthetőséget támogató aspektusainak felderítése kapcsán<sup>99</sup>. A vizsgálatban részt vettek több szakterület képviselői is, ami jól harmonizál a probléma korábban már felvázolt interdiszciplináris jellegével.

---

<sup>97</sup> Jó példa erre az egészségügy, ahol a betegtájékoztatók szövegezésével kapcsolatban például a 30/2005. (VIII. 2.) EüM rendelet 3. § (4) bekezdésének szövegezésében találunk elvi megkötéseket a következők szerint: „A betegtájékoztató véglegesítése céljából a gyógyszer indikációja szerinti betegek célcsoportjaival konzultációt kell lefolytatni annak érdekében, hogy a betegtájékoztató jól olvasható, egyértelmű és könnyen értelmezhető legyen. Ennek eredményét a betegtájékoztató szövegének tükröznie kell.”

Az egészségügy területén hasonló Uniós szabályozások még például:

European Commission: Guideline on the readability of the labeling and package leaflet of medicinal products for human use. Revision 1, 12 January 2009. ENTR/F/2/SF/jr (2009)D/869, valamint

CMDh: Consultation with target patient groups - meeting the requirements of article 59(3) without the need for a full test - recommendations for bridging. Revision 1, April 2009. CMDh/100/2007

<sup>98</sup> 2013.EI.II.JGY.1.1-2

<sup>99</sup> Egy következő munkacsoport büntetőügyekben hozott határozatokat vizsgált, amelyek már nem korlátozódtak kizárólag kúriai döntésekre, hanem alsóbb szintű bírói fórumok határozatait is tartalmazták.

A lefolytatott vizsgálat nemzetközi előzményeként többek között megnevezi a Strasbourgban 2010-ben elfogadott Bírák Magna Chartáját, valamint a Vilniusi Nyilatkozatot. Míg előbbi 16. pontja szerint, mint idézi: „A bírósági iratokat és a bírósági határozatokat érthető, egyszerű és világos nyelvezettel kell megfogalmazni”, addig utóbbi kiemeli, hogy „Az igazságszolgáltatásnak meg kell tennie a szükséges lépéseket annak érdekében, hogy erősödjön a bíróságokba vetett közbizalom. (...) Az ítéletek és más bírósági határozatok világos és érthető nyelvezettel történő megfogalmazása mind elengedhetetlen e cél elérése érdekében” (idézi Orosz 2014: 13).

A munkacsoport informatikai szegmense a határozatszerkesztés gépi támogatási lehetőségeit vizsgálta tüzetesebben. Az implementációs lehetőségek kapcsán főbb megállapításaik egy része a különböző közigazgatási rendszerek közötti integritás javítását érinti, egy nagyobb csoport pedig a kötött szövegváltozatok sokszor ismétlődő, de alapvetően sablonszerű elemeinek gyorsított beillesztésére koncentrált. Ez utóbbihoz sorolják például a per tárgyának listából választhatóságát, vagy a jogszabályhelyek megjelölése után a vonatkozó szövegrészlet automatikus beilleszthetőségét.

A szenzitív adatok gyorsabb anonimizálhatósága mentén utalást tesznek olyan számítógépes nyelvészeti eredmények alkalmazására, mint például a névelem-felismerési algoritmusok használata, amelyek a személynevek, pénzüsszegek, személyi azonosításra szolgáló adatok automatikus megjelölését tennék lehetővé a dokumentumon belül. Ezen felül elszórtan található még utalások más megoldásokra is, például a bekezdések közötti „egy kattintással” hivatkozási rendszer felépítése, de ezek esetében nyelvtudományi relevanciáról nem beszélhetünk (Orosz: 2014, 4-es melléklet).

Az informatikai munkacsoport összefoglalójában megemlíti továbbá, hogy az ajánlásaik alapján megalkotható rendszer egyfajta intelligens szövegszerkesztő funkcióját tölthetné be, amelyet a külföldi gyakorlatban már több területen is alkalmaznak<sup>100</sup>. A szintén az összefoglaló vélemény mellékleteként elérhető, Stíluskönyv néven ismertté vált dokumentum azonban ennél nyelvészeti szempontból messzebb menő javaslatokat is tartalmaz.

---

<sup>100</sup> Ezek egy korai leírását adja például az online jogi szövegek kereshetőségének javítását célzó törekvéseknek Susskind (1988). Az ún. szakértői rendszerek hosszú utat jártak be az angolszász területen, és megítélésük korántsem volt végig töretlen (vö. Leith 2010). Az elmúlt évtizedben főként a gépi tanulásra alapozott rendszerek kerültek előtérbe, ezeken belül népszerűek például a szövegbányászatra alapozó megoldások (Rupali: 2013).

| Lexikai szint  | Szintaktikai szint   |
|--|--|
| <p>Igék esetében:</p> <ul style="list-style-type: none"> <li>igésítés</li> <li>igeidők egységes használata</li> </ul> <p>Stiláris jellegzetességek:</p> <ul style="list-style-type: none"> <li>következetes terminológiahasználat, de szóismétlés kerülése</li> <li>az <i>illetve</i> / <i>illetőleg</i> kifejezések lehetőség szerinti kerülése<sup>101</sup></li> </ul> <p>Többértelmű kötőszavakkal kapcsolatos javaslatok:</p> <ul style="list-style-type: none"> <li>kirekesztő választás indikálása a <i>vagy – vagy</i> szerkezettel az egyszerű <i>vagy</i> alkalmazása helyett</li> </ul> <p>Helytelen használat:</p> <ul style="list-style-type: none"> <li>az <i>amennyiben</i> használata <i>ha</i> értelemben</li> <li>a <i>mégpedig</i> túlhasználata</li> <li>az <i>ugyanakkor</i> használata <i>azonban</i> értelemben</li> <li><i>felé</i> használata <i>-nak/nek</i> rag helyett<sup>102</sup></li> <li>a <i>követően</i> használata az <i>után, nyomán</i> névutók helyett</li> <li>a <i>szemben</i> használata <i>ellen</i> értelemben</li> </ul> <p>Funkcióigék / terpeszkedő kifejezések kerülése, pl:</p> <ul style="list-style-type: none"> <li><i>panasszal él – panaszt tesz</i></li> <li><i>befizetést eszközöl – befizet</i> stb.</li> </ul> <p>Legjellemzőbb kiemelt példák még: <i>helyez, képez, rendelkezik, történő, történt.</i></p> <p>Archaizmusok kerülése:</p> <ul style="list-style-type: none"> <li>pl.: <i>akként, illetén</i></li> </ul> | <ul style="list-style-type: none"> <li>többszörös alárendelések, bonyolult mondat szerkezetek kerülése</li> <li>közbeékelések kerülése</li> <li>alany – állítmány – tárgy – határozók sorrend megtartása<sup>103</sup></li> <li>többszörös birtokos szerkezetláncok mellőzése</li> </ul> |

7. táblázat: A Stíluskönyvben szereplő javaslatok<sup>104</sup>.

Mivel a munkacsoportok szeparáltan vizsgálták az informatikai és nyelvészeti vonatkozásokat, így a kettő határterületét képező korpusznyelvészeti elemzés ekkoriban nem

<sup>101</sup> Az indoklás szerint jogi környezetben szerepelhet a *pontosabban szólva, helyesebben*, valamint *és*, továbbá alkalmanként még *vagy* értelemben is.

<sup>102</sup> pl.: *felperes felé – felperesnek* jelentésben

<sup>103</sup> A dokumentum különösen kerülendőként említi az állítmánynak a mondat végére sorolását.

<sup>104</sup> A felsorolás nem teljes; nem tartalmazza azokat a javaslatokat (például a „dátumok egységes helyesírását”) amelyek inkább helyesírási, mintsem közérthetőségi kérdésként kezelhetők.

készült el. Ennek következtében a nyelvészeti, illetve a szöveg szintaktikai, lexikai jellemzőire vonatkozó javaslatok implementálhatóságának felmérésére sem került sor.

A nyelvészeti munkacsoport javaslatai hasonló kategóriák szerint oszlanak meg, mint a korábban a PLM kapcsán ismertetett ajánlások (vö. 3.3.4 fejezet, 4. táblázat). Külön kiemelendő, hogy nyelvtudományi perspektívából nézve (a vállalt célkitűzéssel összhangban) a „stiláris” javaslatok nem csak a normavilágosságnak az alapvetően a jogalkotó szemszögéből értelmezett nézőpontjából tartalmaznak fontos ajánlásokat, de több ponton összhangban állnak az általánosabb értelemben vett „közérthető fogalmazás” támogatásával is.

Számítógépes nyelvészeti megvalósítás szempontjából a „nyelvi egyszerűsítés, egységesítés lehetőségei” részben tárgyaltak a legjelentősebbek, amelynek lényegi elemeit a 7. táblázat hivatott szemléltetni.

A nyilvánvaló pozitívumok ellenére azonban érdemes megemlíteni, hogy az összefoglaló véleményt a Kúria teljes ülése csak ajánlás jelleggel fogadta el, azaz érdemi kötelezettségről a jogalkalmazás gyakorlatára nézve nem beszélhetünk. Külön érdekesség, hogy a konkrét nyelvi javaslatokat tartalmazó melléklet („Stíluskönyv”) nem kapott egyöntetű támogatást, feltehetően ezért is a nyilvánosság számára mindmáig csak korlátozottan hozzáférhető (Minya – Vinnai 2018: 13).

#### **4.3.2 A „Jog és nyelv” irányzat néhány hazai eredménye**

A hazai kutatások, amelyek a nyelv és a jog kölcsönhatását tették meg vizsgálatuk tárgyául modern értelemben véve az 1980-as években kezdődtek meg Magyarországon az írott szövegekre vonatkozóan, a 2000-es évek elején pedig lassan a beszélt (jogi) nyelvre vonatkoztatva (Vinnai 2014a).

Az első ilyen empirikus kutatás a Miskolci Egyetemen (jogász és nyelvész szakemberek közreműködésével) zajlott, „Nyelvhasználat a jogi eljárásban. Nyelvi átfordítás és ténykonstitúció a jogi tényállás megállapításának folyamatában” címmel. A fő cél azon intralingvális átfordítás tanulmányozása volt, melynek során a büntetőeljárásokban a rendőrségi és bírósági bizonyítás során a laikusok narratívájából jogi (szak)nyelven megfogalmazott állítások válnak. A főbb eredmények között Vinnai (2014b: 62) kiemeli annak a hipotézisnek az igazolását, amely szerint a nyelvi kommunikáció hatékonysága a

ténymegállapítások hatékonyságának is alapvető feltétele, ezért pedig a jogalkalmazásban a nyelvhasználati, pragmatikai törvényszerűségek tudatosítása is szükséges. Ezzel összefüggésben kiemeli, hogy igazolást nyert az a hipotézis is, mely szerint a *tisztességes eljáráshoz való jog* alapvető követelményeként kell tekinteni az érthetőséget, ami hozzájárul a kiszolgáltatottság csökkentéséhez (Vinnai 2014b: 63).

A fent említett tisztességes eljáráshoz való jog követelménye az Emberi Jogok Európai Egyezménye 6. cikk 1. bekezdésében található meg, és a következőképpen szól:

- 10) „Mindenkinek joga van arra, hogy ügyét a törvény által létrehozott független és pártatlan bíróság tisztességesen, nyilvánosan és ésszerű időn belül tárgyalja és hozzon határozatot polgári jogi jogai és kötelezettségei tárgyában. ...”

Ennek összefüggése az eljárások nyelv(használat)i dimenziójával a következők szerint írható le: „Jogi szempontból [n]em lehet ugyanis korrekt (>>fair<<) az eljárás, ha nyelvi, kommunikációs szempontból nem az. Ha tehát a laikusok és a jogi szakemberek közötti interakciókban problémák, hiányosságok, torzítások merülnek fel, akkor az jelentős mértékben kihat a ténymegállapítás folyamatára, és végső soron az eljárás végső eredményére, az ítéletre is.” (Vinnai 2014b: 63).

Ehhez közvetetten kapcsolódik a *joghoz való hozzáférés* problémaköre is. A meghatározás eredeti használatában azon társadalmi erőforrások (pl.: idő, pénz, műveltség stb.) egyenlőtlen eloszlására volt hivatott felhívni a figyelmet, amelyek hiányában a jog segítségül hívása akadályba ütközhet (vö. Szilágyi 2016), azonban a kommunikációnak a laikusokra vonatkoztatott érthetőségét is figyelembe véve kiterjeszhető a (köz)érthető megfogalmazás hiánya esetén a jogérvényesítő képességben fellépő hiányosságokra is.

A fenti kutatás folytatásaként (*A tisztességes eljárás nyelvi aspektusai. A jogi nyelv hatása a joghoz való hozzáférés érvényesülésére* címen) a fő, egymásra épülő kérdések a következők voltak:

- a jogi és a laikus nyelvhasználat között fellelhető olyan eltérések kimutatása, amelyek akadályozhatják a megértést,
- az ilyen eltérések / akadályok (ha azonosíthatók) akadályozhatják-e a tisztességes eljárás követelményének és a joghoz való hozzáférés elvének sérülését,
- a fenti kettő teljesülése esetén megoldási javaslatok kidolgozása (Vinnai: 2014b: 67).

Ennek megfelelően a kutatás során (nagyreszt a 3.5.2 fejezetben bemutatott MJK segítségével) a közreműködők korpusznyelvészeti módszereket is alkalmazva részletesen elemezték a magyar jogi (szak)nyelv nyelvi jellegzetességeit. A projekt zárásaként megjelent kötetben (Szabó – Vinnai 2018) összefoglalt főbb (nyelvészeti vonatkozású) vizsgálati irányok a következők szerint alakultak;

- az alanyi (Balogh 2018) és állítmányi szerkezetek (Dobos 2018), valamint
- az alá- (Sajgál 2018) és mellérendelések (Kurtán 2018) sajátosságai a jogi nyelv rétegeiben,
- szófaji elemzések és doménhasonlósági mérések (pl. köznyelvi szövegekkel összevetve, vö. Vincze 2018),
- a jogi szövegekben alkalmazott központosítás (Kurtán 2018),
- valamint például a jogi eljárásokban (bíróági tárgyalásokon) zajló szóbeli interakciók során elhangzó közlések beható vizsgálata (Vinnai 2018).

A jogi nyelvvel kapcsolatos társadalmi sztereotípiák megalapozottságát (mint például a nominalizált jelleg, a funkciószavak vagy éppen a személytelen szerkezetek fokozott jelenléte) a korpusznyelvészeti elemzés csupán részben igazolta (Vincze 2018: 35-36). Ugyanakkor a megértést nehezítő tényezők közül (bővebben lásd: 3.1.2 fejezet) igazolódott a kanonikus szórendtől (SOV) való, a köznyelvhez képest gyakoribb eltérés, például az ige tárgyi bővítményének a mondat elejére, gyakran a kezdőpozícióba történő kimoztatása, valamint az egyes bővítményeknek az igétől való, a köznyelvihez képest jelentősen nagyobb távolsága (Dobos 2018: 66-68).

#### **4.3.3 Nemzeti Adó- és Vámhivatal – közérthetőségi program I.<sup>105</sup>**

A Nemzeti Adó- és Vámhivatal saját közlése alapján mintegy 10 éve foglalkozik a közérthető kommunikáció kérdésével. A cél egységes programmá 2019-ben állt össze, amikor is a NAV vezetője szervezeti célként fogalmazta meg a hivatal kommunikációjának közérthetőségét, amely után megkezdődött a szervezett közérthetőségi munka nyelvész szakértők bevonása mellett. Mindennek fő célja a jogkövető magatartás elősegítése, a kiadott

---

<sup>105</sup> A fejezet tartalma a NAV munkatársaival e-mail-ben és személyesen folytatott beszélgetések, valamint az általuk a Közérthetőségi Programról küldött írásos összefoglalók feldolgozásán alapszik.

írásos anyagok meg nem értéséből adódó reklamációk és interakciók számának csökkentése, valamint a NAV általános megítélésének javítása volt.

A program fontosabb állomásai között említhetjük például a 2018-ban a Magyar Nyelvstratégiai Intézettel kötött megállapodást, a Közérthetőségi Munkacsoport 2019-es megalapítását, a hivatal belső honlapján (intranet) 2019-ben létrehozott közérthetőségi menüpontot, vagy akár a „kötelező közérthetőségi kontroll” bevezetését, amelynek keretében 2019 októberétől valamennyi adózóknak szánt levél és honlapra kihelyezett tájékoztató kötelezően közérthetőségi felülvizsgálaton esik át.

A program dedikált célkitűzései között szerepel többek között

- a NAV szövegek helyesírásának javítása,
- az alkalmazott mondatok rövidítése, szóhasználatuk egyszerűsítése,
- a szövegek tagoltságának javítása, áttekinthetőségének növelése,
- a jogszabályi hivatkozások kiszervezése a főszövegből lábjegyzetekbe,
- valamint a szövegek lényegének kiemelése azok elejére.

A fenti gondolatok konzisztensek a PLM által leggyakrabban megfogalmazott elvi elvárásokkal. A hivatal szakértőinek beszámolója alapján ezen felül 2021-ig mintegy 20 ezer gépelt oldalnyi szöveg nyert közérthetőbb formát. A munkacsoport az általuk átalakított szövegeket a jelen kutatáshoz is rendelkezésre bocsátotta mind eredeti, mind pedig korrektúrázott verzióban. Ennek a szöveganyagnak a részletes ismertetését, valamint a kutatásban történő felhasználhatóságát részletesen a 6.2.1 fejezet mutatja be.

#### **4.3.4 További eredmények**

Az eddigieken felül megemlítendőek még azok a kezdeményezések, amelyek az olyan központi szervektől érkeznek, mint például az Országos Bírósági Hivatal (OBH). A hivatal döntése szerint 2017 a „közérthetőség éve” volt a magyar bíróságokon, amelynek keretében az ország valamennyi bíróságán kötelező képzés is zajlott „Közérthetőség és szakszerűség a bírósági jogalkalmazásban” címmel. Ennek céljai között már deklaráltan szerepelt a bírósági szervezetek „felhasználóbarát” kommunikációjának elősegítése, valamint a bíróságok szolgáltató jellegének hangsúlyozása.



A képzés tapasztalatait összefoglalva Minya –Vinnai (2018: 15) megállapítja, hogy „a tervek között szerepelt még a teljes bírósági weboldal felhasználóbarát megújítása és a bírósági ügyfélcentrumokban elérhető tartalmak (például ügyfél-tájékoztató kiadványok) olyan szerkesztése is, hogy azok valóban segítséget nyújtsanak a felhasználóknak”. Általános konklúzióként (részben a köznyelviség – szaknyelviség között feszülő ellentétre reflektálva) megállapítja továbbá, hogy „...a jogszabályok nyelvi igényessége megkönnyíti a jogalkalmazást. A fogalmazás nyelvi ellentmondásai viszont veszélyeztetik a jog hatékonyságát. Cél az arany középutra törekvés a köznyelv és a jogi nyelv között” (Minya – Vinnai 2018: 16).

A közérthetőséget érintő, a jogi szakma és a nyelvtudomány képviselői közötti kommunikáció kiemelt fóruma a Magyar Jogi Nyelv folyóirat, amely 2017-es alapítása óta ad lehetőséget a jogi (szak)nyelvvvel kapcsolatos terminológiai, regiszterhasználati, korpuszfejlesztési, nyelvoktatási és nyelvtörténeti kérdések és az ezek kapcsán felmerülő gyakorlati alkalmazási problémák (mint törvényszéki nyelvészet és fordítás-tolmácsolás) megvitatására a tárgykört jellemző interdiszciplináris keretben<sup>106</sup>. Témái között a közérthetőségi kérdések propagálása a kezdetektől kiemelt helyet foglalt el.

A magánszektor kezdeményezései közül kiemelkedő például a Világos Beszéd<sup>107</sup>, amelynek tagjai előadásokkal, képzésekkel segítik az érdeklődő cégek vállalati kommunikációjának közérthetőbbé tételét.

#### **4.4 Összegzés**

A nemzetközi és hazai eredmények tükrében adódik a kérdés, hogy mely tényezők járultak hozzá ahhoz a különbséghez, mely szerint néhány országban az állami hivatalok kommunikációjának a laikus nézőpontjából vett érthetőségét törvény szabályozza, addig például hazánk esetében az eddigi kezdeményezések (bár szimbolikus jelentőségük elvitathatatlan) mindaddig nélkülözik a jogalkotó bármilyen jelentősebb elköteleződését a téma prioritizálása mellett?

---

<sup>106</sup> <https://joginyelv.hu/bekoszo/>

<sup>107</sup> <https://vilagosbeszed.hu/>

A lehetséges okokra más országok példáinak értékelése által következtethetünk, konkrétan azon társadalmi faktorok azonosításával, amelyek az adott országban kedvező kontextust teremtettek ezen törekvéseknek.

A PLM-mel összefüggésben a sikerhez vezető tényezők véleményem szerint az alábbiakban foglalhatók össze:

1. Az Egyesült Államok esetében a nyelvhasználatra, amellyel az egyes hivatalok kommunikálnak, hagyományosan a demokratikus érdekérvényesítő képességnek alávetett médiumként tekintenek. Ebben a kulturális közegben tehát alapelvárás, hogy minden ilyen csatorna törekedjen rá, hogy nyelvhasználatában is képes legyen azokhoz az ügyfelekhez szólni, akik hozzá fordulnak segítségért (hasonlót figyelhettünk meg a skandináv országok esetében is).
2. A jogi szakma relatív nyitottsága az ilyen irányú önreflexióra; ez kiemelten fontos, hiszen az állammal való kapcsolattartásban és a mindennapos ügyekkel összefüggésben egyetlen más szaknyelv sem annyira domináns, mint a jogi szakmáé, érthető tehát, hogy ezzel kapcsolatban fogalmazódnak meg a legalapvetőbb társadalmi elvárások<sup>108</sup>.
3. A jogi nyelv nyelvészeti szempontú tanulmányozása az USA-ban komoly hagyományokkal rendelkezik, az interdiszciplináris megközelítés a területen nem újkeletű, éppen ezért már léteznek kiépült csatornák az érintett szakemberek együttműködésére.
4. A közérthetőség igénye része a jogi kultúrának<sup>109</sup>.

Ezenfelül fontos még kiemelni egy (a hazai példán jól demonstrálható) szempontot is, ez pedig éppen;

---

<sup>108</sup> Kivétel lehet talán az orvosi szaknyelv. A témával dedikáltan foglalkozik például a Magyar orvosi nyelv folyóirat egészen 2001 óta (online: <https://old.orvosinyelv.hu/>).

<sup>109</sup> A jogi kultúra fogalmát bevezető munkájában Friedman a jogrendszert olyan, a társadalom többi intézményétől a legkevésbé sem szeparált jelenségként írja le, amely, jóllehet természetesen rendelkezik önálló belső struktúrával, működésének milyenségére mégis alapvető hatást gyakorol a rajta kívül elhelyezkedő külső társadalmi feltételrendszer is (Friedman 1975). A kettő egymásra hatását olyan kölcsönhatásként mutatja be, amelyben a fennálló jogrendszerre való reflexióként társadalmi-kulturális igények jelennek meg, amelyek sok esetben beépülnek a jogrendszer további fejlődésébe, és hasonlóképpen, a jogrendszer is hatással van az azt körülvevő társadalmi-kulturális környezetre: formálhatja, alakíthatja azt, nem pusztán a szó regulatív értelmében (amennyiben a jogszabályok fő rendeltetése a világ optimális állapotának leírása, az ide vezető előírások deklarálása), hanem a hagyományos értelemben vett kultúra alapértékeinek formálása által is. Eszmetörténeti áttekintésében Fekete Balázs a jogi kultúra fogalmát a fent említett két rendszer közötti közvetítés tereként aposztrofálja, kiemelve, hogy „...az autonóm belső struktúrával rendelkező jogrendszer »milyenségét« döntően meghatározzák a társadalmi-kulturális feltételek, azaz az azzal szemben támasztott különféle, koronként változó társadalmi igények...” (Fekete 2017: 42).

## 5. az állami szerepvállalás.

Ez a fajta szerepvállalás pedig nagyjából a 2010-es évek óta<sup>110</sup> legfőképpen a szervezeti kultúrában gyökerezik, amely egyfajta irányelvek mentén megszabja az adott hivatali rendszer működését, és előrevetíti annak legfontosabb fejlődési irányait. Informálisan úgy is megfogalmazhatjuk, hogy megadja a dolgok „szokásos elintézési módját”. Mindemellett ezt szokás a teljesítmény alapvető mozgatórugójának tartani, amely egyfajta keretrendszert ad a döntéshozáshoz, és az adott szervezet hatékonyságának megítéléséhez is.

Az ilyen kultúrák kialakulásában ugyanúgy érvényesülnek trendek, mint például a digitalizáció trendje az adattárolásban, az adatvezérelt megközelítés és a mesterséges intelligencia fokozott térnyerése számos tudományterületen, vagy például a globalizáció a világgazdaságban.

Azok a hívószavak, amelyek az elmúlt évtizedben leginkább jellemzik az amerikai kormányzat által kialakítani kívánt szervezeti kultúrát, sokszor valamilyen, a versenyszférában már ismert elv alkalmazására szólítanak fel, mint például:

- „nyitott kormányzat”, amely értékeli az átláthatóságot és az együttműködést<sup>111</sup>,
- a tanulás kultúrája: relevancia és hatékonyság priorizálása a környezet változásaihoz való gyors adaptálódás érdekében
- magasan teljesítő szervezetek: a legfontosabb a produktivitás és az innováció, valamint a munkavállalók és az ügyfelek elégedettsége (Customer Experience – CX<sup>112</sup>).

Ez utóbbi, vagyis az ügyfélelégedettség kapcsán pedig két fontos értéktényező van, amelyet az elmúlt években rendre a témához kapcsolnak:

- a legfontosabb, hogy az elégedettség fokmérőjeként a kormányzatba, illetve a kormányzati szervekbe vetett bizalmat emelik ki,
- a másik pedig, hogy a reklamációval, panasszal végződő esetek elkerülésének, az elégedettség és a bizalom kiépítésének alapfeltételeként éppen a közérthető

---

<sup>110</sup> Forrás: U.S. General Services Administration: <https://www.gsa.gov/>

<sup>111</sup> Lásd például: <https://www.gsa.gov/governmentwide-initiatives/gsas-open-government-initiatives>

<sup>112</sup> OMB (Office of Management and Budget) Circular A-11, Section 280 – „Managing Customer Experience and Improving Service Delivery”. Online elérhető: <https://www.performance.gov/cx/assets/files/a11-280.pdf>

kommunikáció alkalmazását tartják kulcsfontosságúnak, vagyis jelentősen felértékelődik a kapcsolattartás nyelvi dimenziója<sup>113</sup>.

Mindezen feltételek összessége ideális környezetet teremt tehát a PLM által képviselt elvek érvényesülésének, ezáltal pedig tágabb értelemben véve a hivatalos / hivatali szövegek közérthetőbbé tételének is.

Az északi államok esetében a fő hangsúly véleményem szerint a demokratikus berendezkedés alapvető feltételrendszerének biztosításán van, amelynek direkt eszköze a hivatali kommunikációnak az átlagemberhez igazítása.

Több szerzőnél megjelenő érv, hogy demokratikus szempontból a bonyolult nyelvezet használata azért jelent komoly problémát, mert megfosztja a polgárokat a politikai döntéshozatalban való részvétel és a döntéshozatal befolyásolásának lehetőségétől. A képviseleti demokrácia eszménye az, hogy minden polgárnak joga és lehetősége van arra, hogy a választott tisztségviselők meghallgassák. E tekintetben a (nem megfelelő hivatali) nyelvhasználat jelentheti az egyik legfontosabb akadályát a demokratikus részvételnek és a polgárok közszférához való hozzáférésének (pl. Johannessen – Berntzen – Ødegård 2017, Lutz 2016).

A hazai kontextusban is kiemelt fontosságú annak hangsúlyozása, hogy a jogi nyelvet nem ismerők számára alapvető érdekérvényesítő képességük egyik eszköze a jogi szövegek megértése, hiszen, ha nem akarnak kiszolgáltatottá válni intézményeknek, és a jogrendszer által részükre kínált lehetőségeket, jogokat és köteleességeket behatóan szeretnék megismerni, kihasználni és betartani, ahhoz a jogi nyelv(ezet) megértése is elengedhetetlenül szükséges (Vinnai 2014b: 66).

---

<sup>113</sup> <https://digital.gov/resources/customer-experience-toolkit/>

## 5. A közérthetőség kritikái

A lehetséges (és elérendő) előnyök mellett a közérthetőségi törekvéseket (angolszász területen természetesen főként a PLM-et<sup>114</sup>) a kezdetektől számos kritika érte az általa választott problémára adható válasz leegyszerűsítéséért, a „közérthető” fogalmának homályban hagyásáért vagy éppen az alkalmazott módszerek hatékonyságának bizonyítatlansága (esetleg bizonytalansága) miatt. Hasonló kritikai észrevételek a hazai szakirodalomban is jelen vannak, legyen szó akár a „jog és nyelv” irányzat képviselőinek meglátásairól, akár a jogtudomány más területein tevékenykedők kritikai észrevételeiről.

Tekintettel a jelen dolgozat témaválasztására, a hivatali nyelvhasználat közérthetőbbé tételét célzó törekvések kritikáinak tárgyalása e helyütt is elengedhetetlen.

### 5.1 A probléma leegyszerűsítése

Assy (2011: 379-380) korábbi forrásokra hivatkozva<sup>115</sup> kiemeli, hogy a PLM képviselőinek álláspontja szerint a közérthető nyelvhasználatnak alapvetően azt a célt kell szolgálnia, hogy a jogot (ideértve a jogszabályokat is) a címzettjei számára olyannyira érthetővé tegye első olvasatra is, hogy az értelmezéshez ne legyen szükségük jog(ász)i segítségre. Meglátása szerint a PLM-et propagáló közösségen belül akár az olyan álláspontok is könnyen támogatásra lelhetnek, mint hogy a jogalkotásnak megfogalmazásában a legkevésbé kompetens olvasókat kell megszólítani a norma lehetséges címzettjei közül annak érdekében, hogy a jog valóban mindenkire szólhasson. Véleményem szerint ez elvi szinten könnyen indokolható, azonban a megvalósíthatóság már számos gyakorlati és technikai kérdést vet fel.

Ellenérvként két fő témát ad meg, melyek közül az első szerint az ilyen javaslatokban azok támogatói jellemzően nem jelölnék meg kritériumokat az érthetőség optimális, elérendő fokára vonatkozóan, a második szerint pedig a hivatalos (jogi) szövegek közérthetővé tételének okait nem specifikálják megfelelően (Assy 2011: 382-384). Érvelése alapján a jog maga túlságosan komplex természetű ahhoz, hogy az azt leíró szövegek (néhány egyszerűbb esettől eltekintve, mint például a közlekedési szabályok, vagy a termékek garanciális feltételei) jogi segítség vagy képzettség nélkül értelmezhetőek legyenek a laikusok számára.

---

<sup>114</sup> Korábbi terminológiában: Plain English Movement, pl. Felsenfeld – Cohen – Fingerhut (1981) munkájában.

<sup>115</sup> Például legjellemzőbben: Law Reform Commission of Victoria (1987).

Hozzáteszi továbbá, hogy véleménye szerint, ha az egyszerűsített nyelvhasználatnak vannak gyakorlati előnyei, azok főként abban érhetők tetten, hogy a jogi szövegeket értelmező, jogvégezett szakemberek számára bírhat értelmezéskönnyítő szerepkörrel, ezáltal növelve azt a „szolgáltatási minőséget”, amelyet az érintett szakemberek a munkájuk során a hozzájuk forduló ügyfeleknek nyújtani képesek. Álláspontja szerint legcélszerűbb lenne a PLM céljait ebben az irányban újra definiálni, amely egyúttal nem aláaknázná, hanem realiztikusabbá és fontosabbá tenné a mozgalom törekvéseit.

A fenti gondolatmenet szellemiségében sokban emlékeztet a hazai AB gyakorlatban a normavilágosság fogalmának értelmezése kapcsán kialakult szemléletbeli aszimmetriára a laikus-jogász opposícióban, az utóbbi javára (vö. 2.3.3 fejezet), kiemelve, hogy a normavilágosság koncepciója hangsúlyosan a kontinentális jogrendszerben van jelen, míg a PLM alapvetően az angolszász országok precedenciális jogrendszereinek talaján nőtte ki magát, amelynek következtében teljesen eltérő értelmezési tartománnyal is rendelkeznek.

A szerző állításával egyetértek abban a tekintetben, hogy a közérthetőségi törekvéseknek a jogi szövegek teljes spektruma nem lehet alanya. Hasonlóan irreális az a cél, hogy az érintett szövegeket bárki első olvasására, szakértői segítség nélkül megértse, feltéve ismét, hogy az elvárás az összes hivatali / jogi szövegre egységesen kívánjuk alkalmazni (vö. 2.4.2 fejezet). A jogi domén megfelelő szegmensének kiválasztása esetében azonban (akár egy erre a célra kialakított jogi metanyelv alkalmazása által) a lehetséges társadalmi előnyök messze felülműlják a befektetendő munkát. A kulcs a megfelelő szövegtípusok meghatározása, amelyek esetében az ilyen és ehhez hasonló érthetőbbé tétel nem pusztán relevanciával bír, de a szaknyelvi szempontok sérülése nélkül érvényre is juttatható.

## **5.2 A közérthetőség fogalmának alulspecifikáltsága**

Eltávolodva az elvi szinten elérendő céloktól, gyakorlati szempontból a legfontosabb áthidalandó probléma annak meghatározása, hogy egy szöveget pontosan kinek a szemszögéből tekintünk könnyen vagy nehezen érthetőnek. Ez a szempont szintén megjelenik a PLM-et és általában a közérthetőség keresését célzó kezdeményezések kritikai olvasataiban is (Assy 2011: 383).

Tekintettel a társadalom már említett heterogenitására, ezt valóban rendkívül problémás lehet pontosan meghatározni, nem feltétlenül lehetséges rá egyetlen, minden olvasóra jól

illeszkedő modellt alkotni. A gyakorlatban, például a közérthető nyelvhasználat már ismertetett ISO sztenderdjében ez a feladat a szöveg megalkotójára hárul. Ez logikus abból a szempontból, hogy legtöbb esetben ő ismerheti leginkább azt a célközönséget, akihez a szövegnek később szólania kell.

Egy másik megoldást azon „univerzális” jegyek megkeresése jelentheti, amelyekről a vonatkozó szakirodalom egységesen állítja, hogy általában véve a megértést képesek könnyíteni vagy nehezíteni. Ezek összevetésével, valamint az érthetőségre útmutatók, lexikai és szintaktikai jegyek halmazával korábban választ adó irányzatok (bővebben lásd: 3.2, 3.3 és 4.3.1 fejezetek) javaslatainak számbavételével véleményem szerint olyan nyelvi jegyeknek egy olyan halmaza határozható meg, amelyek figyelembevételére jó közelítést jelent a könnyebb olvashatóság mibenlétére vonatkozóan.

Ugyanakkor a módszer korlátjaként fontos kiemelni a jogi szövegek értelmezése kapcsán például a 2.3.1 és 2.3.2 fejezetben felvázolt komplexitást, amelyet pusztán a szövegek módosításával nem lehetséges feloldani.

A fentiek összhangban állnak Zódi (2018: 247-248) a témában tett megállapításaival is. A szerző a megértés hierarchikus jellege kapcsán kifejti, hogy az ennek folyamatában általánosan jelen levő vertikális tagolódás, mely szerint például a szójelentések elérhetlensége esetén (lexikai szint) eleve elérhetetlen lesz a mondatjelentés (szintaktikai – szemantikai szint) megkonstruálása, különösen fontos probléma a jogi nyelv(ezet) esetében. Ennyiben tehát igazolva látja a PLM által támasztott ajánlásokat, amelyek főleg stiláris, szintaktikai és lexikai javaslatokat tartalmaznak (vö. 3.3.4 fejezet). A módszer kritikája kapcsán a szerző is azt az aspektust emeli ki, amely szerint a jogi szövegek bonyolultságát a PLM kizárólag a bennük előforduló (olykor a köznyelv szempontjából atipikus) kifejezések alapján értékeli<sup>116</sup>.

A jogi szövegeknek a jog tudományából, és nem a szöveg szervezéséből adódó értelmezési nehézségeit Zódi szerint semmilyen nyelvi egyszerűsítés nem képes hatékonyan megoldani. Bár ennek vizsgálata további empirikus bizonyítást igényelne, ilyen, vagy amely az állítást cáfolná, nem áll rendelkezésre egyértelműen. Az azonban biztos, hogy a szöveg befogadója

---

<sup>116</sup> Példaként olyan mondatokat hoz, amely nem rendelkezik egyetlen olyan szerkezeti, lexikális jeggyel sem, amelyet hagyományosan az „érthetetlen” kategóriába sorolhatunk, ellenben értelmezésük (az általa végzett webes kutatás tanúságai alapján) mégsem triviális, pl.: „Aki valakiről más előtt a becsület esorbítására alkalmas tényről állít, híresztel, vagy ilyen tényre közvetlenül utaló kifejezést használ...” Btk. 226. § (1) bek., Zódi (2018: 249) 3. példa.

számára hasznos, és több helyzetben önmagában is elegendő lehet, ha a szövegszervezési, fogalmazási aspektusok megválasztása az értelmezést nem megnehezíti, hanem megkönnyíti.

Az ilyen („optimalizációs”) megközelítésnek véleményem szerint a legfontosabb gyakorlati előnye abban áll, hogy az értelmezést megnehezítő, de valóban pusztán nyelvi jelenségek zavaró hatásának megszüntetésével a tisztán jogi természetű komplexitás a szövegben nyilvánvalóbbá válik, megkönnyítve akár a valós jogi természetű probléma azonosítását is. Ilyen módon a nyelvi megfogalmazás bonyolultsága és a jogi jellegű probléma kétrétegű összetettsége szétválasztható, illetve előbbi kiküszöbölésével (vagy legalábbis a lehetőségekhez mért minimalizálásával) a kettő egymással interferáló, zavaró jellege is megelőzhető.

A 11) példa kirívó esete annak, amikor a szövegben hordozott jogi (normatív) tartalom önmagában nem különösebben összetett, a nyelvi forma ugyanakkor még a tapasztalt olvasó számára is különösen megnehezíti az értelmezést:

- 11) „Az úti okmány iránti kérelem előterjesztésekor a kérelmező a kérelmezett úti okmány adattartalmát képező személyes adatait, azok ellenőrzésére szolgáló okiratokat, lakcímét, illetve értesítési címét, továbbá arcképmását és saját kezű aláírását köteles az eljáró útlevélfeladó rendelkezésére bocsátani. Felhívjuk azonban a figyelmet arra, hogy az arcképmás, aláírás, és ujjnyomat a kérelem benyújtásakor kerül felvételezésre.”

A korábban a hazai PLM képviselői által adott alternatív megfogalmazás tükrében világosan kivehető, hogy a nyelvi tényező önmagában is felelős lehet az értelmezési nehézségekért:

- 12) „Új útlevel igényléséhez hozza magával
  1. Lakcímkártyáját
  2. Érvényes személyi igazolványát, kártya formájú jogosítványát, vagy útlevelét. Az útlevelhez fotót készítünk Önről, továbbá rögzítjük az aláírását és ujjlenyomatát.” (A 12) és 13) példa forrása: <https://vilagosbeszed.hu/vilagos-dij/>, 2016-os díjazottak – Az érthetetlen szövegek.)

Mindez jól mutatja, miért is nevezhetjük a jogi domén szövegeinek közérthető(bb)re alakítását egyféle optimalizálásnak, hiszen, noha ez a folyamat minden felmerülő értelmezési nehézséget nem képes önmagában megoldani, mégis „helyzetbe hozza” a



szöveg befogadóját a tényleges (doménfüggő) probléma meglátásához. Jó példa lehet erre a Nemzeti Adó- és Vámhivatal tájékoztató anyaga az adószám és adóazonosító jel kiváltásáról<sup>117</sup>. Egy kiragadott mondat a tájékoztatóból:

13) „Kinek kell adószámot kiváltania?

Adószámot kell kiváltania

- a gazdasági tevékenységet folytató magánszemélynek,
- az egyéni vállalkozóknak,
- a cégeknek,
- a bíróság által nyilvántartott jogi személyeknek, mint például az alapítványok, ha gazdasági tevékenységet folytatnak,
- valamint az úgynevezett törzskönyvi jogi személyeknek, mint például a költségvetési szervek.

A NAV-hoz az adóköteles tevékenység megkezdése előtt kell bejelentkezni.<sup>118</sup>”

Összehasonlításban a 11), 12) -vel, a 13) egyértelműen a közérthető példához konvergál szövegezésében. Ennek köszönhetően (legalábbis lexikai szinten) azonnal világos, ha az olvasó egy számára ismeretlen fogalommal vagy koncepcióval találkozik. Másképpen szólva a figyelmet nem osztja meg a mondat (morfo-)szintaktikai relációinak értelmezése, vagy éppen a túlbonyolított, esetenként archaikus szóhasználat.

Mindemellett persze fontos megjegyezni, hogy a hasonló tájékoztatók sem lehetnek homogének a megfogalmazás egyszerűségét tekintve; egy cégalapítást például nem lehet a jogban ehhez kapcsolódó, gyakran a laikus számára nehezen átlátható fogalom és viszonyrendszerekre való utalás nélkül elképzelni. A 14) példa az egyéni vállalkozók szja- és járulékkötelezettsége címen elérhető tájékoztatóból származik:

14) A személyi jövedelemadó rendszerében egyéni vállalkozó:

- 8) az egyéni vállalkozóról és az egyéni cégről szóló törvény szerinti egyéni vállalkozói nyilvántartásban szereplő magánszemély, az ott rögzített

---

<sup>117</sup> Elérhető: [https://www.nav.gov.hu/nav/inf\\_fuz/2020](https://www.nav.gov.hu/nav/inf_fuz/2020)

<sup>118</sup> Kiemelés a szerzőtől; a fontosabb jogi fogalmak a szövegben.

tevékenysége(i) tekintetében azzal, hogy nem minősül egyéni vállalkozónak az a magánszemély, aki

aa) az ingatlan-bérbeadási,

ab) a magán szálláshely-szolgáltatási tevékenységéből származó bevételére az önálló tevékenységből származó jövedelemre, vagy a tételes átalányadózásra vonatkozó rendelkezések alkalmazását választja, kizárólag a választott rendelkezések szerinti adózási mód alapjául szolgáló bevételei tekintetében;...”

Ebben az esetben például a szöveg teljes egészében az 1995. évi CXVII. törvény (a továbbiakban: Sza tv.) 3§ 17. pontjának idézése. Erre alapvetően a szaknyelvi kritériumok teljesülése végett kerülhet sor, hiszen amennyiben a törvényi meghatározásból akár egyetlen szót is elhagy a tájékoztató, akkor (tekintettel a kiragadott bekezdés definíciós jellegére) az eredmény pontatlanná válik. Ugyanakkor érdekes szempont az is, hogy míg a jogalkotó például egy adószám igénylése esetén nem számíthat arra, hogy a folyamat menedzseléséhez az ügyfélnek szakértői segítség áll rendelkezésére, egyéni vállalkozók adózása esetében a legtöbbször könyvelő, vagy egyéb szakértő közreműködésére is sor kerül. Felmerülhet tehát a(z elvi) kérdés, hogy ilyen esetben valóban elvárható-e, hogy ne a szakembert tekintsük a szöveg elsődleges (egyben leginkább plauzibilis) címzettjének? Véleményem szerint elvárható lenne, azonban ez a fajta szemlélet jelenleg nem tükröződik világosan a jogalkotásban (vö. 2. fejezet bevezető).

Az összevetést ezen felül valamelyest torzítja a 14) kapcsán említett „idézet-jelleg” is, hiszen ebben az esetben valójában nem is az ügyfeleknek szánt tájékoztató szövegről, hanem törvényszövegről van szó (még ha egy tájékoztató anyag részét képezi is a jelen esetben), amely két kategória természetesen nem eshet azonos elbírálás alá.

### **5.3 A pragmatikai kontextus figyelmen kívül hagyása**

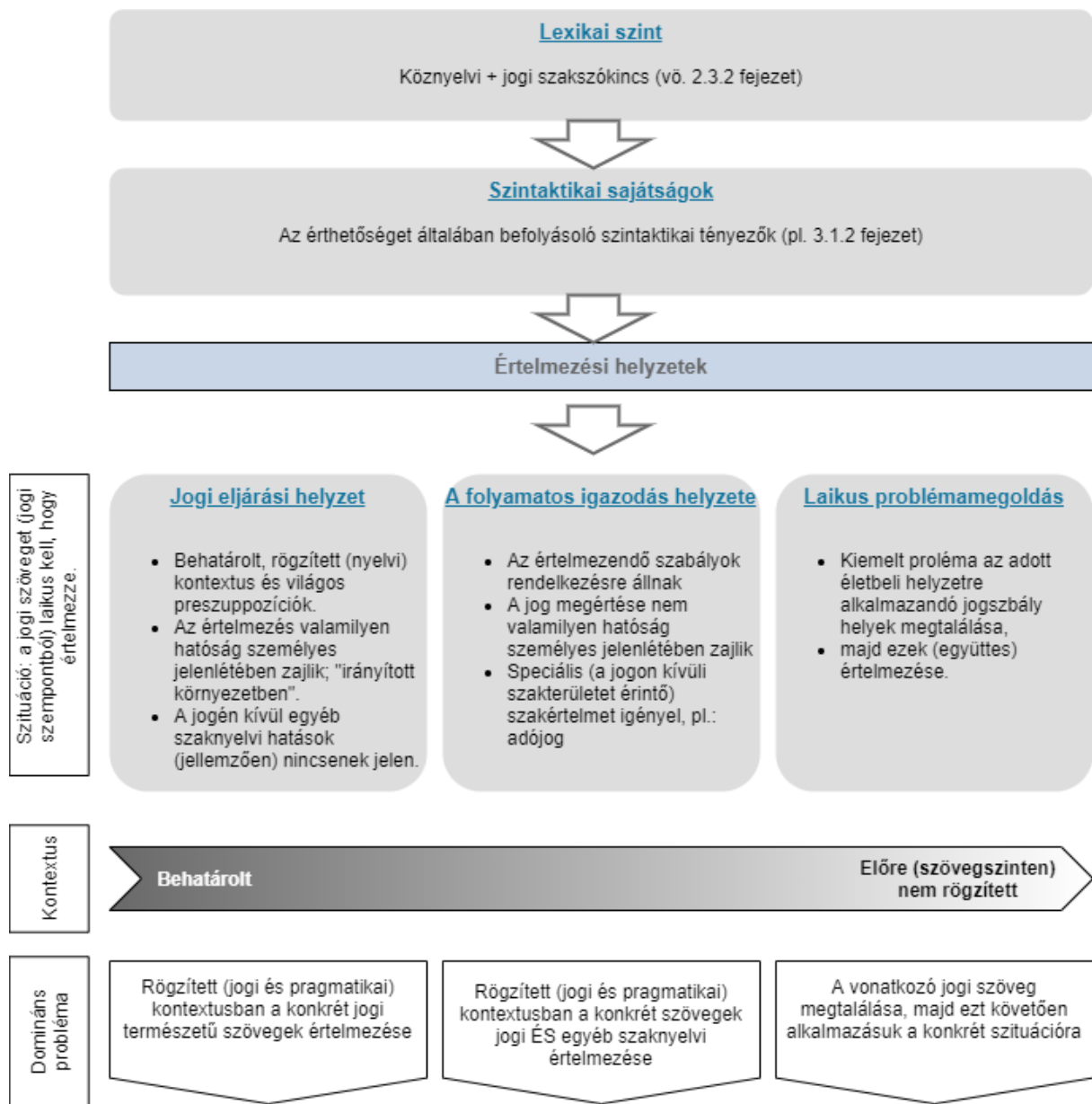
Az 5.2 fejezetben már részben utaltam rá, hogy a közérthetőségi törekvéseket leggyakrabban érő kritikák között fontos szerep jut azoknak is, amelyek az ilyen ajánlásokat azok idealizált, egyszerűsített jellege miatt bírálják, ezzel pedig nagyban átfed a jogi szövegek pragmatikai kontextusának figyelmen kívül hagyása is.

Zódi (2018: 253-254) a probléma egyik legfőbb okaként azt a jelenséget azonosítja, amely szerint a PLM-mel összefüggő kutatások (egyben az első, hazai viszonylatban úttörő

jelentőségű jog és nyelv vizsgálat is) egy pragmatikai értelemben viszonylag homogén helyzetcsoportot kutatott, amelyet a szerző „eljárási helyzetnek” nevez. Ez alatt olyan szituációkat ért, amikor az „egyszerű laikus” találkozik a „jogi gépezettel”, mint amilyen helyzet például a bírósági tárgyalások vagy az azt megelőző tanúkihallgatások esete a büntetőperек során (vö. Vinnai 2010, 2011). Ebben a helyzetben a fő hangsúly Zódi értelmezésében (2018: 255) a szituációban lecsapódó alá- és fölérendeltségi viszonyokon, valamint a szűk, behatárolt preszuppozíció-készleten van, amelyet nyelvi értelemben a kontextus képvisel. Ezekben az esetekben a kiinduló helyzetet alapvetően ismertnek és érthetőnek tekinti, hiszen például a vádlott már tudatában van, hogy ellene büntetőeljárás folyik, így a kontextus világosnak tekinthető. Fontos továbbá, hogy ebben a helyzetben a fő hangsúly nem az alkalmazandó jogszabályok megtalálásán van, ez ugyanis szintén előre adott.

A központi kérdés sokkal inkább a konkrét szöveg értelmezése a joghoz nem értő laikus által, vagyis mindazon jogértelmezési mechanizmusok (vö. 2.3.1 és 2.3.2 fejezet), amelyek a norma összekapcsolását hivatottak elvégezni a konkrét, valós életbeli szituációval, már előzetesen lezajlottak, ilyen formán tehát az értelmezés körülményei is leegyszerűsödtek, a laikus szempontjából optimalizáltak.

Zódi (2018: 254-257) további két tipikus értelmezési helyzetet említ, amelyek közül az elsőt a „folyamatos igazodás helyzete”-ként, a másikat pedig a „laikus problémamegoldás” szituációjaként írja le. Az első típusnál a prototipikus helyzet az, amikor a jogszabály nem csak a jogi szaknyelv megértési nehézségeit támasztja a vele kapcsolatba kerülő olvasó elé, hanem még valamilyen egyéb szakterület ismeretanyagára is épít. Ez olyankor fordul elő, ha egyébként az általa szabályozott kérdéskör ennek a másik szakterületnek a hatáskörébe tartozik, mint például az adójogi jogszabályok esetében. A jogértelmezésnek itt is adott egy (nem közvetlenül megjelenő) autentikus forrása, a problémát ezért valójában a jogitól eltérő szakterület ismeretanyagának felfejtése okozza. A másik típusnál fontos eltérés az eddigi kettőhöz képest, hogy a jogértelmezésnek nincs jelen ilyen autentikus forrása. Ez az az eset, amikor a probléma megfogalmazása is a laikustól ered, éppen ezért az ahhoz vezető út megtalálása is egy önálló folyamatként írható le. Itt a cél a probléma megoldásában irányadó jogforrások azonosításán túl az alkalmazás / applikáció mikéntjének megtalálása, vagyis a „mit kell tennem a személyi igazolványom elvesztése esetén” típusú kérdéstől az illetékes hivatalig, vonatkozó űrlapig vezető út felfejtése.



6. ábra: Különböző kontextusok a jogi szövegek értelmezése kapcsán.

Az ilyen módon szétválasztott helyzetcsoportok példáján jól tetten érhető, hogy a joggal való találkozás egyes kontextusaiban, habár a megoldandó részproblémák azonosak, a főhangsúly mégis dominánsan eltér. Minden esetben szükséges folyamat tehát:

- az adott, konkrét szituációra irányadó rendelkezések megtalálása, amelyhez - tekintettel arra, hogy a jog saját belső logikája mentén, és nem a hétköznapi helyzetek kategóriái szerint szerveződik – szükségesek a jogértelmezés módszerei is,

- a megtalált rendelkezések értelmezése (ez állhat legközelebb a puszta szövegértéshez, ugyanakkor fontos megjegyezni, hogy a legkevésbé sem azonos azzal),
- esetlegesen az egyéb szakterületek ismeretanyagának értelmezése.

#### **5.4 A közérthetőség helye a jogi domén szövegeiben**

Az egyes helyzetcsoportok alapján tagolt (pragmatikai) kontextusok alapján tehát jól kivehető, hogy a szövegen végzett nyelvi változtatások hatása nem lesz homogén minden helyzetben, mint ahogyan maga a jogi domén sem egységes.

A MJK kapcsán (3.5.2 fejezet) már láthattuk, hogy a jogi szövegeket típusokba sorolhatjuk például azok származási helye alapján;

- kódexjog / jogszabályok,
- bírósági és törvényszéki ítéletek szövegei,
- metanyelvi / magyarázó szövegek,
- jogi fórumok szövegei,
- bírósági tárgyalásokon és rendőrségi kihallgatásokon készített átiratok (beszélt nyelvi szövegek).

Tóth (2016: 47-48), ezzel szemben inkább a használat körülményeire tekintettel osztályozza a jogi szaknyelv rétegeit. Hozzá hasonlóan Vinnai (2011: 90-91) is lényegileg hasonló csoportosítást használ, habár esetében a szervezőelvet a szaknyelvi szempontok szigorú érvényesülése adja; esetében a 1.-től 4.-ig haladva például a szakszó jelentés-behatároltsága egyre lazábbá, a stílus pedig egyre kötetlenebbé válik.

1. „A jogi norma nyelve;
2. A jogtudomány, a jogi szakirodalom nyelve, amely bizonyos fokig követi a jogszabályok fogalmi rendszerét, de éppen annak hibáit, következetlenségeit dogmatikai alapozással kiegészíti, korrigálja, rendszerbe foglalja;
3. A jogalkalmazók nyelve, amelyet gyakran eljárási nyelvnek is neveznek, amikor konkrét formába öntik a jogszabályi célokat, jogelveket a szóbeliség érdekében; valamint
4. A jogász és a laikus közti kommunikáció nyelve, amelyben leginkább azt a szakadékot kell áthidalni, amely a köznyelv és a szaknyelv között található (például

az ügyvéd védencének elmagyarázza, mit jelent a fegyveregyenlőség elve vagy a hallgatás joga az éppen folyó eljárásban, mit tehet az ügyfél”.

A fenti osztályozás alapján véleményem szerint a pusztán nyelvi tényezőket figyelembe vevő közérthetőségi javaslatok, habár minden stílusrétegben segíthetik a nyelvileg jól követhető és transzparens szövegek megalkotását, mégis legnagyobb szerepük egyértelműen a Zódi (2018) által eljárási helyzetnek nevezett, Tóth (2016) és Vinnai (2011) alapján pedig a jogász és laikus közötti kommunikáció stílusrétegeként megnevezett típusba tartozó szövegek esetén lehet.

Ez alapján a létrehozni kívánt szoftver tipikus használati esete lehetnek például olyan szövegeken végzett szakértői / fogalmazói munka támogatása, amelynek során hivatali tájékoztatók, bírósági ítéletek indoklásai, vagy éppen űrlapok kitöltési útmutatóinak elkészítése történik.

## **5.5 Gyakorlati lehetőségek – elméleti összefoglalás**

Az eddigieket egybevetve látható, hogy a közérthető fogalmazás mibenlétét, helyét a jogi domén szövegein belül, valamint a szükséges megszorításokat számos tudományterület kísérelte már meg magyarázni legyen szó akár a nyelvtudomány egyes ágairól, akár a jogtudomány részterületeiről.

A közérthető kommunikáció morális előnyei kapcsán már több ponton is említésre került a tisztességes eljáráshoz való jog, valamint a joghoz való hozzáférés elősegítése (4.3.2 fejezet), illetve az az érv is, mely szerint a közérthető hivatalos szövegek segítenek a jogbiztonság megőrzésében, továbbá növelik az adminisztratív folyamatokba vetett bizalmat és jótékonyan hatnak a demokratikus érdekérvényesítő képességre is (4.2.2.1 fejezet). Emellett praktikussági szempontokat figyelembe véve nem elhanyagolható az sem, hogy a kommunikáció hatékonyságának növelése / fenntartása például vállalati környezetben hozzájárul a reklamációk, állami szervek esetén a panaszok számának csökkentéséhez, ezáltal jelentős pénzügyi és humán erőforrás spórolható meg, amelyet egyéb esetben ezek kezelésére kellene fordítani (3.3 fejezet).

Az ellenérvek (~a megoldást nehezítő faktorok) közül az 5. fejezetben mindeddig azon körülményeket tárgyaltam részletesen, amelyek miatt a jogi / hivatalos szövegek nyelvi átdolgozásától nem várhatunk egységesen jelentős eredményeket minden szövegtípus esetén

(5.1 – 5.4 fejezet). A jogi domén rétegzettségét (pl. 5.4 fejezet), valamint a közérthetőségi törekvések eddigi eredményeit áttekintve (4. fejezet) az is nyilvánvalóvá vált, hogy a dolgozat célkitűzésében vállaltak teljesíthetősége nagyban függ a jogrendszerétől, amelyben a közérthetőségi törekvés megfogalmazódik, illetve az ilyen igények társadalmi beágyazottságától, amely az adott országot e tekintetben jellemzi.

Jó példa erre, hogy az egyesült államokban törvény garantálja a hivatalos szervek kommunikációjának ilyen irányú felülvizsgálatát, a skandináv államokban sok esetben a konkrét törvények szövege is törekszik az egyszerű és világos megfogalmazásra, hazánkban pedig az ilyen irányú kezdeményezések mindösszesen szórványosak a jogalkotó részéről. Ennek ellenére a NAV Közérthetőségi Munkacsoportjának tevékenysége rávilágít talán a legjellemzőbb területre, amely esetében Magyarországon is sikerre vihető lehet egy közérthetőségi szempontú ellenőrzési folyamat bevezetése a hivatalok tájékoztató anyagai, tágabb értelemben a funkcionális szövegek kapcsán. A fentiek egyben igazolják azt a kiinduló hipotézist (ii.), amely szerint a jogi nyelvnek létezik olyan rétege, amelyben a közérthetőség, mint célkitűzés a gyakorlatban is érvényesíthető, mégpedig a funkcionális szövegek csoportja.

A dolgozat célkitűzésében vállaltak szerint szükséges egyrésztől konkrétan definiálni, hogy mely nyelvi jegyek tekinthetők valamiképpen a közérthető fogalmazás indikátorainak. Erre vonatkozóan konkrét javaslatokat ismertetett a 3.1.3 fejezet a pszicholingvisztika, a 3.2.2 az olvashatósági formulák készítői, a 3.3.4 a PLM képviselői, illetve a 4.3.1 a Stíluskönyv íróinak nézőpontjából. Az egyes területek álláspontjai integrálhatók egy egységes „kézzel írt” szabályrendszerbe, amely egy helyesírásellenőrző működéséhez hasonlóan konkrét javaslatokkal lenne képes támogatni egy-egy dokumentum fogalmazóját. Hasonló alkalmazások léteznek már például az angol nyelv esetén (pl. Hemingway Editor<sup>119</sup>), amelyek az érthetőséget javítani kívánó stiláris javaslatokat is képesek tenni a felhasználó számára.

A szakirodalomban feltárt eddigi eredmények ennek megfelelően igazolták az a kiinduló hipotézist (i.), amely szerint egy szöveg érthetőségének mértéke (bizonyos korlátokkal) kapcsolatba hozható benne előforduló konkrét nyelvi jellegzetességekkel. Ezen jellegzetességek számbavételére minden fent említett kutatási irányzat kísérletet is tesz.

---

<sup>119</sup> <https://hemingwayapp.com/>

Az ilyen módon definiált szabályok azonban nem aknázzák ki az NLP területén a gépi tanlás elterjedése óta elért eredményeket. A szabályalapú rendszerek jellemzője, hogy az a logika, amely mentén a kód szerveződik, a kezdetektől fogva adott, éppen ezért a továbbiakban kevésbé rugalmas is egyben. Voltaképpen az ilyen rendszerek lényege, hogy a megalkotásukkor rendelkezésre álló szakértői tudást kódolják afféle „ha-akkor” szabályok formájában. Gépi tanulás esetében a döntéshez szükséges szabályokat nem ember, hanem egy megfelelően nagy adathalmaz birtokában tanított algoritmus fogja létrehozni. Itt voltképpen teljesen automatikusan történik meg az adatelemzés, valamint az egyes adatpontok közötti összefüggések feltárása. A cél ez esetben a kapcsolat megteremtése a rendelkezésre álló adatokból levonható következtetések, illetve a jövőbeli kimenetek között.

Az ember készítette szabályhalmaz előnye, hogy alkalmazásával kikényszeríthetünk (például szövegek esetében) olyan kimenetet, amely mindenképpen rendelkezik egy-egy adott tulajdonsággal. Ha például a célunk az, hogy csak bizonyos karakterszám alatti mondatokat tekintsünk megfelelőnek, egyszerű szabályt írhatunk rá, amely jelzi a felhasználónak, amennyiben ezt a korlátot túllépi.

A gépi tanuló algoritmusok ezzel szemben felépítésüknek köszönhetően olyan összefüggések megtalálására is képesek lehetnek, amely az ember számára nem, vagy csak kevésbé nyilvánvaló.

A fentiekből következően a kézi szabályok előnye a könnyű interpretálhatóság, hiszen már meglevő tudást képeznek le algoritmikus alapokra. Ezzel szemben a gépi tanuló algoritmusok (és különösen igaz ez a neuronháló alapú megoldásokra) gyakran feketedoboz (black box) jelleggel működnek. Ez a gyakorlatban azt jelenti, hogy az adat és kimenet közötti összefüggés csak nehezen, vagy egyáltalán nem feltárható.

A dolgozat céljainak a két megközelítés ötvözése felel meg leginkább, amennyiben a szakirodalomban megtalálható, algoritmikusan megfogható és a magyar nyelv esetében releváns szabályok manuálisan jönnek létre, és ezen felül egy gépi tanuló modell is segíti a közérthetőségi javaslattételt. Utóbbi szerepe leginkább a szövegek előszűrésében indokolható, amennyiben a modell kiválogatja azon szövegrészeket, amelyek a tanító adatok alapján problémásak lehetnek, majd a külön definiált szabályrendszer már az eredeti szövegnek ezen a szűrt részhalmazán fut le.



A következő fejezetek ennek megfelelően ismertetik a rendszer tervezett felépítését, a fontosabb elvégzett gépi tanítási kísérleteket és azok eredményeit, ezt követően bemutatják a kézi szabályrendszer elemeit, végül pedig azok fontosabb implementációs részleteit.

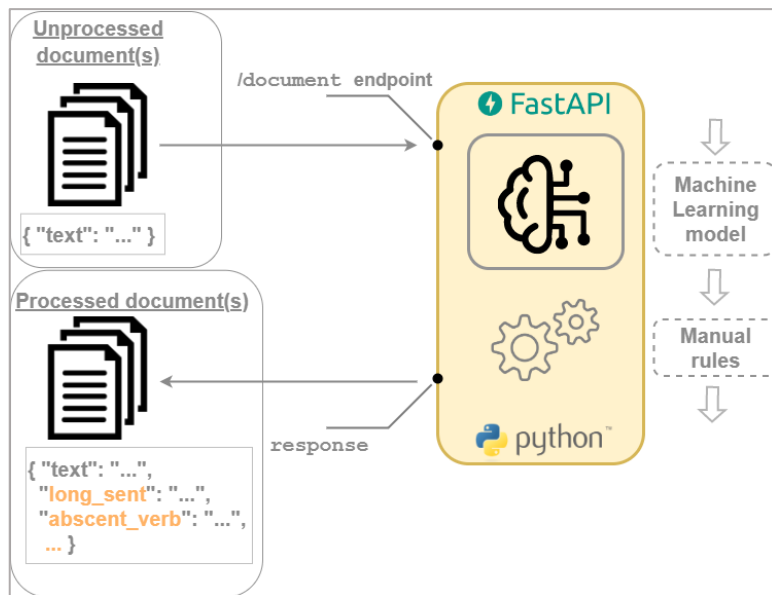
## 6. Gépi tanítási kísérletek előkészítése

Az eddigiekből, főként a 3. fejezetben vázoltak alapján világosan kirajzolódik egyfajta kép arról, hogy egy közérthető fogalmazást támogató szoftver pontosan milyen szabályok megfogalmazásán alapulhat. Ezzel kapcsolatban azonban még hátralévő feladat, hogy a szakirodalomban feltárt szabályszerűségeket rendszerezzük az implementálhatóság szempontjából is. Ez voltaképpen azt jelenti, hogy a lehetőségek közül ki kell választanunk azokat, amelyek a nyelvtechnológia mai eszköztárával lefordíthatók „a gép számára is érthető” formára. Ezt a feladatot a 12. fejezet végzi el.

Az 5.5 fejezetben már említetteknek megfelelően ezen felül szükséges valamilyen típusú gépi tanult modell építése is. Egy ilyen modellel egyrészt hatékonyan lehet olyan fajta tudást beépíteni a rendszerbe, amely csak a rendelkezésre álló adatokban jelenlevő, de az emberi szem számára nem hozzáférhető mintázatok alapján sajátítható el. Másrészt bizonyos modell típusok erősségeként leginkább a nyelvre jellemző szemantikai ismeretek hatékony kódolását szokás megemlíteni. Ahogyan a 12. fejezetben látható lesz, a már ismertett közérthetőségi szabályok főként a szintaxis és a szemantika szintjén hordoznak konkrétan megvalósítható javaslatokat, ezért a gépi tanulás a szemantikai tulajdonságok alapján történő osztályozás szempontjából is kulcsfontosságú lehet. Egy ilyen osztályozási feladat leginkább úgy képzelhető el, hogy a modell feladata a szövegek / bekezdések / mondatok szétválogatása közérthető és javításra szoruló egységekre.

Az elkészített szoftver működését ez alapján tehát két logikai komponens jellemezheti. A szövegek előzetes feldolgozásáért egy gépi tanult modell felelős, amely eldönti, hogy az adott szövegegységen szükséges-e további szabályok alkalmazása. Amennyiben a döntés igen, úgy egy manuális szabályrendszer fut le, amely eldönti, hogy pontosan mely javaslatok figyelembevétele lehet szükséges a kérdéses szövegegység közérthetőbbé tételéhez.

Egy ilyen működés előnye lehet, hogy hatékonyan képes ötvözni az ember számára is interpretálható szabályokat azokkal a gépi tanuló modell által elsajátítottakkal.



7. ábra: A megvalósítandó alkalmazás sematikus felépítése.

A 7. ábra az imént megfogalmazott koncepció szerinti sematikus működést illusztrálja. Az alkalmazás egy teljes dokumentumot kap, amelyen elsőként gépi tanult modellel kiválogatja a problémás szövegrészeket, majd pedig ezekhez a szabályrendszer alapján bizonyos javaslatokat is társít. Az ábrán szereplő legtöbb technikai részletet a 13. fejezet ismerteti.

A modell kiválasztása kapcsán több tényezőt is számításba kell venni. Talán az első és legfontosabb, hogy magyar nyelvre egyelőre semmilyen hasonló sem létezik. Éppen ezért, lévén, hogy szakirodalmi tapasztalatok nem állnak rendelkezésre, érdemes több lehetőséget is számításba venni, mielőtt az alkalmazásba véglegesen bekerülőt ezek közül kiválasztanánk. Ezek a lehetőségek magától értetődően régebbi és újabb megoldások is lehetnek, ami több érdekes következtetés levonását is lehetővé teszi.

Egyrésztől figyelemmel követhetjük, hogy az egyre újabb technikai megoldások egyre jobban teljesítenek-e a szövegek közérthetőségi osztályozási feladatán, és ha igen, milyen mértékben? Az előzetes várakozás természetesen az, hogy minden újabb megoldás teljesítményben szignifikánsan megelőzi majd az őt megelőzőt. Emellett nem szabad azonban elfelejtenünk, hogy a jelen írás nem titkoltan egyfajta tervrajzot is kíván nyújtani a közérthetőség automatizálása iránt érdeklődőknek. Ezzel kapcsolatban fontos tudni, hogy a különböző modellek használata erősen eltérő felkészültséget követel meg a fejlesztést végzőtől. Egy hagyományos modell (például Szupport Vektor Gép) futtatása ma már alig

pár sor kóddal megoldható egy szinte tetszőleges számítógépen. Egy BERT-modell finomhangolása valamilyen feladatra azonban nem pusztán összetett programozási feladat, de jelentős hardveres kapacitást (főleg GPU-t) is igényel. Az egyes modell típusok bemutatására részletesen a vonatkozó (7-10) fejezetekben kerül sor.

Az itt bemutatott modellek kipróbálása tehát nem pusztán a legjobb megoldás keresésére irányul, de egyben betekintést is enged abba, hogy a nagyobb informatikai szaktudást és infrastruktúrát igénylő modellek valóban annyival jobban teljesítenek-e, amely megéri a befektetett plusz erőforrásokat.

Mindemellett a gépi tanulási projektek, kísérletek során szokás egy „baseline” megoldás kijelölése is. Ez általában egy olyan modellt jelent, amelytől előzetesen gyengébb teljesítmény várható, mint a mellette kipróbált egyéb megoldásoktól. Egy ilyen modell eredményei egyfajta természetes referenciát jelentenek, amelynek tükrében a további megoldások teljesítménye könnyebben kontextusba helyezhető és értékelhető. Mivel magyar nyelven (ismét csak) egyetlen közérthetőségi osztályozást végző modell sem létezik a dolgozat írásának pillanatában, így ez a fajta egymáshoz viszonyított kiértékelés az egyes modellekre vonatkozóan máshogyan nem is lenne objektíven elvégezhető.

Ennek megfelelően tehát a kipróbált gépi tanuló algoritmusokat, a velük elért eredményeket és azok értékelését a már említett 7–10 fejezetek ismertetik, az alkalmazásba kerülő szabályrendszert a 12. fejezet, annak konkrét implementációs részleteit pedig a 13. fejezet mutatja majd be részletesen.

A jelen fejezetben a gépi tanításhoz felhasznált korpusz részletezése, a szükséges előkészületi lépések, valamint az adathalmaz néhány alapvető jellemzőjének ismertetésére kerül sor.

## **6.1 Felügyelt és felügyelet nélküli gépi tanítás**

A gépi tanulási módszereket a gyakorlatban két nagy csoportra szokás bontani, amelyek közül az elsőbe a felügyelt- (supervised), a másodikba pedig a felügyelet nélküli (unsupervised) gépi tanuló algoritmusokat soroljuk. Előbbiek legfontosabb jellemzője, hogy működésükhöz szükség van valamilyen fajta előre címkézett tanítóadatra, míg utóbbiak esetében az osztályozás ezek hiányában megy végbe. (Jordan – Mitchell 2015)

A tanítóadatok voltaképpen bármilyen digitális formátumban elérhető adatállomány elemeiből kikerülhetnek a konkrét feladatnak megfelelően. Ha például a cél beszédfelismerés vagy a kulcsszavak felismerése hangzó szövegben (spoken term detection), akkor az inputot alapesetben hangfelvételek adják. Mindkettő esetében az alkalmazott algoritmus feladata a leirat és a hanganyag közötti megfeleltetés megtalálása (Gosztolya – Tóth 2011). Digitális képfeldolgozás esetén a bemenetet képek jelentik (pontosabban az egyes pixelekhez tartozó intenzitás mátrix), a cél pedig ezek osztályozása például az ábrázolt tartalom szerint (Pak – Kim 2017).

Természetesnyelvi szövegek osztályozása során törekedhetünk például szövegek osztályozására emóciótartalom alapján (Ring et al. 2022), vagy az egyes tokenek szófaji címkézésére (Chiche – Yitagesu 2022). Ilyen esetben az algoritmus természetesnyelvi szövegek adott részleteit kapja tanítóadatként a megfelelő címkékkel ellátva.

A nem neurális háló alapú megoldások közül gyakran alkalmazzák felügyelt gépi tanítási problémák megoldására például a Naive Bayes modellt (NB), a Support Vector Machine (SVM) vagy a Logistic Regression (LR) algoritmust, amelyek alkalmazása kiterjed például a szentimentelemzésre (pl. Chauhan 2017) vagy éppen a spam-detektálásra, amely során egy adott üzenetről kell eldönteni (jellemzően emailek esetén), hogy azok tartalma kéretlenül kategorizálható-e (Sun et. at. 2020).

A felügyelet nélküli esetben előre címkézett tanítóadat nem áll rendelkezésre, a gépi tanítást végző szakembernek „mindössze” a kimenetként elvárt halmazok számát kell definiálnia, majd az algoritmus maga végzi el a csoportosítást a rendelkezésre álló jellemzők alapján. Ilyen tipikus feladat például, amikor a cél nagyobb szövegegységek osztályozása a bennük előforduló témák alapján (topik-modellezés). Ebben az esetben talán a leginkább széleskörben elterjedt megoldásnak a K-közép klaszterezés (k-mean clustering), valamint az LDA (Latent Dirichlet Allocation) topik modellezés tekinthető (Sebők et al. 2021).

## **6.2 Korpusz közérthetőségi osztályozáshoz**

Az ismertetett két nagyobb osztály közül az alkalmazás elkészítéséhez szükséges gépi tanult modell előállításához felügyelt gépi tanuló algoritmusok használatát tűztem ki célul. Ennek háttérében az állt, hogy a kutatás során sikerült olyan korpuszra szert tenni, amely alapján

(annak előfeldolgozása után) lehetséges volt egy megfelelő tanítóhalmaz képzése a rendelkezésre álló adatokból.

### 6.2.1 Nemzeti Adó- és Vámhivatal – Közérthetőségi Program II.

A Nemzeti Adó- és Vámhivatal Kommunikációs Főosztályának Médiaosztálya a 4.3.3 fejezetben ismertetetteknek megfelelően immár négy éve működtet dedikált Közérthetőségi Programot, amelynek keretében három szakértő a hivatal valamennyi kommunikációs anyagát ellenőrzi, és azt (a közérthetőséget, valamint többek közt a PLM által propagált elveket figyelembe vevő) javaslatokkal korrektúrázza.



8. ábra: A Közérthetőségi Programiroda tevékenységének helye az ügyfélértékelő anyagok készítésében.

A folyamat része még ezen felül, hogy az átvizsgált dokumentumok egy végső szakmai ellenőrzésen esnek át, amely lezárulta után hozzák nyilvánosságra a hivatal kommunikációs felületein (a folyamatot a 8. ábra szemléleti). Ezeken a változatokon ismételt ellenőrzés a Közérthetőségi Program munkatársai részéről már nem történik. A hivatal munkatársai a jelen kutatáshoz rendelkezésre bocsátották a teljes 2021-2022-ben korrektúrázott anyagot, amelyben követhetők az újrafogalmazás előtti és a már átírt anyagok is.

Az itt említett dokumentumok végső, szakmai szempontú ellenőrzésen átesett változatai elérhetőek a NAV honlapján, az „Információs füzetek” menüpont alatt<sup>120</sup>. Fontos tehát, hogy ezek nem esnek egybe a jelen kutatás által felhasznált verziókkal. Az eredeti dokumentumok bármilyen közérthetőségi ellenőrzés előtti állapotot jelentenek (amelyek szükségszerűen

<sup>120</sup> [https://nav.gov.hu/ugyfeliranytu/nezzen-utana/inf\\_fuz](https://nav.gov.hu/ugyfeliranytu/nezzen-utana/inf_fuz) (Elérés: 2024. 01. 29.)

nem nyilvánosak), az átfogalmazott verziók pedig a 8. ábra tanúsága alapján nem biztos, hogy teljesen egybeesnek a korpuszban szereplő változatokkal.

Ez értékes betekintést enged tehát a szövegek olyan változataiba, amelyek a fent leírt együttműködés nélkül nem lettek volna hozzáférhetők. A közreműködésért és a segítségért ezúton is köszönettel tartozom a NAV munkatársainak.

### **6.2.2 Bemeneti fájlok**

A dokumentumokból minden esetben rendelkezésre állt tehát két változat; egy eredeti és egy szakértők által korrektúrázott verzió. Előbbire a továbbiakban eredeti -ként, míg utóbbira átfogalmazott-ként utalok. Mivel a jelen kutatás nagyban kapcsolódik a PLM által propagált elvekhez, így érdemes megemlíteni, hogy a szakértőkkel folytatott beszélgetés, valamint az általuk a NAV munkatársai számára készített segédletek, belső kommunikációs anyagok alapján az átfogalmazáskor alkalmazott elvek túlnyomó többsége metszetet képez a PLM elvárásaival (pl. az olvasó szempontjainak középpontba helyezése, a funkcióigék kerülése, jogszabályhivatkozások kiszervezése a főszövegen kívülre stb.).

A fentieket alapul véve a keletkezett korpusz ideális választásnak tűnt annak vizsgálatára, hogy egy olyan absztrakt és nehezen körülhatárolható fogalom, mint a "közérthetőség" megragadható-e bármilyen szinten korpusznyelvészeti és gépi tanulási módszerek alkalmazásával is?

A teljes korpusz több mint 600 .doc / .docx kiterjesztésű fájlt tartalmazott. A fájlok három csoportban álltak rendelkezésre, amelyek az eredeti, az átdolgozott és egy technikai változatot tartalmaztak. Ez utóbbi szerepe az volt, hogy benne a módosítások (a MS Word korrektúra funkciójának köszönhetően) nyomon követhetők legyenek.

Az eredeti korpuszt első lépésként a következők szerint osztottam részekre:

- Az eredeti és átfogalmazott változatok összetartozását a fájlok elnevezési konvenciója kódolta; a közös, korpuszban egyedi prefix után "A" jelölte az eredeti, "B" pedig az átfogalmazott szöveget (pl.: A44A és A44B),
- a fájlokat könnyebben feldolgozható, .txt állományokká konvertáltam,
- a dokumentumpárokat ezt követően mondatokra szegmentáltam,

- az így kapott mondatok közül eltávolítottam azokat, amelyek a dokumentumpár mindkét tagjában azonosan fellelhetők voltak.

A megmaradt mondatok a szülő fájlok fájlneveinek posztfixe ("A" vagy "B") alapján kerültek besorolásra vagy az eredeti, vagy az átfogalmazott szövegek részkorpuszába.

### 6.3 Adatelőkészítés

Ahhoz, hogy a gépi tanuló algoritmusok számára szükséges jellemzővektor előállhasson, a kapott természetesnyelvi szövegeknek tehát számos feldolgozási lépésen kell átesnie, amelyek részleteit a következő fejezetek ismertetik.

#### 6.3.1 Mondatszegmentálás

A fenti lépések közül különös jelentősége volt a rendelkezésre álló szövegek mondatszegmentálásnak. Tekintettel arra, hogy az osztályozás alapegységeként a mondatszintet határoztam meg, ennek a lépésnek a megbízható kimenetele alapfeltétel minden későbbi folyamat kapcsán és különös figyelmet igényelt.

A NAV ügyféltájékoztató szövegei – csakúgy, mint a jogi természetű szövegek általában – jelentős arányban tartalmaznak olyan hivatkozásokat, rövidítéseket és szövegtagolási elemeket, amelyek megnehezítik ezt a feladatot.

Magyar nyelvű szövegek mondatszegmentálására többféle megoldás is elérhető. Mindazonáltal érdemes megemlíteni, hogy a magyar nyelvű jogi szövegek esetében ez a feladat nem tekinthető tökéletesen megoldott problémának, ezért a rendelkezésre álló eszközökkel némi kísérletezésre volt szükség a legpontosabb megoldás megtalálása érdekében. Az általános célú nyelvfeldolgozó eszközök tekintetében a magyar nyelv meglehetősen előnyös helyzetben van, mivel számos automatikus elemző eszköz létezik mondatszegmentáló modullal. Ezek egyike a Szegedi Tudományegyetem HUN-REN-SZTE Mesterséges Intelligencia Kutatócsoportja által a 2010-es évek elején kifejlesztett magyarlanc nyelvfeldolgozó eszközkészlet (Zsibrita et al. 2013). Egy másik lehetőség például a HuSpaCy<sup>121</sup> csomag (Orosz et al. 2022), amely jelenleg a Mesterséges Intelligencia Nemzeti Laboratórium (MILAB) kutatási programja keretében áll fejlesztés alatt. Ezek a

---

<sup>121</sup> <https://github.com/huspacy/huspacy>



megközelítések főként gépi tanulási megoldásokon alapulnak. Egy másik módszer a mondatszegmentálás kezelésére heurisztikus algoritmus alkalmazása, amelyre jó példa lehet a `sentence-splitter`<sup>122</sup> python nyelvű függvénykönyvtár.

Az előzetes vizsgálatok azt mutatták, hogy bár a gépi tanuláson alapuló megoldások számítanak a legkorszerűbbnek a természetesnyelv-feldolgozási feladatok megoldása tekintetében, a gyakorlatban nem feltétlenül ezek alkalmazásától várhatjuk a leginkább megbízható eredményt az olyan speciális szövegtípusok esetén, mint a jogi szövegek.

Annak érdekében, hogy a szegmentálás várható pontosságát becsülni tudjam, készítettem egy külön erre a célra szolgáló, kisméretű korpuszt, amely 3, a jogi doménre jól jellemző szövegosztályból, valamint egy, az általános szövegeket reprezentáló szövegtípusból tartalmaz osztályonként 100-100 mondatot.

Az egyes osztályokba tartozó mondatok válogatásakor a fő szempont nem az volt, hogy azok a szegmentálás során esetlegesen problémát okozó jelenségeket (pl. jogszabályhelyek hivatkozásai, bírósági határozatok hivatkozásai stb.) a doménre jellemző valós előfordulási gyakoriság szerinti arányban tartalmazzák, hanem hogy a lehetőségekhez mérten minél nagyobb arányban tartalmazzanak ilyen eseteket<sup>123</sup>. Ennek oka az volt, hogy a korpusz kis mérete ellenére a kiválasztott mondatok minél inkább reprezentálhassák a potenciálisan problémás eseteket, vagyis, hogy a szegmentálási eredmények kiértékelésével az egyes algoritmusok működése ezekben a problémás helyzetekben a leginkább megjósolhatóvá váljon. Éppen ezért az itt bemutatott eredmények nem tükrözik az egyes megoldások „általános esetben” várható teljesítményét, azonban jól bemutatják a jogi doménre jellemző, problémás mondatok esetében tanúsított viselkedésüket. A korpuszban jelenlevő mondatok relatív alacsony elemszámát az indokolja, hogy a kapott eredmények értékelése megbízhatóan csak manuálisan volt elvégezhető, vagyis szükséges volt egyfajta egyensúly megteremtése a reprezentativitás és az ellenőrizhetőség között.

Az erre a célra összeállított korpusz néhány alapvető adatát a 8. táblázat ismerteti.

---

<sup>122</sup> <https://github.com/mediacloud/sentence-splitter>

<sup>123</sup> A mondatok válogatása mögött impliciten az az elgondolás is állt, hogy a nem speciálisan jogi szövegek esetében az egyes eszközök teljesítménye stabilan kiemelkedő volt; a huspaCy esetében például a Szeged korpuszon 97,54% pontosság volt mérhető a mondatszegmentálásban (Orosz et al. 2022: 66).

|                      | Határozatok | NAV tájékoztatók | Törvéyszöveg | Hírek |
|----------------------|-------------|------------------|--------------|-------|
| <b>Token</b>         | 4544        | 3580             | 7900         | 2196  |
| <b>Mondat (kézi)</b> | 100         | 100              | 100          | 100   |

8. táblázat: A mondatsegmentálási hatékonyság ellenőrzésére szolgáló teszt korpusz.

Az egyes csoportokhoz a mondatok a következő forrásokból kerültek kiválasztásra:

- határozat: Pf.V.20.012/2020/9., Gf.V.40.002/2020/10. (bíróági határozatok<sup>124</sup>)
- NAV tájékoztatók: 4. Információs füzet – Magánszemélyek külföldről származó jövedelme<sup>125</sup>
- törvéyszöveg: 2011. évi CCIV. törvény a nemzeti felsőoktatásról<sup>126</sup>
- hírek: 24.hu véletlenszerűen kiválasztott cikkei<sup>127</sup>.

A híreket tartalmazó részkorpusz szolgált „kontroll” adatként, vagyis az itt megtalálható mondatok nem bírtak semmilyen, a jogi szövegeket jellemző komplexitással. Az egyes eszközök által adott, szegmentált mondatokhatárok pontosságát a 9. táblázat mutatja be számszerűen.

|                                      | Határozatok | NAV tájékoztatók | Törvéyszöveg | Hírek     |
|--------------------------------------|-------------|------------------|--------------|-----------|
| <b>magyarlanc</b>                    | 37          | 82               | 52           | <b>90</b> |
| <b>Sentence-splitter</b>             | <b>84</b>   | <b>88</b>        | <b>84</b>    | 82        |
| <b>hu_core_news_lg<sup>128</sup></b> | 76          | 60               | 63           | 88        |

9. táblázat: A mondatsegmentálás eredménye (**kiemelés**: az adott részkorpuszban legjobb eredmény).

Az eredményekből az látszik, hogy a vizsgált eszközök mindegyike hajlamos volt a bemenetként kapott mondatokhoz helytelen szegmentálási eredményt hozzárendelni.

A speciálisan a jogi doménre jellemző szövegtípusok esetében a legmegbízhatóbban konzekvensen a sentence-splitter teljesített. Ez annak köszönhető, hogy a többi megoldással ellentétben ez a csomag lehetőséget ad a felhasználó számára, hogy definiáljon egy ún. kivétellistát azon rövidítésekkel, amelyek szövegbeli előfordulása esetén az

<sup>124</sup> <https://eakta.birosag.hu/anonimizalt-hatarozatok>

<sup>125</sup> [https://nav.gov.hu/ugyfeliranytu/nezzen-utana/inf\\_fuz/2022](https://nav.gov.hu/ugyfeliranytu/nezzen-utana/inf_fuz/2022) (Elérés: 2022.07.11.)

<sup>126</sup> <https://net.jogtar.hu/jogszabaly?docid=a1100204.tv> (Elérés: 2022.07.11.)

<sup>127</sup> <https://24.hu/kultura/2022/07/07/meghalt-bradford-freeman-elit-alakulat-101-hadosztaly-ii-vilaghaboru/>  
<https://24.hu/fn/gazdasag/2022/07/07/rona-peter-forintgyengules-kormany-mnb-hibaja/>  
<https://24.hu/kultura/2022/07/07/obi-wan-kenobi-kritika-star-wars-sorozat-disney-plus/> (Elérés: 2022.07.11.)

<sup>128</sup> A HuSpaCy-hez elérhető magyar nyelvű modellek közül ellenőrzésre választott elnevezése.

algoritmus biztosan nem szükséges, hogy új mondatot kezdjen (pl. törvények rövidítései, mint Btk.). A felhasznált kivétellista alapja a Kúria honlapján elérhető<sup>129</sup> (eredetileg a Stíluskönyv mellékletét képező) összefoglalás volt a hazai joggyakorlatban leggyakrabban hivatkozott jogszabályok rövidítéseivel. Ezt olyan alapvető felsorolásokkal egészítettem ki, mint például római számok vagy sorszámnevek, amelyek szintén nagy valószínűséggel előfordulhatnak, és kontextustól függetlenül nem jelölnek mondatatárt.

A gyakorlatban, vagyis a NAV szakértői által a rendelkezésemre bocsátott korpusz szegmentálása során az előzetes eredmények alapján legjobbnak mutatózó `sentence-splitter` ezen felül eredetileg jelentős mértékben túlszegmentálta a mondatokat. Az ilyenkor keletkező (valójában fals) „mondat jelöltek” viszonylag egyszerűen felismerhetőek voltak a kisbetűs mondat kezdésről, így ezeket a fals eredményeket végül összeillesztettem egy egyszerű algoritmussal valós mondatokká<sup>130</sup>.

Végül a szegmentált mondatok határait reguláris kifejezések is ellenőrzik, és amennyiben egy mondat-jelölt végződése és a következő eleje együtt tipikus, jogszabály hivatkozásra jellemző mintázatot alkot, utolsó lépésként a két mondat jelölt összeillesztése ugyancsak automatikusan megtörténik.

### 6.3.2 Szétválasztás „eredeti” és „átfogalmazott” mondatokra

A 6.2.2 fejezetben már említett, absztrakt szétválogatást a gyakorlatban meghatározó elvek megadására két főbb lehetőség állt rendelkezésre.

Az elkészített korpusz esetében egy python-nyelven írt kód automatikusan megkereste a dokumentumpárokat (A, és B posztfixekkel, lásd: 6.2.2 fejezet), majd meghatározta azokat a mondatokat, amelyek csak az eredeti dokumentumban vannak jelen (ezek kerültek a későbbi „eredeti” alkorpuszba), továbbá azokat, amelyek csak az átfogalmazott változatban vannak jelen (későbbi „átfogalmazott” alkorpusz)<sup>131</sup>. Ezt a módszert a 9. ábra `Method1` szemlélteti.

---

<sup>129</sup> [https://kuria-birosag.hu/sites/default/files/joggyak/3\\_melleklet.pdf](https://kuria-birosag.hu/sites/default/files/joggyak/3_melleklet.pdf) (Elérés: 2022.07.11.)

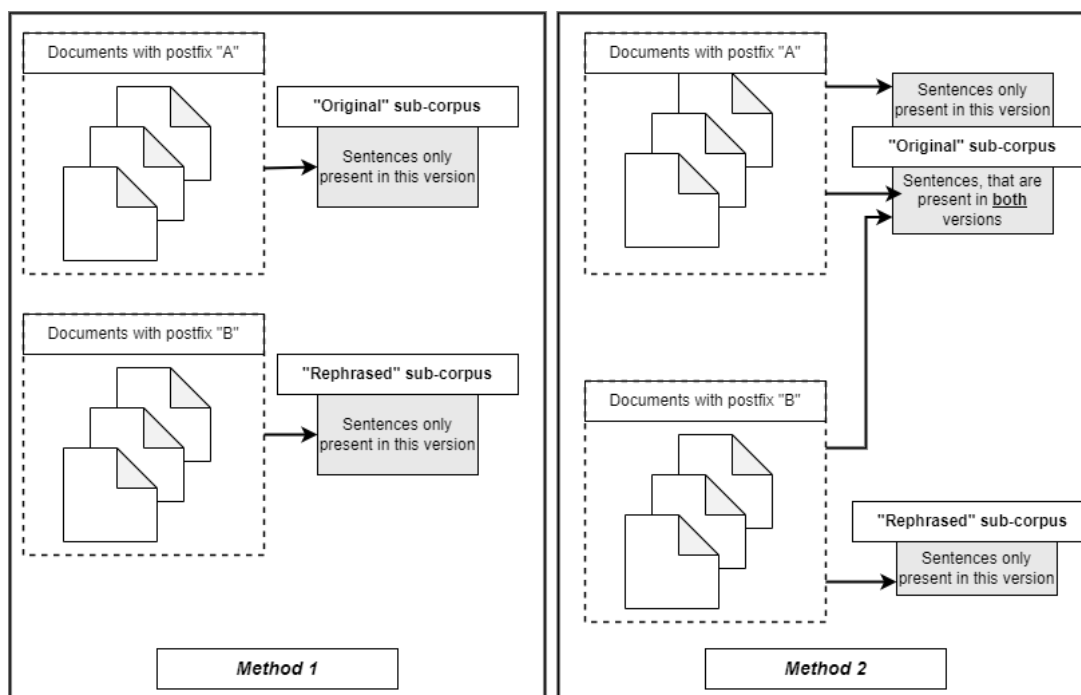
<sup>130</sup> A kisbetűs mondat töredékeket rendre az őket megelőző mondathoz illesztettem mindaddig, amíg a következő „mondat-jelölt” már nagy kezdőbetűvel nem kezdődött.

<sup>131</sup> Halmazelméleti jelöléssel;  $A' = A \setminus (A \cap B)$ , és  $B' = B \setminus (B \cap A)$ , ahol  $A'$ ,  $B'$  a kapott részkorpuszok,  $A$  az eredeti,  $B$  pedig az átfogalmazott mondatok halmaza.

A két részkorpusz különböző mondatainak megkereséséhez egy hash-függvénnyel egyedi értéket rendeltem minden szegmentált mondatához, majd az ismertett halmazműveletet a hash értékek halmazain végeztem el.

Egy másik megoldás<sup>132</sup> lehetett volna az összes olyan mondat kiválasztása, amely mindkét dokumentumban jelen van, kiegészítve azokkal, amelyek csak az eredeti dokumentumokban vannak jelen; ez lett volna az eredeti mondatok(nak a jelenleginél lényegesen nagyobb) halmaza, míg az átfogalmazott mondatok halmaza változatlan maradt volna (ezt a 9. ábra Method2 mutatja).

A kettő közötti leglényegesebb eltérés a hozzájuk kapcsolható kérdésfeltevésben mutatkozik meg. A Method1 esetében a kérdést, amelyre a korpusz „eredeti” mondatainak vizsgálata által választ kaphatunk, nagyjából a következők szerint foglalhatjuk össze; melyek azok az eredeti mondatok, amelyek esetében szükséges volt szakértői revízió? A Method2-ra tekintettel ez sokkal inkább a következők szerint hangozhatna: melyek a kapott adathalmaz megfelelő, és melyek annak problémás mondatai? Ebből látszik, hogy a dolgozat alapvető vizsgálati irányának a Method1 felelt meg inkább.



9. ábra: A tanítóadatok kiválasztásának lehetséges módszerei.

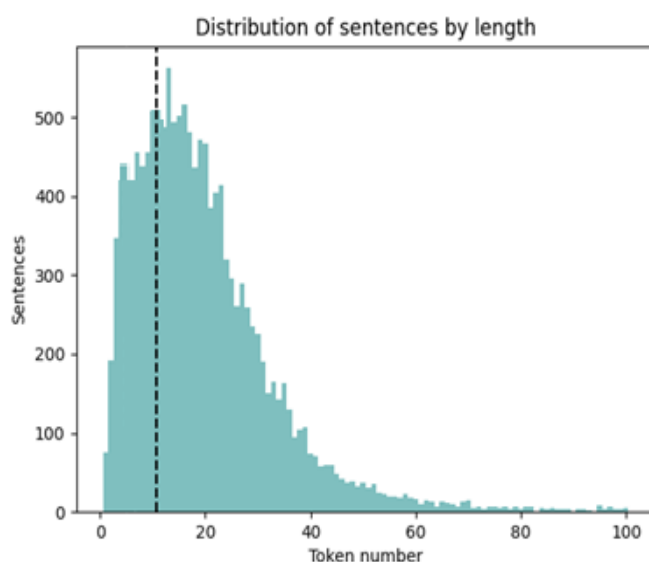
Fontos érv ez utóbbi mellett az is, hogy ilyen módon kiegyensúlyozott korpuszt kapunk, amelyben mindkét címke nagyjából azonos arányban fogja képviseltetni magát. Ez amiatt fontos, hiszen a kiegyensúlyozatlan adathalmazok esetében fennáll annak a veszélye, hogy egyes osztályozók (pl. a 7. fejezetben bemutatott SVM) csak a nagyobb számosságú osztály jellemzőit tanulják meg. Ez végső soron a konkrét osztályozási probléma feletti általánosítás

<sup>132</sup> Az iméntiek szerint;  $A' = A$ ,  $B' = B \setminus (B \cap A)$ .

képességének csökkenéséhez és alacsonyabb hatékonysághoz vezet (Üveges et al. 2022: 521-22). Ezt a jelen esetben mindenképpen célszerűbb volt elkerülni.

### 6.3.3 Adattisztítás és előfeldolgozás

Az utolsó lépés az adatokban jelenlévő zaj eltávolítása, vagy legalábbis arányának a lehető legkisebbre redukálása volt. Ebben a konkrét esetben ez a helytelenül szegmentált mondatok eltávolítását jelentette, valamint azok kivételét, amelyek ugyan valódi mondatok, de nem hordoznak ténylegesen releváns információt a téma szempontjából (pl.: hamisan szegmentált „mondatjelöltek”, amelyek eredetileg csak felsorolások elemei voltak, fejezetcímek, lábjegyzetek stb.). Az adatok kézi vizsgálatát követően a leghatékonyabb és legegyszerűbb megoldásnak a 10 tokennél rövidebb mondatok eltávolítása tűnt<sup>133</sup>.



10. ábra: A korpuszban jellemző mondathosszok (tokenzám).

**A Hiba! A hivatkozási forrás nem található.** az eddig leírtak szerint létrehozott részkorpuszban előforduló mondathosszok eloszlását mutatja. Érdeemes megemlíteni, hogy ugyan a leghosszabb szegmentált mondat 304 token hosszú volt, de mivel az ilyen szélsőségek nagyon ritkának bizonyultak, a jobb áttekinthetőség érdekében az ábra csak az [1-100] intervallumba eső mondathosszok hisztogramját szemlélteti. A függőleges szaggatott vonal a kiválasztott mondathossz 10 tokenes határát mutatja.

<sup>133</sup> A tapasztalatok szerint ezek jellemzően fejezetcímek és lábjegyzetek szövegei voltak, amelyek, bár az olvasó szempontjából hordoznak releváns tartalmat, mégsem tekinthetők olyan egységeknek, amelyekre közérthetőségi javaslatok alkalmazása feltétlenül praktikusnak látszott.

A megoldás előnye (egyszerűsége mellett), hogy relatíve alacsony adatvesztés mellett (az eldobott mondatok az összesített tokenzámnak csak minimális részéért felelősek) jelentős javulást lehetett vele elérni az adatok minőségére tekintettel.

#### 6.4 A végleges korpusz

Az eddig ismertetett lépések után fennmaradó korpusz néhány tulajdonsága a 10. táblázatban látható, ahol is az utolsó sor a részkorpusz százalékos arányát mutatja a teljes korpuszhoz képest.

Az eredeti, Word fájlokból kialakított korpusz kezdetben mintegy 40057 mondatot tartalmazott. Ezek közül összesen 14123 mondatban történt valamilyen mértékű változtatás a Nemzeti Adó- és Vámhivatal szakértői által elvégzett felülvizsgálat során, vagyis ennyi mondatnak volt esélye bekerülni a tanítóadatok közé.

Valamennyi fentebb említett kiválasztási és feldolgozási folyamat után, a tanításhoz kiválasztott korpuszban 10883 mondat maradt, amelyek eloszlása a két lehetséges címkét figyelembe véve szinte tökéletesen kiegyensúlyozottan alakult<sup>134</sup>.

Ez alapján tehát a kiindulási szövegek 35,26%-án történt módosítás, a fennmaradó mondatok változatlanok maradtak. Az átfogalmazáson átesett mondatoknak pedig mintegy 77,06%-a maradt bent a végleges korpuszban a 10 tokennél rövidebb mondatok kizárását követően is<sup>135</sup>.

|                            | <b>Eredeti</b> | <b>Átfogalmazott</b> |
|----------------------------|----------------|----------------------|
| <b>Mondatszám</b>          | 5445           | 5438                 |
| <b>Tokenzám</b>            | 139704         | 127945               |
| <b>Átlagos mondathossz</b> | 25,66          | 23,53                |
| <b>Arány</b>               | 50,03%         | 49,97%               |

10. táblázat: Az elkészített korpusz alapvető mutatói.

Mivel az elsőként kipróbált klasszikus gépi tanuló algoritmusok esetén a szövegeken célszerű előfeldolgozást végezni még azok vektor formára alakítása előtt, ezért a

---

<sup>134</sup> A korrektúrázott verziók kézi ellenőrzése során látható volt, hogy egyes esetekben egy-egy eredeti mondat két-, vagy több átfogalmazott mondatra lett felosztva, illetve egyes átfogalmazott mondatok több eredeti mondat rövidítése és összevonása által keletkezett, ezért a számosságuk nem végeredményben nem teljesen azonos.

<sup>135</sup> A kiindulási korpusznak a Method1 szerint leválogatott változata tehát a teljes korpusznak mintegy 27,16%-át teszi ki.

modelltanítás előtti utolsó lépés a megmaradt mondatok átalakítása volt néhány szokásos lépéssel;

- kisbetűsítés,
- számok és írásjelek szűrése,
- lemmatizálás.

## 7. Modelltanítás SVM-mel

A 6.1 fejezetben írtaknak megfelelően a későbbi (egyben korszerűbb) gépi tanuló algoritmusok kipróbálása előtt szükséges volt valamilyen baseline eredmény, amelyhez azok teljesítménye mérhető lesz. Erre a célra a `scikit-learn` (`sklearn`<sup>136</sup>) Szupport Vektor Gép (Support Vector Machine – a továbbiakban: SVM) algoritmusát választottam hiperparaméter optimalizálással egybekötve a lehető legjobb eredményt hozó beállítások megtalálása érdekében.

A gépi tanuló algoritmusok általában két paraméterkészlettel rendelkeznek; a modellparaméterek tanulása (vagy becslése) a gépi tanulási folyamat része, míg a hiperparaméterek nem tanulhatók közvetlenül a rendelkezésre álló adatokból, így optimalizálásuk manuálisan végezhető el. Ez utóbbi tulajdonságuk ellenére beállításuk nagy hatással lehet az algoritmus kimenetére és ezáltal teljesítményére is. Az SVM esetében ilyen hiperparaméterek lehetnek például:

- a  $C$  (regularizáció - amely az egyes osztályokat elválasztó hipersík margójának méretét szabályozza),
- a  $\Gamma$  (amely azt befolyásolja, hogy a potenciális elválasztási határtól távol eső pontok mennyire vesznek még részt a döntésben) vagy
- a kernel, avagy magfüggvény (amely a modell tanulása során az egyes osztálycímék között konstruált hipersík számítási módjára lesz befolyással).

A jelenlegi kísérletben a TF-IDF vektorizálás (Sammur és Webb 2011) került alkalmazásra az adatok (mondatok) végső, gépileg olvasható formátumba alakításához.

Az említett hiperparamétereket a következő értékek összes lehetséges kombinációja ( $4^3$ ) szerint ellenőriztem:

- $C$ : [0.1, 1, 10, 100]
- $\gamma$ : [1, 0.1, 0.01, 0.001]
- kernel: ['rbf', 'poly', 'sigmoid', 'linear'].

---

<sup>136</sup> <https://scikit-learn.org/stable/> (Elérés: 2024. 01. 29.)



Tekintettel arra, hogy a tanítóadatok mennyisége viszonylag alacsony, a modelleket jellemző szokásos metrikákat (pontosság, fedés és F1-érték) az adatok 10-szeres keresztvalidálása<sup>137</sup> után értékeltem ki. A tanító- és tesztadatok eloszlása minden egyedi esetben 90%-10% szerint lett kialakítva, továbbá minden tesztalmazba az egyes osztálycímkével jelölt mondatok 10-10% -a került<sup>138</sup>.

## 7.1 Eredmények

A 11. táblázat az eredeti- és az átfogalmazott mondatok elkülönítésében legjobbnak bizonyuló 10 modell eredményét mutatja be azok beállításával együtt. Az eredmények azt mutatják, hogy általánosságban a legjobb kernelfüggvénynek a "lineáris" kernel tűnik (az élre rangsorolt 10 modelltől 6 ezzel a beállítással rendelkezett), míg a regularizáció (C) a 10-es értékre beállítva bizonyult optimálisnak (a legjobb 4 modellt ez az érték jellemezte). A Gamma paraméter kivételnek tekinthető abból a szempontból, hogy nem mutat egyértelmű tendenciát; a [0.01, 1] intervallumon belüli értékei látszólag véletlenszerűen fordulnak elő a modellek rangsorolása során, ezek közül pedig kizárólag a 0.001-es érték optimalitása zárható ki egyértelműen. A kiválasztott modellek kiegyensúlyozottnak tűnnek az eredeti és a már átfogalmazott mondatok felismerése tekintetében.

Az általános teljesítményen felül további fontos kérdés volt, hogy mely modellek képesek a legpontosabban azonosítani az eredeti mondatokat, hiszen a szoftver célja, amelybe a legjobban teljesítő modell be fog kerülni, a szöveg problémás pontjainak azonosítása lesz a szakértők számára. A 12. táblázat ezt a szempontot tükrözi, feltüntetve azt az 5 modellt, amelyek a legjobban teljesítettek az eredeti mondatok azonosításában; a rangsorolás alapját ez esetben a korpusz „eredeti” mondatainak azonosítása során mért pontosság (P) képezte.

---

<sup>137</sup> A keresztvalidálás olyan eljárás, amelynek során minden adatpont pontosan egyszer lesz a tesztalmaz része, 9 alkalommal pedig a tanítóadatok között fog szerepelni. Ennek jelentősége, hogy mivel minden adat tanító és tesztadatként is szerepel, lényegesen kiegyenlítődik a tanító- és tesztalmazok véletlenszerű felosztása során keletkező „szerencsefaktor”. Más szóval, a keresztvalidálás során kapott eredmények átlagolása sokkal realisabb képet fest az algoritmus valós teljesítményéről, mint az egyszerű (és egyszeri) felosztás.

<sup>138</sup> Ez a scikit-learn könyvtár `train_test_split` függvényének `stratify` paramétere által volt szabályozható.

| Paraméterek                            | Osztály       | P (AVG)     | P (STD) | R (AVG)     | R (STD) | F1 (AVG)    | F1 (STD) | F1 (AVG - 2 osztályra) |
|--|---------------|-------------|---------|-------------|---------|-------------|----------|------------------------|
| C: 10, gamma: 1,<br>kernel: linear     | eredeti       | <b>0.66</b> | 0.02    | <b>0.69</b> | 0.05    | <b>0.68</b> | 0.04     | <b>0.68</b>            |
|  | átfogalmazott | <b>0.68</b> | 0.03    | <b>0.65</b> | 0.03    | <b>0.66</b> | 0.02     |                        |
| C: 10, gamma: 0.1,<br>kernel: linear   | eredeti       | <b>0.66</b> | 0.02    | <b>0.69</b> | 0.05    | <b>0.68</b> | 0.04     | <b>0.68</b>            |
|  | átfogalmazott | <b>0.68</b> | 0.03    | <b>0.65</b> | 0.03    | <b>0.66</b> | 0.02     |                        |
| C: 10, gamma: 0.01,<br>kernel: linear  | eredeti       | <b>0.66</b> | 0.02    | <b>0.69</b> | 0.05    | <b>0.68</b> | 0.04     | <b>0.68</b>            |
|  | átfogalmazott | <b>0.68</b> | 0.03    | <b>0.65</b> | 0.03    | <b>0.66</b> | 0.02     |                        |
| C: 10, gamma:<br>0.001, kernel: linear | eredeti       | <b>0.66</b> | 0.02    | <b>0.69</b> | 0.05    | <b>0.68</b> | 0.04     | <b>0.675</b>           |
|  | átfogalmazott | <b>0.68</b> | 0.03    | <b>0.65</b> | 0.03    | <b>0.66</b> | 0.02     |                        |
| C: 100, gamma: 0.1,<br>kernel: rbf     | eredeti       | <b>0.66</b> | 0.03    | <b>0.68</b> | 0.05    | <b>0.67</b> | 0.03     | <b>0.67</b>            |
|  | átfogalmazott | <b>0.68</b> | 0.03    | <b>0.65</b> | 0.02    | <b>0.67</b> | 0.02     |                        |
| C: 100, gamma: 0.1,<br>kernel: sigmoid | eredeti       | <b>0.66</b> | 0.02    | <b>0.69</b> | 0.05    | <b>0.67</b> | 0.03     | <b>0.67</b>            |
|  | átfogalmazott | <b>0.68</b> | 0.03    | <b>0.65</b> | 0.03    | <b>0.67</b> | 0.02     |                        |
| C: 100, gamma:<br>0.01, kernel: rbf    | eredeti       | <b>0.68</b> | 0.03    | <b>0.66</b> | 0.07    | <b>0.67</b> | 0.05     | <b>0.66</b>            |
|  | átfogalmazott | <b>0.67</b> | 0.04    | <b>0.68</b> | 0.03    | <b>0.67</b> | 0.03     |                        |
| C: 1, gamma: 1,<br>kernel: sigmoid     | eredeti       | <b>0.69</b> | 0.04    | <b>0.61</b> | 0.06    | <b>0.65</b> | 0.05     | <b>0.65</b>            |
|  | átfogalmazott | <b>0.65</b> | 0.03    | <b>0.73</b> | 0.04    | <b>0.69</b> | 0.03     |                        |
| C: 1, gamma: 1,<br>kernel: linear      | eredeti       | <b>0.68</b> | 0.04    | <b>0.62</b> | 0.06    | <b>0.65</b> | 0.05     | <b>0.65</b>            |
|  | átfogalmazott | <b>0.65</b> | 0.04    | <b>0.71</b> | 0.05    | <b>0.68</b> | 0.03     |                        |
| C: 1, gamma: 0.1,<br>kernel: linear    | eredeti       | <b>0.68</b> | 0.04    | <b>0.62</b> | 0.06    | <b>0.65</b> | 0.05     | <b>0.65</b>            |
|  | átfogalmazott | <b>0.65</b> | 0.04    | <b>0.71</b> | 0.05    | <b>0.68</b> | 0.03     |                        |

11. táblázat: A 10 legjobban teljesítő modell eredményei (P: Pontosság, R: Fedés, F1: F1-érték, keresztvalidálás során kapott értékek: AVG: átlag, STD: a megfigyelt szórás).

| Paraméter készlet                   | P (AVG)     | P (STD) | R (AVG) | R (STD) | F1 (AVG) | F1 (STD) |
|-------------------------------------|-------------|---------|---------|---------|----------|----------|
| C: 0.1, gamma: 1, kernel: rbf       | <b>0.79</b> | 0.28    | 0.03    | 0.03    | 0.05     | 0.06     |
| C: 0.1, gamma: 1, kernel: sigmoid   | 0.72        | 0.05    | 0.45    | 0.08    | 0.55     | 0.07     |
| C: 0.1, gamma: 1, kernel: linear    | 0.71        | 0.04    | 0.47    | 0.08    | 0.57     | 0.07     |
| C: 0.1, gamma: 0.1, kernel: linear  | 0.71        | 0.04    | 0.47    | 0.08    | 0.57     | 0.07     |
| C: 0.1, gamma: 0.01, kernel: linear | 0.71        | 0.04    | 0.47    | 0.08    | 0.57     | 0.07     |

12. táblázat: A legjobban illeszkedő modellek az átfogalmazandó mondatok előrejelzésére (jelölések a 11. táblázatnak megfelelően).

## 7.2 Következtetések

A 11. és 12. táblázatban kiválasztott modell-halmazok szinte teljes átfedést mutatnak a legjobb modellt kivéve, amely képes volt átlagosan 0.79 -es pontosság elérésére is az eredeti mondatok felismerése során. Azonban az itt megfigyelt szórás extrém magas értéke arra enged következtetni, hogy az egyes validációs halmazokon a teljesítmény rendkívül széles skálán mozgott, ami csökkenti a megbízhatóságot a többi paraméter készlettel rendelkező modellel szemben, melyek esetében a szórások egymással homogén módon alakultak.

Éppen ezért gyakorlati alkalmazás szempontjából a rangsor második helyén álló modell tűnik a legalkalmasabbnak. Az optimális hiperparaméterek ebből a szempontból tehát a következők szerint alakultak; a regularizáció esetében 0,1, a gamma tekintetében pedig az 1-es érték. A kernelfüggvényre a legjobb választásnak ez esetben a sigmoid tűnik, habár a többi pozíció esetében ismét csak a lineáris kernel dominált.

## 8. Modelltanítás a fastText alkalmazásával

A FastText egy viszonylag új (nyílt forráskódú) gépi tanulási megoldás, amely alkalmazható szóbeágyazások tanulására és felügyelt szövegosztályozásra is. Az algoritmust a Facebook AI Research (FAIR) laboratórium hozta létre (Joulin et al. 2016).

A módszer jelentős gyakorlati előnye más neurális hálón alapuló megoldásokkal szemben a könnyű használhatóság (a kipróbált modellek paraméterkészleteinek inicializálásán felül a tanítási folyamat nem igényel mélyebb matematikai ismereteket). Ezzel azonban járulékosan együtt jár az a hátrány is, hogy a modellek konstruálásáért felelős osztályok közel sem nyújtanak annyi testreszabhatóságot, mint például a `scikit-learn` LSTM (amely alkalmas például az egyes rejtett rétegek tulajdonságaira vonatkozó egyedi beállítások megadására is).

Lényeges különbség a 7. fejezetben ismertetett megoldáshoz képest, hogy míg az SVM esetében az algoritmus számára a szövegeket reprezentáló vektorizációt TF-IDF-fel állítottam elő, addig a fastText esetében a modell bemenetét (ahogy a neuronháló alapon működő gépi tanuló algoritmusok esetében általában, úgy itt is) szóbeágyazások alkotják<sup>139</sup>.

A projekt honlapján 157 különböző nyelvre állnak rendelkezésre előzetesen betanított vektorok. Ezek közül sok olyan, amelyekre a "nagy nyelvekhez", például az angolhoz mérten jelentősen kevesebb NLP-eszköz érhető el.

A könnyű kezelhetőség mellett a fastText használatának másik előnye az általa használt szószint alatti (sub-word level) reprezentációs forma. Míg a TF-IDF nem tudja kezelni a tanítóhalmazban nem látott, azaz szótáron kívüli (OOV - Out of Vocabulary) szavakat, addig a fastText modellek képesek reprezentálni az ilyen elemeket is azáltal, hogy az azokat alkotó karakter n-grammokból rendelnek hozzájuk reprezentációt.

---

<sup>139</sup> Ezeket jellemzően szintén egy kiinduló neurális hálózatból kaphatjuk meg olyan módon, hogy az adott szó one-hot encoded reprezentációját odaadjuk a kiinduló hálózatnak, és a rejtett rétegekben kapott súlyokat a tanítási folyamat után kivonjuk (skip-gramok esetén). A rejtett rétegek száma fogja megadni a kapott beágyazások végleges dimenziószámát.

## 8.1 Eredmények

A kutatáshoz a fastText hivatalos weboldaláról származó, magyar nyelvre előzetesen betanított szövektorokat<sup>140</sup>, valamint a saját korpuszon tanított szövektorokat is kipróbáltam. Az előbbiek előnye, hogy sokkal nagyobb adathalmazon tanított modellekből származnak, míg az utóbbiak feltehetőleg jobban tudják reprezentálni a doménspecifikus, vagy akár csak az aktuális feladatra jellemző jelentésbeli különbségeket az egyes szavak között.

A hangolásra kiválasztott hiperparaméterek ez esetben a következők voltak:

- learning rate ('lr'): a modell tanulási sebességének szabályozására szolgál. Alacsonyabb érték esetében a tanítóhalmazra való rátanulás mértéke lassítható. A FastText hivatalos oldalának ajánlása szerint<sup>141</sup> optimális értéke 0,1 és 1 közötti érték közé becsülhető.
- n-gramok megengedett mérete ('wordNgrams'): Ha ezt  $\geq 1$ -re állítjuk, meghatározhatjuk általa a tanítási fázisban figyelembe veendő szó kollokációk (együtt előforduló tokenek) méretét. A legegyszerűbb esetben unigramokat (csak egyedi tokenek; wordNgrams = 1) használunk, de olyan összetett problémáknál, mint a hivatalos szövegek érthetősége, ennél magasabb érték is előnyös lehet.
- maximális epochok száma ('epoch'): az epochok maximális számának beállításával szabályozható, hogy a modell hányszor lásson egy-egy tanítópéldát a tanulás során. Az alapértelmezett érték 5, de magasabb szám beállítása is megfelelő lehet.
- Előre tanított, vagy a korpuszon tanított szóbeágyzások használata ('pretrainedVectors'): a szóbeágyzások előállítása az adott adathalmazon, vagy előre tanított vektorok alkalmazása.
- A minimális előfordulási gyakoriság, amelyet egy tokennek el kell érnie ahhoz, hogy a tanítás során figyelembe vegyünk ('minCount').
- A használt szóbeágyzások dimenziója ('dim'): a szóbeágyzások dimenziójának csökkentésével különböző kísérleteket lehet végezni kisebb dimenziójú szövektorokkal is (pl. a jelen esetben 100, 50 dimenziós esetek az alapértelmezés szerinti 300-on felül).

---

<sup>140</sup> A hivatalos weblapon a szövektoroknak egy .bin formátumú verziója érhető el, azonban ezek használatához a modellnek egy .vec formátumra van szüksége. Az átalakításhoz szükséges konverziót külön meg kellett valósítani, amelyhez az alap ötletet a fastText GitHub repozitóriumának egyik fórumbejegyzése adta: <https://github.com/facebookresearch/fastText/issues/1082>.

<sup>141</sup> <https://fasttext.cc/docs/en/supervised-tutorial.html>

Az SVM-mel ellentétben itt 80%-20%-ban osztottam fel a korpuszt tanító- és tesztadatokra. A modelltanítás előtt kérdés volt, hogy az előre tanított szóvektoroktól jobb teljesítmény várható-e, mint a korpuszon betanítottaktól. Mindkét lehetőség mellett szólhatnak érvek. Tekintettel az előre tanított vektorokhoz felhasznált, több nagyságrenddel kiterjedtebb adathalmazra, indokolható lenne, hogy ezek hozzanak jobb teljesítményt. A gépi tanulási kísérletek során azonban előfordulhat, hogy a konkrét korpuszban jelenlévő speciális jelenések sokkal inkább megfoghatók a korpuszon tanított szóbeágyazásokkal. Előfeltevés volt, hogy az egyszerre több tokenet is figyelembe vevő n-grammok felhasználásával jobb teljesítmény várható, mint pusztán unigramok figyelembevételével. A fent felsorolt többi paraméter értékeire vonatkozóan nem állt rendelkezésre hasonló prekonceptió.

Mivel a fastText véletlenszerűen inicializálja az összes súlyt a modellen belül minden egyes futtatáskor, illetve nem ad módot ennek a viselkedésnek a szabályozására (mint amilyen például sklearn modellek esetében a 'random state' paraméter), indokolt volt minden paraméterkészletet epochonként több futtatással is tesztelni.<sup>142</sup> Ehhez az 5 futás / epoch értéket választottam, majd a kapott eredményeket átlagoltam, hogy a predikciós pontosságról realisabb kép alakulhasson ki.

A felügyelt gépi tanulásban a fő cél az, hogy olyan modellt tudjunk létrehozni, amely nemcsak a tanítóadatokon hatékony, hanem képes általánosítani a korábban soha nem látott adatok jellemzői felett is. Az ilyen jellegű feladatok megoldása lényegében egy optimalizálási problémát eredményez, amelyet strukturális kockázatminimalizálásnak (Structural Risk Minimization, SRM) nevezünk. Ennek során a cél az, hogy általánosan a leghatékonyabb modellt tudjuk megalkotni mindössze véges számú tanítóadat rendelkezésre állása mellett (Vapnik 2006).

Ebben a minimalizálási problémában a veszteségfüggvény (loss function) lesz az az összetevő, amely segít meghatározni a modell tényleges kimenete és a várható kimenet közötti távolságot, vagyis amelynek segítségével a modell módosítható a jobb eredmények elérése érdekében (Wang et al. 2020).

A veszteségfüggvény tanító és validálási halmazokon elért értékeinek mérésével a mélytanuló hálózatok két tipikus problémájára, az alul- és túlillesztésre is következtethetünk. Az előbbi esetben a veszteségfüggvény a tanításkor meghatározott epochok végét követően

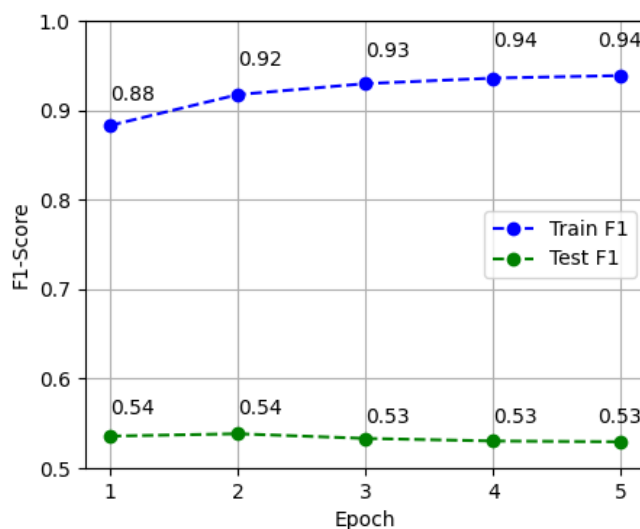
---

<sup>142</sup> Gépi tanulás során egy epoch alatt az a teljes ciklust értjük, amelynek során a modell minden rendelkezésre álló tanítóadatot legalább egyszer „látott”.

még mindkét halmazon csökkenő tendenciát mutat - ekkor további tanításra van szükség (under fit). Az utóbbi esetben a kezdeti csökkenés után a validációs halmazon a veszteségfüggvény értéke növekedni kezd, miközben a tanítóhalmazon értéke tovább csökken; ezen a ponton a modell már nem képes hatékonyan általánosítani az új adatokon (overfit).

Az előre tanított vektorokkal végzett előzetes kísérletezés után (mindössze 5 epochos tanítás mellett) kiderült, hogy az ezeken a vektorokon betanított modellek minden esetben (azaz valamennyi kipróbált paraméterkészlet mellett) már az első epoch után is túlillesztenek. Ez olyan esetben fordul elő, amikor a modell mintegy „bemagolja” a tanítóadatokat, ezáltal pedig elveszti azt a fajta általánosítási készséget, ami a korábban még sosem látott példák helyes osztályozásához lenne szükséges<sup>143</sup>. A jelenséget érdemben nem befolyásolta egyik fent említett paraméter változtatása sem. A 11. ábra egy tipikus eredményt mutat be, miután a modellt mind a tanító-, mind pedig a teszhalmazon kiértékeltem.

Miután világossá vált, hogy az előre tanított szövektorok nem képesek általánosítani az adatokon (generalizálási hiba), a kísérletezés következő fázisában csak a konkrét korpuszon betanított vektorokat alkalmaztam.

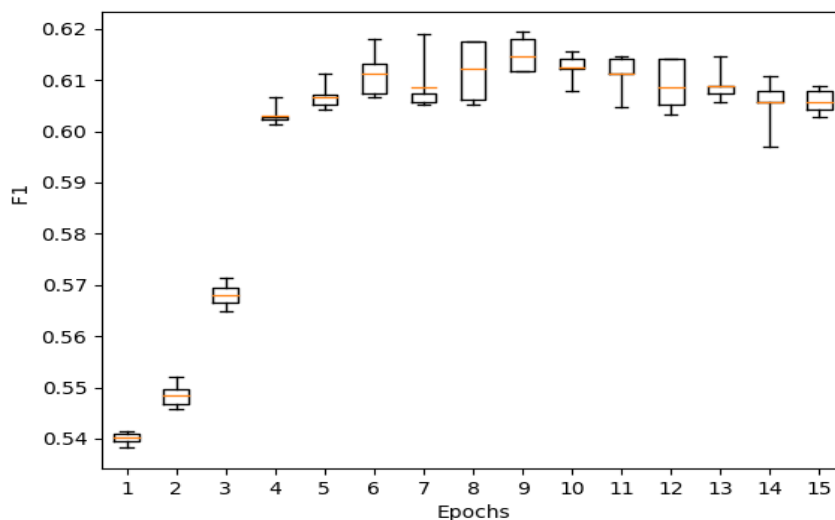


11. ábra: A mért F1-értékek {"lr": 0.3, "wordNgrams": 1, "minCount": 1, "epoch": 5} paraméterkészlet mellett.

<sup>143</sup> vö.: <https://www.ibm.com/cloud/learn/overfitting>

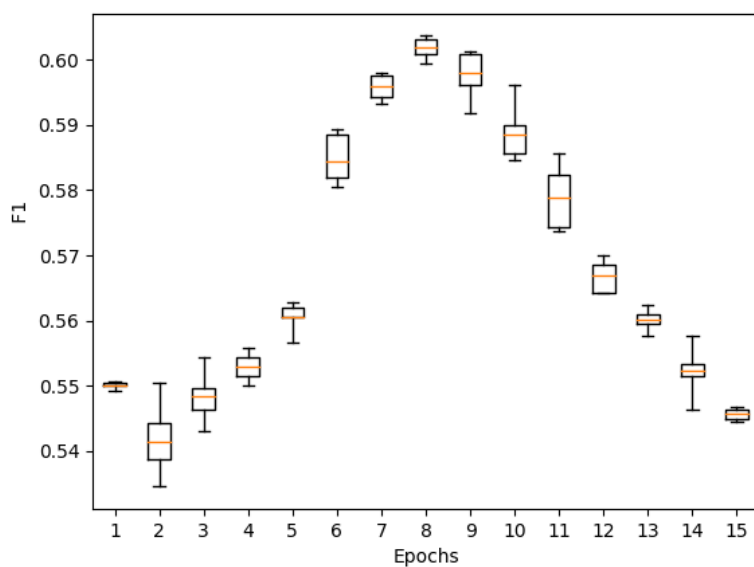
A lehetséges paraméterek különböző kombinációinak vizsgálatából a következő főbb következtetések voltak levonhatók:

- A túlillesztés elkerülése érdekében a 'learning rate' paramétert az ajánlott minimális értékre (0,1) volt célszerű beállítani,
- az n-grammok használata az unigramok helyett jelentősen rontotta a tesztalmazon mért eredményeket minden egyéb paraméter kombináció esetében,
- azt az értéket, amely lehetővé tette, hogy egy bizonyos előfordulási gyakoriság feletti tokeneket vegyünk csak figyelembe az osztályozás során (minCount), a legcélszerűbb volt  $\geq 10$ -es értékre állítani (magasabb értékre, pl. 20-ra vagy 30-ra állítva ez már nem javította jelentősen a tesztalmazon kapott eredményeket),
- a használt szóbeágyazások dimenziószámának megváltoztatása nem volt jelentős hatással az eredményekre, de úgy tűnt, hogy a dimenziószám növelése az egyébként 5 epoch után elérhető platót az F1-értékek grafikonján jelentősen redukálja a későbbi epochok során (túlillesztést eredményez magasabb epochszám mellett).

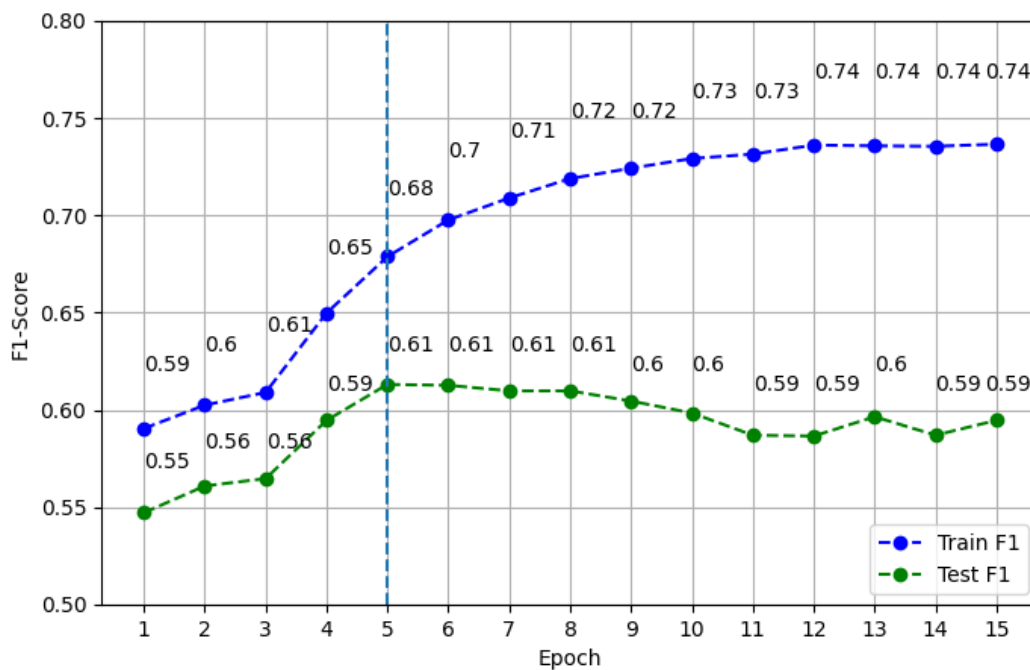


12. ábra: F1-értékek a legjobbnak bizonyult paraméter készlettel: (a min, max, q1 és q3 kvartilisek és az átlag az epochonként 5 különböző futtatás átlagát tekintve).





13. ábra: Annak illusztrálása, hogy a beágyazási dimenzió 50-ről 300-ra történő változtatása nem befolyásolta jelentősen a legjobb eredmény elérését, de a következő epochok során erősen erős túlillesztést váltott ki.



14. ábra: Annak illusztrálása, hogy a beágyazási dimenzió 50-ről 300-ra történő változtatása nem befolyásolta jelentősen a legjobb eredmény elérését, de a következő epochok során erősen erős túlillesztést váltott ki.

A legjobb eredményeket a 'learning rate' 0,1-re, az 'ngram-size' 1-re, a minCount 20-ra és a beágyazások dimenziójának 50-re (a továbbiakban "legjobb paraméterkészlet") történő beállításával voltak elérhetőek. A kapcsolódó F1-értékeket a 12. ábra szemlélteti<sup>144</sup>.

A 14. ábra mutatja be a korábban említetteknek megfelelően a beágyazások dimenziójának növelése által kiváltott hatást (tútanulás magasabb epochok esetében). Ebben az esetben a 'learning rate' ismét 0,1, az 'ngram-size' 1, a 'minCount' 20, a beágyazások dimenziója azonban 50 helyett 300-ra lett beállítva. A legjobb paraméterkészlet értékeitől ez csak a beágyazások dimenziószámában tért tehát el.

A tútanulás megjelenése az optimális paraméterkészlet mellett leolvasható például a 14. ábra adataiból.

A train és a teszt F1-eredmények (5 különböző futtatás átlaga) összehasonlításával látható, hogy az optimális epochaszám ebben az esetben 5, hiszen ekkor lehetett a legjobb eredményt elérni a tanítóhalmazon úgy, hogy a különbség a tanító- és teszthalmazon mért F1-értékek között még viszonylag alacsony marad. Ezután az F1-érték lassan csökkenni kezd a teszthalmazon, míg a tanítóadatokon továbbra is folyamatosan emelkedik, ami a túlillesztés egyértelmű indikátora.

Az utolsó lépés annak ellenőrzése volt, hogy a 6.4 fejezetben említett standard előfeldolgozási lépéseken túl a szövegek lemmatizálása mennyivel javítja tovább a teljesítményt?

A lemmatizáció a természetes nyelvfeldolgozásban (NLP) gyakran használt technika, amikor morfológiailag gazdag nyelvekkel foglalkozunk. Fő különbsége a szótövezéstől (stemming, amely például az angol nyelv esetében gyakran használt technika) azzal írható le, hogy a szótövezés főként csak eltávolítja a szavak végéről a különböző képzőket, jeleket, ragokat (ezért nem feltétlenül eredményez értelmes szóalakot). Ezzel szemben a lemmatizálás mindig a szó szótári alakját adja vissza eredményként (Jurafsky - Martin 2022: 20)<sup>145</sup>.

Mindkét megközelítést a vektortérmodell építése során arra használják, hogy az egyes szavak különböző szóalakjait egy közös kanonikus formára hozzák, lehetővé téve azok

---

<sup>144</sup> Megjegyzendő, hogy a kísérletezés során számos paraméter beállítás mellett kiértékeltem az eredményeket, azonban azok rendre az itt látható tipikus mintázat szerint alakuló F1-értékeket eredményeztek, így ezeket külön nem ábrázoltam.

<sup>145</sup> Vesd össze például: „bokrokot” – szótó (képzők, jelek, ragok csonkolása utáni alak): „bokr”, de lemma: „bokor”.

egységes reprezentációját. Ez a gyakorlatban azt jelenti, hogy a szóalakhhoz kapcsolódó információ ugyan elvész, de a vektortér dimenziója jelentősen csökkenthető. Az angol nyelvben ez a csökkenés elérheti a 40-70%-ot, míg a magyarban egyes megfigyelések szerint akár 90%-nyi is lehet (Tikk 2007: 41). Az, hogy az egyes szóalakok külön reprezentációja felesleges zaj vagy értékes információ, erősen domén- és alkalmazásfüggő lehet.

Annak érdekében tehát, hogy a tesztalmazon magasabb F1-érték legyen elérhető, a rendelkezésre álló mondatok lemmatizált formáját is kiértékeltem fastText modellekkel (ugyanazon, optimálisnak tekintett paraméterkészletek mellett, mint korábban). A lemmatizálást a spaCy<sup>146</sup> magyar nyelvi modelljével végeztem el (hu\_core\_news\_lg). Az eredmények azt mutatták, hogy az izolált szóalakok számának drasztikus csökkenése (-68,74%) ellenére a felismerési pontosság és a fedés a tesztalmazon nem változott érdemben.

| Epoch | A                 |        |                        |        | B                 |        |                        |        |
|-------|-------------------|--------|------------------------|--------|-------------------|--------|------------------------|--------|
|       | Simple preprocess |        | Simple + lemmatization |        | Simple preprocess |        | Simple + lemmatization |        |
|       | Train             | Test   | Train                  | Test   | Train             | Test   | Train                  | Test   |
| 1     | 0.5750            | 0.5402 | 0.5748                 | 0.5409 | 0.5890            | 0.5497 | 0.5899                 | 0.5511 |
| 2     | 0.5810            | 0.5486 | 0.5776                 | 0.5488 | 0.5883            | 0.5497 | 0.6033                 | 0.5686 |
| 3     | 0.6135            | 0.5679 | 0.6143                 | 0.5680 | 0.5955            | 0.5539 | 0.6052                 | 0.5620 |
| 4     | 0.6694            | 0.6031 | 0.6699                 | 0.6047 | 0.6416            | 0.5950 | 0.6060                 | 0.5575 |
| 5     | 0.7026            | 0.6067 | 0.7030                 | 0.6068 | 0.6890            | 0.6075 | 0.6392                 | 0.5872 |
| 6     | 0.7168            | 0.6113 | 0.7165                 | 0.6123 | 0.7132            | 0.6103 | 0.6685                 | 0.6024 |
| 7     | 0.7270            | 0.6085 | 0.7277                 | 0.6150 | 0.7223            | 0.6187 | 0.6961                 | 0.6059 |
| 8     | 0.7334            | 0.6121 | 0.7333                 | 0.6121 | 0.7315            | 0.6159 | 0.7119                 | 0.6056 |
| 9     | 0.7375            | 0.6146 | 0.7378                 | 0.6123 | 0.7362            | 0.6101 | 0.7304                 | 0.5981 |
| 10    | 0.7416            | 0.6125 | 0.7419                 | 0.6079 | 0.7400            | 0.6096 | 0.7439                 | 0.5930 |

13. táblázat: A 2 paraméterkészlet mért F1-értékeinek összehasonlítása a lemmatizálás hatásának szempontjából – ismét 5 futás / epoch átlagát tekintve.

A 13. táblázat két modellpár eredményeit mutatja be, amelyek közül az egyik minden esetben lemmatizált szöveget kapott bemenetként, a másik pedig a vektortér csökkentését szolgáló előfeldolgozási lépések nélkül előkészített szöveget (a többi paraméter mindkét esetben azonos volt). A dimenziószám hatásának ismételt teszteléséhez a modell tanítást itt

<sup>146</sup> A spaCy egy nyílt forrású nyelvtechnológiai keretrendszer, amely több nyelv nyelvtechnológiai feldolgozásához is standardizált kereteket biztosít (<https://spacy.io/>). A korábban is említett huSpaCy voltaképpen ennek egy többé-kevésbé függetlenül fejlesztett, magyar változata (<https://huspacy.github.io/>).

is 50 és 300 dimenziós szóbeágyazások mellett is elvégeztem. Az alkalmazott paraméter készletek a táblázat jelölései szerint:

- A: {"lr": 0.1, "wordNgrams": 1, "minCount": 20, "epoch": 10, "dim": **50**}
- B: {"lr": 0.1, "wordNgrams": 1, "minCount": 20, "epoch": 10, "dim": **300**}

Az eredmények alapján megállapítható, hogy a két modell F1-értékei közötti különbség általában kisebb, mint .01, ami elhanyagolhatónak tekinthető.

## 8.2 Következtetések

A kapott adatok alapján adódik a következtetés, hogy a lemmatizálás (noha jelentősen csökkenti a reprezentálandó szóalakok arányát) érdemben nem változtatott az eredményeken. Ez valamelyest ellentétes a szakirodalomban korábban fellelt adatokkal. Összevetve ezt a fastText, valamint az SVM használatával elérhető maximális F1 értékekkel, adódik a következtetés, hogy a modell szokatlan viselkedésének oka feltehetőleg a használt szóbeágyazásokban, azok „minőségében” keresendő. Érdekes megfigyelés, hogy sem a több tokenes kollokációk (n-gramok) figyelembevétele, sem pedig a használt szóbeágyazások dimenziószámának növelése nem hat ki pozitívan az osztályozási minőségre.

A legjobb eredményhez fastText modellek esetében világosan a korpuszon tanított szövektorok alkalmazása vezetett (szemben az előretanított beágyazásokkal). Emellett szintén előnyösen hatott az osztályozás során figyelembe vett tokenek előfordulási gyakoriságának egy minimális értékhez rögzítése is.

## 9. Modelltanítás BERT-tel

Az elmúlt években az előre tanított (pre-trained) nyelvmodellek váltak a legtöbb NLP feladat esetében a legkorszerűbb megoldásokká. A több százmillió hangolható paraméterrel rendelkező modellek, mint például az ELMo (Peters et al. 2018), GPT (Radford – Narasimhan 2018), BERT (Devlin et al. 2018) vagy a RoBERTa (Liu et al. 2019) jelentős teljesítmény növekedést eredményeztek számos NLP-feladatban.

Az ilyen, ún. transformer alapú modellek alkalmazásával számos területen sikerült a korábbi gépi tanuló megoldásokhoz képest kimagasló eredményeket elérni, legyen szó akár (aspektusszintű) szentiment- és érzelelemzésről (Üveges et.al. 2022), gépi fordításról (Zhu et al. 2020) vagy automatikus szövegenerálásról, amilyen például az absztraktív összefoglalók készítése egy szöveg tartalmának rövid reprezentálása érdekében (Yang Zijian et. al. 2020).

Konkrétnan a BERT (Bidirectional Encoder Representations from Transformers) modellek egyik erőssége, hogy a bennük felhasznált transformer-architektúra miatt (Vaswani et al. 2017) egyrészt azok betanítása a korábbinál sokkal hatékonyabban párhuzamosítható, másrészt a „figyelmi-mechanizmus”-nak köszönhetően (attention mechanism) azok képesek megragadni a nyelvi elemek (szavak) közötti kapcsolatokat is. Előbbi azért fontos, hiszen ezt kihasználva lehetőség nyílik korábban (főleg performancia okokból) elképzelhetetlen méretű korpuszok alkalmazására tanítóadatként, utóbbi pedig lehetővé teszi, hogy az elkészült nyelvmodell képes legyen tárolni a benne reprezentáció szinten előforduló elemekhez – szótöredékekhez kapcsolt szemantikai információt is.

### 9.1 Finomhangolás

Általánosan kétféle módszer létezik az említett előre betanított modellek adott (specifikusabb) feladatra történő felhasználására: a jellemző-alapú (feature-based) és a finomhangoláson alapuló (fine-tuning) módszerek. Az előbbinél a feladatspecifikus architektúrát általános nyelvi reprezentációval egészítik ki, míg az utóbbinál a feladatspecifikus komponensek számát csökkentik, és a kívánt eredményt az előre betanított paraméterek finomhangolásával érik el. A BERT modellek általában a GPT nyelvi modellhez hasonlóan a finomhangolási megközelítést valósítják meg

Ahhoz, hogy egy „közérthetőségi osztályozó” elkészülhessen, elsőként a megfelelő nyelvmodellt volt szükséges kiválasztani. A jelen vizsgálat során ezt a huBERT (Nemeskey 2020) jelentette<sup>147</sup>, amely a BERT-base modell első magyar nyelvű implementációja, amely tehát szintén a BERT kapcsán említett finomhangolási megközelítést valósítja meg (Howard – Ruder 2018).

A finomhangolás lényegében a modell paramétereinek olyan változtatását jelenti, hogy a modell a lehető leghatékonyabban teljesítsen az adott megoldandó problémán. A modellteljesítmény mérésének gyakori módja ez esetben is a veszteségfüggvény (loss) kiértékelése mind a tréning-, mind a validálási halmazon. Ebből a célból az eredeti korpuszt ezúttal 80%-os train, 10%-os validációs és 10%-os teszhalmazra osztottam. Ezek közül a validációs halmazt a finomhangolási folyamathoz használtam, a teszhalmazt pedig a modell végső metrikáinak értékelésére ismét csak pontosság, fedés és F1 tekintetében.

## 9.2 Adatok előkészítése

A legtöbb nyelvmodell önmagában nem alkalmas klasszifikálásra; ennek a feladatnak a megoldásához egy minimális neuronháló architektúrát szükséges definiálni a modell köré, valamint a modell bementét az architektúrához mérten előfeldolgozni.

Elsőként definiálni kell a modell bementét, amely BERT modellek esetében jellemzően a hozzá tartozó tokenizáló által előállított token id-ekből, valamint az adott szöveget jellemző attention mask-okból áll.

A token id-k a huBERT modell által kezelt, és a finomhangolási folyamat során is felhasznált, szótöredék-szintű reprezentációkat jelentik. A BERT saját belső reprezentációjában a kontextusfüggő beágyazások nem klasszikus token-szinten, hanem szótöredékek szintjén (subword level) keletkeznek. A fastText-hez kapcsán már említetteknek megfelelően ez a működés segíti a modellt többek között a korábban még nem látott (Out Of Vocabulary - OOV) szavak kezelésében is azáltal, hogy az ismeretlen szavakat is képes lesz a modell „előállítani” azok alkotórészeiből. Ezt a működést a 16) példa szemlélteti. Az a) részben egy hagyományos tokenizáló sematikus kimenete látható, míg a

---

<sup>147</sup> A nyelvmodell és kapcsolódó tokenizáló szabadon elérhető a Hugging Face felületén: SZTAKI-HLT/hubert-base-cc

b) rész a huBERT tokenizálójának output-ját mutatja be az „A Szegedi Egyetemen folytatott doktori képzések során született dolgozatok tárháza.” mondatra vonatkozóan.

- 15) a) ['A', 'Szegedi', 'Egyetemen', 'folytatott', 'doktori', 'képzések', 'során', 'született', 'dolgozatok', 'tárháza', '.']
- b) ['A', 'Szegedi', 'Egyetemen', 'folytatott', 'doktori', 'képzések', 'során', 'született', 'dolgozat', '##ok', 'tár', '###háza', '.']
- c) [2, 2038, 10934, 16346, 9784, 16834, 20704, 3001, 5173, 18384, 2040, 2915, 7872, 4575, 3]

A „dolgozatok”, valamint a „tárháza” szavak esetében jól látszik, hogy a modell azokat nem egyben, hanem két-két töredékre bontva ismerte fel. Minden egyes ilyen töredék rendelkezik egy egyedi azonosítóval, amely a modell tanítása során alakult ki; ezek a 15). c) -ben kísérhetők figyelemmel. Az „a” névelő például a 2038-as token id-vel rendelkezik, míg a „tárháza” szó a 7872 („tár”) és a 4575 („###háza”) töredékekből konstruálható meg.

A 15) c) -ben szemléltetett vektorban annak első tagja (2-es token id-vel) a [CLS]-tokent, míg az utolsó („3”) a [SEP] tokent jelenti. Előbbi egy olyan speciális token, amely nem a konkrét mondatból származik – valójában minden szövegegység elejére beszúrja a tokenizáló -, hanem a klasszifikáláshoz szükséges jellemzők akkumulálása a célja a modell számára. Utóbbi funkciója pedig, hogy szekvencia klasszifikálás esetén az egyes szövegegységek elválasztásában segítse a modellt (esetünkben ez jelzi a mondatok végét).

A modell bemenetét fix hosszúságú vektorok kell, hogy képezzék. A maximális méret, amely a modellnek megadható egy-egy szekvenciára nézve, 512 token (subword-szinten értve), e fölött a bemeneti adatok csonkolása történik<sup>148</sup>. A rövidebb mondatokat a tokenizáló kiegészíti „üres” tokenekkel a megadott hosszúságra. Ahhoz, hogy a ténylegesen az eredeti adatból származó, és a csak kitöltésre használt tokenek megkülönböztethetők legyenek, az attention mask nyújt segítséget. Ez valójában egy olyan bináris vektor, amely egyeseket tartalmaz azon vektorbeli pozícióknál, ahol a szöveg eredeti szavai szerepelnek (továbbá a [CLS] és [SEP] tokenek pozícióján is), és nullákat ott, amelyek csak a fix hosszúság elérése érdekében kerültek be. A fenti példánál maradva, ha ezt a hosszt például 20-ra állítjuk be, akkor az attention mask a 15) példában már vizsgált mondatra nézve a következők szerint alakul:

---

<sup>148</sup> A korpusz esetében azon mondatok aránya, amelyek 512-nél több szótöredék szintű tokent tartalmaztak kevesebb mint 0.5% volt, így ez nem eredményezett jelentős adatvesztést.

16) [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

A 16) példában az első 15 token a mondat valódi szótöredékeit jelenti, míg a fennmaradó 5 pozíció tehát a fix hosszúság miatt került be (ezeket a klasszifikáló nem veszi figyelembe a finomhangolás során).

A bemenetek meghatározása után a huBERT modell megkapja minden egyes szöveghez a token id + attention mask vektorokból álló párosításokat (ezek jelentik a bemeneti – input réteget).

A modell ezt követően egy 768 dimenziós vektort ad vissza eredményként, amelyet tovább „örökíthetünk” egy tetszőlegesen választott neuronhálónak. A jelen esetben ez egy 2 neuronból álló kimeneti réteg, amelynek az a szerepe, hogy a one-hot enkódolt címkékhez a modell által prediktált valószínűségi eloszlást megjelenítse.

A modellek esetében alkalmazott kötegméret (batch size), azaz, hogy a modell egy-egy iteráció során milyen számosságú tanító példát láthat egyszerre, GPU RAM takarékosági okokból 16-os értékre került beállításra<sup>149</sup>. A kimeneti réteg aktivációs függvényeként a 'relu' került kiválasztásra.

## 9.2 Eredmények

A performancia növelésére (a batch méretének változtatásán túl) további lehetőség lett volna, hogy a huBERT esetében nem mind a 12 rejtett réteg marad tanítható, hanem például csak az utolsó 4. Ezt a megközelítést azonban a predikciós teljesítmény maximalizálása érdekében a jelen kísérlet nem használta ki. A predikciók helyességét a BinaryCrossEntropy függvény követte nyomon. Ennek célja, hogy értékelje a modell által alkotott döntések helyességét az alapján, hogy azok mennyiben térnek el az elvárt címkéktől. A függvényt a tanító és a validációs halmazon is kiértékelve meghatározhatjuk azt az optimális epoch számot, amely még nem vezet generalizálási hibához.

A 14. táblázat ezen loss függvénynek a tanító és validációs halmazokon mért értékeit szemlélteti epochonként, a vonatkozó accuracy-vel együttesen. Általánosságban igaz, hogy ameddig az accuracy értékek nem válnak lényegesen magasabbá a tanítóhalmazon, mint a validációs halmazon, valamint a loss érték nem lesz lényegesen alacsonyabb a tanító

---

<sup>149</sup> A kísérletek során az átlagos GPU RAM használat ezen érték mellett ~32 GiB volt.



halmazon a validációshoz képest, addig a modell jó eséllyel nem vesztette el a generalizálási képességét.

| <b>Epoch</b> | <b>Validation loss</b> | <b>Validation Accuracy</b> | <b>Train loss</b> | <b>Train Accuracy</b> |
|--------------|------------------------|----------------------------|-------------------|-----------------------|
| <b>1</b>     | 0.5138                 | 0.7424                     | 0.6225            | 0.6396                |
| <b>2</b>     | 0.5374                 | 0.7332                     | 0.5088            | 0.7485                |
| <b>3</b>     | 0.5410                 | 0.7393                     | 0.4280            | 0.8004                |
| <b>4</b>     | 0.6411                 | 0.7127                     | 0.3591            | 0.8416                |
| <b>5</b>     | 0.6833                 | 0.7119                     | 0.2931            | 0.8737                |

14. táblázat: A tanító és validációs halmazokon megfigyelt accuracy és loss értékek.

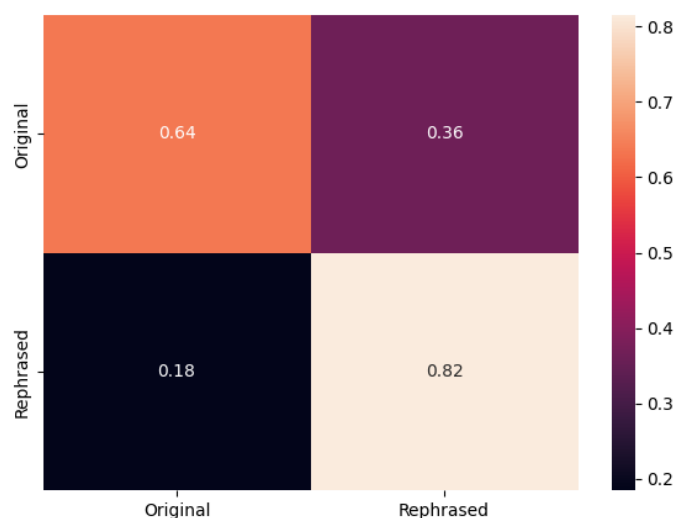
Az eredmények alapján az látható, hogy a 3. epochig a validációs halmazon mért loss érték görbéje alacsony meredekségű, majd a 4. epochtól kezdően meredek emelkedésnek indul. Ezzel párhuzamosan a validációs halmaz accuracy értéke is itt lesz monoton csökkenő. Emiatt ez az epoch szám tekinthető egy (közel) optimális megállási feltételnek a modell finomhangolása során.

A végső modell ezek alapján tehát 3 epochon át került finomhangolása, majd pedig ennek kiértékelése történt meg a korábban leválasztott 10%-nyi adatot tartalmazó teszhalmazon. Az így nyert eredményeket a 15. táblázat ismerteti.

|                      | <b>precision</b> | <b>recall</b> | <b>f1-score</b> | <b>support</b> |
|----------------------|------------------|---------------|-----------------|----------------|
| <b>Átfogalmazott</b> | 0.71             | 0.79          | 0.75            | 647            |
| <b>Eredeti</b>       | 0.76             | 0.67          | 0.71            | 666            |
| <b>accuracy</b>      |                  |               | 0.73            |                |
| <b>macro avg</b>     | 0.74             | 0.73          | 0.73            | 1313           |
| <b>weighted avg</b>  | 0.74             | 0.73          | 0.73            | 1313           |

15. táblázat: Az optimális modellel mért metrikák a teszhalmazon.

Az itt elért macro átlag F1 mintegy 0.11-gyel megelőzi a fastText kontextusfüggetlen beágyazásaival elért legjobb értéket (0.6165).



13. ábra: : Normalizált konfúziós mátrix a finomhangolt modell teljesítményének értékelésére.

A modell teljesítményét egy másik aspektusból szemlélhetjük konfúziós mátrix, valamint ROC-görbe számításával. Mindkettő alkalmazása a gépi tanulási kísérletek során a kiértékelt modellek hibáinak megjelenítésére szolgáló szabványos megoldás, amelyek lényegileg egymással ekvivalens információt hordoznak (Zheng 2019).

Alapértelmezés szerint egy konfúziós mátrixban annak y-tengelye az adatok valós címkéit, míg az x-tengelye a modell által prediktált címkéket jelöli, a mátrix elemei pedig az egyes címkék szerint helyesen vagy tévesen besorolt elemek számát mutatják. Ennek egy változata az, amikor a mátrix elemeit valamilyen módon normalizáljuk. Ahhoz, hogy az eredeti mátrixból egy normalizált változatot kapjunk, egyszerűen eloszthatjuk az egyes sorelemeket a teljes sor összegével. Mivel minden sor az egyes osztálycímkék tényleges értékeinek teljes számát jelenti, ezen a módon a végleges normalizált mátrix minden egyes pozíciójában egy százalékos érték fog megjelenni. Bináris esetben ez azon értéket jelenti, amit a modell az adott címkéhez mérten helyes osztályozott a főátló elemei esetében, a fennmaradó pozíciók pedig a tévesen osztályozott példák százalékos arányát mutatják meg. Nagy elemszámú halmazok esetében ez szemléletesebb lehet a pusztán számszerűsített értékek ábrázolásánál.

A 13. ábra a 3 epochon át tanított modellnek az imént leírt módon normalizált konfúziós mátrixa. Ez alapján a már átfogalmazott mondatok azonosítása mintegy 82%-ban helyesnek bizonyult, azonban a dolgozat célkitűzését figyelembe véve kevésbé szerencsés módon az átfogalmazásra váró mondatok azonosítása ehhez képest csak 64%-ban volt helyes.

## 10. Gépi tanulási kísérletek összefoglaló értékelése

A két prediktálandó osztály alapján számított átlagos F1 érték alapján a leggyengébben a fastText kontextusfüggetlen beágyazásai teljesítettek (~0.62), ezt követte a SVM-mel elérhető 0.68, a leginkább megbízható teljesítményt pedig a huBERT finomhangolása hozta (~0.73). Utóbbi rendelkezik azzal az előnyös tulajdonsággal is, hogy az átfogalmazásra szoruló mondatokat 0.76-os pontosság mellett képes prediktálni, ami jelentősen jobb, mint az SVM-mel stabilan (alacsony szórás mellett) elérhető 0.72.

Ha tehát a 7 – 9. fejezetekben bemutatott modelleket tekintjük, minden szempontól (a várakozásoknak megfelelően) a kontextusfüggő beágyazásokat alkalmazó, finomhangolt huBERT bizonyult a leginkább ígéretes megoldásnak. A modelleknek az iménti bekezdésben vázolt sorrendisége meglepő, tekintettel arra, hogy a történetileg sokkal korábbi SVM jelentősen jobb eredményt ért el, mint a fastText. Hasonlóan szokatlan az is, hogy a BERT alapú modell csak mintegy 0.05-tel teljesített jobban a hiperparaméter hangoláson átesett SVM modellnél.

Erre egyszerű magyarázat lehet a közérthetőségi osztályozás problémájának inherens nehézsége. Másrészt a korpusz tüzetes tanulmányozása után nyilvánvalóvá vált, hogy az abban elvégzett javítások nem pusztán a PLM javaslatain alapultak, de sok esetben „egyszerű” helyesírási korrekciók is történtek. Erre konkrét példát a 11. fejezet ismerteti. Tény, hogy egy helyesírásban nem helyes szöveg kevésbé lehet közérthető, mint egy helyes. Az azonban a jelen adatok alapján nem világos, hogy a helyesírási ÉS közérthetőségi szempontú változtatások a szövegben valójában egy, vagy két osztályozási feladatot jelentettek a modell számára. Utóbbi esetben az adatok „keveredése” befolyásolhatta a huBERTPlain teljesítményét.

A modellek eredményei ettől függetlenül részben igazolják a iii. kiinduló hipotézis állítását. Eszerint közérthetőségi szempontú szövegosztályozásra lehetséges gépi tanult modell létrehozása, mégpedig (és ez kulcsfontosságú) megfelelő hatásfokkal. A betanított neuronháló alapú modellek, valamint egyszerűbb gépi tanuló algoritmusok teljesítményének értékelése ugyanakkor megfelelő viszonyítási pont hiányában rendkívül nehézkes. Éppen ezért annak eldöntéséhez, hogy az elkészített modellek hatékonysága elegendően jó-e egy gyakorlati alkalmazáshoz, a 11. fejezetben keresem a választ. Itt két szakértő által elvégzett kézi annotációval vetem össze a legjobban teljesítő huBERTPlain modell eredményeit.

Megemlítendő, hogy míg az olyan (klasszikus) NLP feladatok esetében, amilyen például a szentiment elemzés, a szakirodalomban összehasonlítási alapul szolgáló adatok széles spektruma áll rendelkezésre, addig közérthetőségi osztályozásra tekintettel hasonlóról nem beszélhetünk; az ilyen jellegű vizsgálatok szórványosak és nem kötődnek kifejezetten a jogi doménhez.

Kanungo – Orr (2009) például a keresőmotorok elterjedésének kezdetén a találati listákban megjelenő oldalak automatikus összefoglalóinak olvashatóságát elemezte. Vizsgálatukban kézzel címkézett tanítóadatokból nyertek ki jellemzőket (feature), pl. Flesch – Kincaid olvashatósági pontszám, szó(tag)szám, ellipszis jelenléte az összefoglaló elején vagy végén stb. Az így kapott jellemzőkön ezután döntési fákat tanítottak be, és megvizsgálták, mely jellemzők korreláltak leginkább az adatok kézi címkézésével.

A korszerűbb megoldások ugyanakkor főleg az egészségügy/orvostudomány területére jellemzőek, amely a jog mellett a másik olyan jelentős szféra, ahol a laikus és professzionális nyelvhasználat közötti különbség a leginkább szembeűnő. Ezek azonban célkitűzésükben jelentősen eltérnek az értekezés témájától. Erre jó példa Guo (et al. 2021) írása, amelyben a szerzők orvosi cikkek tartalmi összefoglalójának közérthető átiratait kísérelték meg létrehozni többek között a BART (Mike et al. 2020) nyelvgenerálásra alkalmazható modell segítségével. Kísérletükhöz olyan összefoglalókat használtak fel, amelyek esetében rendelkezésre állt egy a tudományos közösségnek szóló, valamint egy laikusok számára írt verzió is ugyanazon cikk kivonatából. A korábban szakértők által laikusok számára írt összefoglalókat, és a BART segítségével generáltakat több metrika mentén is összevetették, amelyek közül fontos kitétel volt az emberi visszajelzések begyűjtése is<sup>150</sup>. Ennek során a kivonatokat 4 fő szempont szerint kellett az annotátoroknak értékelnie egy-egy 1-5-ig terjedő skálán, ahol 5 jelentette a legjobb értékelést a következők szerint:

---

<sup>150</sup> Az automatikus értékelési metrikák között szerepelt például a BART által generált összefoglalók ROUGE metrikák szerinti összevetése az eredeti szövegekkel, valamint azok Flesch – Kincaid olvashatóság szerinti értékelése is. A ROUGE metrikák (ROUGE-1, ROUGE-2, ROUGE-L) például a gépi fordítással készült szövegek, vagy automatikus tartalmi összefoglalókat készítő szoftverek teljesítményének számszerűsítésére alkalmazott szokásos megoldások. ROUGE-1 esetében a pontosság és fedés értékek egyszerűen a generált, és a kézzel készített összefoglalók mondatainak közös szókészlete alapján számolható. ROUGE-2 esetében az összevetés alapját bigramok jelentik (amelyek esetében a szavak sorrendje is kötött, tehát az eredeti előfordulási sorrendnek megfelelő kell, hogy legyen). A ROUGE-L ezekhez képest annyiban tér el, hogy itt az értékelés alapját a leghosszabb közös részszekvencia jelenti (LCS). Ez a mondatoknak azon szóhalmaza, amelyek (akár megszakításokkal, de) azonos sorrendben fordulnak elő mind a generált, mind pedig az elvárt kimenetben. A ROUGE metrikák hátránya, hogy a jelentéstartalmat nem, csak a felszíni szerkezetet képesek figyelembe venni.

- grammatikalitás: nyelvtanilag helyesek-e a kivonat mondatai,
- jelentésmegőrzés: a szakértőknek szóló összefoglalóban szereplő hasznos információk mennyiben őrződtek meg a laikusoknak szóló kivonatban,
- érthetőség: könnyebben érthető-e a laikusoknak szóló, mint a szakértői kivonat, illetve
- lényeges információk helyessége: a laikusoknak szóló kivonat hitelessége az eredeti, szakértői változathoz viszonyítva.

A szerzők fő következtetése szerint, habár az automatikusan generált összefoglalók hitelesen megőrizték a szakértői kivonatok főbb tartalmi elemeit, érthetőségük nem feltétlenül javult, illetve az általuk készített modellek jelenleg még nem alkalmasak automatizált közérthető kivonatok készítésére. Megjegyzik továbbá, hogy ugyanakkor az általuk finomhangolt BART modell esetében már tetten érhető néhány olyan típusú változtatás, amelyet a szakértők is előszeretettel alkalmaznak a szövegek egyszerűsítése során (jellemzően a mondat szerkezet egyes módosításai). Az ilyen változtatások azonban több esetben a mondat lényegi jelentésének módosulását is maguk után vonták, amely (különösen olyan területeken, mint az orvostudomány vagy éppen a jog) nem megengedhető hibafaktor.

A helyzet természetesen még összetettebb, ha nem tartalmi kivonatokot készítünk, de a teljes szöveget szeretnénk újraalkotni közérthetőbb változatban.

Mindent összevetve a dolgozat elkészültekor (a szerző tudomása szerint) nem állt rendelkezésre olyan adat, amellyel az itt elkészített gépi tanulási modellek teljesítménye közvetlenül összemérhető lett volna. Ismert ugyanakkor, hogy a nagy nyelvmodellekre alapozott osztályozók tipikusan a legjobb teljesítményt nyújtják a rendelkezésre álló algoritmusok közül a legtöbb NLP feladat esetén; (feladatfüggően) akár 0.8 – 0.9 feletti F1 értéket is elérve. Az értekezés keretében tanított modell ehhez képest jelentősen alacsonyabb teljesítményt ért csak el, aminek a már említetteken kívül is több lehetséges oka van.

Problémát okozhatnak például a mondatsegmentálásban bekövetkező hibák. Ezek közül több olyan tipikus hiba kijavítása utólagosan megtörtént (vö. 6.3.2 fejezet), ahol a szegmentáló valamilyen fals mintázat mentén kezdett új mondatot, azonban a teljes (tízezres nagyságrendű) adathalmaz kézi validálása nem volt megoldható.

Összetettebb probléma, hogy a konkrét korpuszban sok esetben jellemzők a mondatokon végrehajtott olyan módosítások, amelyek esetében az eredeti, és az újrafogalmazott mondatot például egyetlen szó vagy kifejezés cseréje választja csak el. Ilyen lehet a funkcióiágék cseréje aktív szerkezetre, de akár a 17) a) – b) mondatpárjai is jól szemlélteti a jelenséget.

17) a) Eredeti: „**Jelen** felhasználói kézikönyv eddigi verziói a következő időpontokban jelentek meg:”

Átfogalmazott: „**A** felhasználói kézikönyv eddigi verziói a következő időpontokban jelentek meg:”

b) Eredeti: „Az alkalmazás a “NAV-azonosító alkalmazás”-ra **való keresést követően** érhető el.”

Átfogalmazott: „Az alkalmazás a “NAV-azonosító alkalmazás”-ra **keresve** érhető el.”

Habár a BERT egyik erőssége éppen az egyes szavak jelentésének reprezentálási képessége a kontextus függvényében, kérdéses, hogy egyetlen határozott névelő cseréje (amely az érthetőséget valóban javítja) mennyire jelent tetten érhető különbséget a modell számára.

Ez egyben magyarázat lehet arra is, hogy a TF-IDF vektorizálással, és uni- illetve bigramokkal dolgozó SVM miért ért el kimagasló eredményt az adatok osztályozásában. Feltételezve ugyanis, hogy a jogi szövegek közérthetőre alakítása kapcsán alapelvárásként tekintett jelentésbeli statikusság<sup>151</sup> nem sérült, ebből következik, hogy a lexikai értelemben csak minimálisan (egy-két szó vagy kifejezés erejéig) megváltoztatott mondatok osztályozásában jól teljesíthetnek az olyan algoritmusok is, amelyek valóban kizárólag a lexikai elemek, illetve kollokációk megléte, vagy hiánya alapján végzik el a kitűzött osztályozási feladatot.

Hasonló a helyzet azzal a tipikus esettel is, amikor eredetileg a főszövegben helyet kapó jogszabályhivatkozások lábjegyzetbe kerülnek, de a mondat maga változatlan marad. Egyrésztől nem tisztázott ezek zavaró, vagy éppen segítő szerepe a BERT belső reprezentációjában a jelentés modellezése során, másrészt pedig kérdéses, hogy egy mondat, amelyből kiemelünk egy szerkezetet, a fennmaradó részt pedig változatlanul hagyjuk, emiatt mennyiben fog eltérő jelentéssel bírni? Ennek kapcsán felmerül, hogy amennyiben a cél a

---

<sup>151</sup> Azaz az átalakított és eredeti mondatpárok azon tulajdonsága, hogy az átalakítás nem módosította a mondatjelentést, illetve az eredetileg kifejezett jogi norma tartalmát.

dokumentumok újraszerkesztése során éppen az eredeti jelentés megtartása, ez a szempont mennyiben ütközik a klasszifikációs célkitűzésekkel? Hasonló kérdésekre az egyes neurális hálók belső szerkezetének mélyebb ismeretében, esetleg külön kutatásban, empirikus úton lenne csak lehetséges választ találni.

Végző soron megemlítendő, hogy a finomhangoláshoz használható korpusz mérete nem feltétlenül lehetett elégséges az ennyire árnyalt jelentésbeli különbségek megragadásához a modell számára. Összetettebb NLP feladatok esetében nem ritka, hogy a felhasznált tanítóadatok százezres, vagy akár milliós nagyságrendben is rendelkezésre állnak, ilyen adatmennyiség azonban a dolgozat céljához illeszkedő szövegekből nem állt, és feltehetőleg a közeljövőben nem fog rendelkezésre állni.

A neurális háló alapú megoldásokkal kapcsolatos előzetes elvárás az volt, hogy azok (legalábbis bizonyos mértékben) képesek lehetnek nem csak szemantikai, de pragmatikai információkat is figyelembe venni. Annak kapcsán, hogy például a BERT nyelvmodell működése során milyen mértékben inkorporálja ezeket a faktorokat, a legegységesebb vizsgálatot mindeztáig Ettinger (2020) folytatta le.

## 10.1 Pragmatika

A jogi szövegek közérthetőségi szempontú felülvizsgálata során a fogalmazónak mindenekelőtt azt a világtudást és előismeretet kell rekonstruálnia, amellyel a befogadó (a NAV esetében például az „átlagos adózó”) rendelkezik / rendelkezhet. Ezt mi sem mutatja jobban, mint hogy a hivatal Közérthetőségi Programjában is deklarált cél „az egyszerű, adózói logika szerinti kommunikáció”. Ez valójában azt jelenti, hogy ebben a speciális, írott formában történő (tájékoztató anyagok esetén egyoldalú, email váltások során már kétoldalú) kommunikációban a társalgás koherenciáját a befogadó nézőpontjába való behelyezkedéssel lehet a leghatékonyabban fenntartani. Azt, hogy a gyakorlatban ez hogyan valósulhat meg, a legegyszerűbben Grice-nak a társalgások elemzése során lefektetett maximái kapcsán vizsgálhatjuk meg<sup>152</sup>.

---

<sup>152</sup> A pragmatikai kutatások az elmúlt több, mint fél évszázadban számos irányvonalon és több szempontból is vizsgálták a nyelvhasználatot, elég csak a retorika, a beszédaktus-elmélet és a diskurzus-elmélet területeire gondolni, de érintőlegesen ide köthető még a kommunikációelmélet, a narratológia vagy éppen a szociolingvisztika is. Annak oka, hogy a pragmatikai kutatásokkal a jelen értekezésben mindösszesen egyetlen dedikált fejezt foglalkozik, a dolgozat célkitűzésében keresendő. Ez ugyanis olyan algoritmizálható (de legalábbis a jelenleg létező nagy nyelvi modellek jelentésreprezentációs képességeivel jól megfogható)

Grice társalgásemeléte élesen megkülönbözteti egymástól a megnyilatkozás jelentését (szemantikáját), valamint mindazt a hozzáadott információt, amit a beszélő a megnyilatkozás lévén implikál. Ebben a keretben előbbi a konvencionális jelentés, valamint a befogadó általi kontextusfüggő egyértelműsítés segítségével ragadható meg, utóbbi pedig a társalgást irányító racionális elvek és maximák megértése segítségével (Grice 1967).

Nézőpontja szerint az ilyen hozzáadott jelentések (implikatúrák) „kiszámítása” gyakorlatilag a beszélő által elmondottakkal kapcsolatos általános ismereteken alapul, illetve az adott megnyilatkozás nyelvi és nyelven kívüli kontextusán, az általános világtudáson, valamint az általa kooperatív (vagy együttműködési alap)elvnek (Cooperative Principle) nevezett szempontrendszer figyelembevételén. E szerint „a társalgási hozzájárulásod legyen olyan, amelyet a beszélgetésnek abban a fázisában, amikor az történik, az adott beszélgetés elfogadott célja vagy iránya megkíván” (Grice 1989: 26).

Az együttműködési alapelvet a gyakorlatban leginkább négy maxima érvényesülése által szokás leképezni az alábbiak szerint (Grice 1989):

- Mennyiségi maxima: csak annyit mondj, amennyit a társalgás pillanatnyi céljai megkövetelnek, és ne légy a szükségesnél kevésbé vagy túlzottan informatív!
- Minőség maximája: próbáld hozzájárulásodat igazzá tenni! Ezt értelmezhetjük két aleset összességként is;
  - Ne mondj olyat, mint hamisnak hiszel!
  - Nem mondj olyat, aminek megfelelő bizonyítéka nem áll rendelkezésre!
- Kapcsolat – odatartozás maximája: Légy Releváns!
- Módozat – stílus (esetleg mód) maximája: Légy érthető! Ez a megkövetés szintén felbontható;
  - Kerüld a kifejezésmód homályosságát!
  - Kerüld a kétértelműséget!
  - Légy tömör!
  - Légy rendezett!

---

megoldások felderítését célozza, amelyek hatékonyan implementálhatók. Annak ellenére, hogy a számítógépes pragmatika (computational pragmatics) napjainkban már egyre inkább kutatott és gyorsan fejlődő terület, ilyen megoldások (talán éppen a feladat inherens komplexitása miatt) kevésbé állnak rendelkezésre. Éppen ezért a dolgozat célja e téren annyiban merül ki, hogy valamelyest felmérje a jelenleg rendelkezésre álló számítógépes nyelvészeti eszközök kapacitásainak határait, valamint, hogy rávilágítson azok hiányosságaira.



Ha a dolgozat keretét adó kutatást vesszük alapul, akkor a mennyiség maximájának szerepe marginális, hiszen a beszélő (esetünkben fogalmazó) nem dönthet arról, milyen mennyiségű információt kell közölnie, ha feltesszük, hogy a cél a teljeskörű tájékoztatás. A maxima relevanciáját esetleg a „felesleges szószaporítás” elkerülés jelentheti, azaz a tömör és világos megfogalmazás. A minőség maximája nehezen értelmezhető hivatalos kontextusban, hiszen amikor egy közhivatal tájékoztató anyagait olvassuk, eleve élünk azzal az előfeltevéssel, hogy az ott olvasottak igazak (itt: koherensek a hatályos jogszabályokkal, ügymenettel). A kapcsolat maximája szintén egyfajta alapvetés, amelyben nemigen képzelhető el mozgástér; ahogyan egy adózáshoz köthető úrlapon sem kell megadni például egészségügyi adatokat, úgy a konkrét téma köré szervezett hivatali tájékoztatók is (elméletben) csak releváns információkat tartalmazhatnak.

Mindezek alapján hivatalos szövegek újraalkotása során a legfontosabb szerep vélhetőleg a mód maximájának jut, hiszen ez írja elő a beszélő számára, hogy hozzájárulását a társalgásban résztvevők nézőpontjának figyelembevételével tegye meg.

Optimális esetben egy olyan modelltől, amely az átlagos befogadó szemszögéből könnyen érthető, vagy éppen további átalakításra szoruló szövegek elhatárolását végzi, joggal lenne elvárható, hogy a tanítási folyamat során valamiképpen képes legyen reprezentálni a címzett világtudását, de legalábbis figyelembe venni a szövegben megjelenő, mondathatáron túlnyúló kontextust is egy ilyen predikciós helyzetben.

Az, hogy a pragmatikai kontextust a ma létező nagy nyelvi modellek (LLM – Large Language Models) képesek-e egyáltalán figyelembe venni, és ha igen, milyen mértékben, nem egyértelműen eldöntött kérdés. Megjegyzendő, hogy nyelvi modellek esetében nem beszélhetünk olyan összetett kognitív folyamatok direkt meglétről, amelyek az emberek esetében például a pragmatikai alapú következtetéseket jellemzik. Ezek a modellek alapvetően statisztikai alapokon működnek, így a kérdés végső soron az, hogy a bennük létrejövő statisztikai alapú eloszlások mennyiben képesek reprezentálni az emberi intuícióhoz hasonló, szituációfüggő kimenetek előállításának képességét.

## 10.2 Az LLM-ek „pragmatikai kompetenciája”

A BERT 2018-as megjelenése sok szempontból áttörést jelentett a nyelvmodellezés, valamint a jelentések reprezentálása terén, amit a már említett transformer-alapú modellek elterjedése és az alkalmazásukkal elért *SOTA* eredmények is jól példáznak.

Azt, hogy az ilyen modellek (specifikusan a BERT eredeti architektúra variánsai) mennyiben képesek „egyszerű” szemantikai jelentéseken túli jelentéskomponenseket tárolni és alkalmazni predikciós időben, Ettinger (2020) vizsgálja részletesen. Kutatásában azon alapvetésből indul ki, miszerint az előtanított nyelvi modellek eredményei azt sugallják, hogy ezek a tanítás során értékes, generalizálható nyelvi kompetenciát sajátítanak el, arról azonban kevés empirikus adat áll rendelkezésre, hogy pontosan milyen faktorokra is terjed ki ez a fajta kompetencia. Módszertana szerint a nyelvmodellek nyelvi kompetenciáinak mibenléte (vagy ezen kompetenciák hiánya) kimutatható célzott tesztek alkalmazásával. A választott tesztek a pszicholingvisztikában bevett kérdéssorokon alapulnak és kiterjednek a hétköznapi tudás (commonsense), pragmatikai alapú következtetés (pragmatic inference), szemantikai szerepek (semantic roles), eseményekkel kapcsolatos tudás (event knowledge<sup>153</sup>), szemantikai kategóriákhoz tartozás (category membership), valamint a tagadás felismerésének vizsgálatára is.

A BERT tanítása során (a maszkolt nyelvi modell készítése által) alapfeladat volt, hogy a modell egy [MASK] tokennel ellátott, „üres” pozícióra képes legyen az odaillő szót prediktálni; ez képezi a vázolt tesztek alkalmazhatóságának sarokkövét. A tesztek során a modell feladata az egyes mondatok utolsó pozícióján elhelyezkedő, [MASK] tokenrel „kitakart” szó prediktálása volt<sup>154</sup>. Ilyen esetben a modell kimenetét szavak (tokenek) alkotják, a hozzájuk tartozó valószínűség értékekkel párhuzamosan, amelyek alapján azok egyszerűen sorba rendezhetők. Ezek közül a legnagyobb valószínűségi értékkel rendelkezőket tartja a modell a leginkább megfelelőnek az adott kontextushoz.

---

<sup>153</sup> Az általános eseményekkel kapcsolatokkal tudás (generalized event knowledge) azzal az elképzeléssel foglalkozik, hogy a lexikon nem szótárként, hanem inkább hálózatként szerveződik, ahol a szavak a lexikális tudás mellett a pragmatikai tudás által befolyásolt várakozásokat váltanak ki a következő inputra vonatkozóan. Ezért a mondatmegértés úgy fogalmazható meg, mint annak az eseménynek az azonosítása, amely a legjobban magyarázza a bemenetben használt nyelvi jeleket (Kuperberg – Jaeger 2016).

<sup>154</sup> Az utolsó mondatbeli pozíció különös jelentőséggel bír, mivel a BERT architektúrája bidirekcionális, azaz párhuzamosan képes információkat feldolgozni a szöveg elejéről és végéről indulva. Ez aránytalan előnyt jelentene a modell számára, amennyiben azonban a kérdéses szó a végső pozícióban helyezkedik el, úgy ez a fajta hozzáadott előny nem érvényesül, és a modell kimenete azonos információkon fog alapulni, mint ahogyan az ember megjósolná a hiányzó szót.

A teszteléshez választott modellek a BERT-base, illetve a BERT-large voltak, amelyek kiértékelése három adathalmazon valósult meg az alábbi szempontok szerint (Ettinger 2020: 37 - 39):

#### **CPRAG-102** (Federmeier – Kutas 1999)

- Problémakör: hétköznapi tudás, pragmatikai következtetések.
- Leírás: mondatpárok, ahol a második mondat erősen generikus, ezért nem támogat a hiányzó szóra vonatkozó konkrét következtetést. A mondatok továbbá nem tartalmazzák a kiegészítendő szót, ezért koreferencia feloldás helyett a modellt egyfajta „világtudás” felhasználására kényszerítik.
- Vizsgált jelenség: emberi kitöltők esetében a helytelen kiegészítések közel 0 valószínűséggel szerepeltek, az N400 teszt azonban mutatott némi várakozást olyan szavakra is, amelyek bizonyos szemantikai kategórián osztoztak a helyes befejezéssel. Utóbbi a neurolingvisztikában alkalmazott EEG-vel mért teszt, ami az agyi elektromos aktivitás intenzitásából von le következtetéseket az elhangzott szó kontextusba illeszkedéséről<sup>155</sup>. A kérdés ez esetben az, hogy a modell nagyobb eséllyel prediktál-e a helyes kitöltéssel valamely szemantikai kategórián osztozó (de a kontextusban mégis helytelen) szót?
- Példa:  
„He **complained** that **after she kissed him**, he couldn't get the **red color** off his face.”  
„He finally just asked her to **stop wearing that** [MASK].”  
Kiegészítés: lipstick (elvárt), mascara, bracelet (helytelen).  
A mindennapi tudás alapján a vörös szín a rúzs -szóhoz köthető, a pragmatikai következtetés lényege pedig, hogy a panasz ennek viseléséhez köthető.

#### **ROLE-88** (Chow et al. 2016)

- Problémakör: eseményekkel kapcsolatos tudás és szemantikai szerepek.
- Leírás: mondatpárok, ahol a helyes predikcióhoz szükséges, hogy a modell értelmezze a szintaktikai struktúrából következő szemantikai szerepeket, valamint

---

<sup>155</sup> A gyakorlatban a stimulust követő 3-500 ms időintervallumban bekövetkező, elektroencefalográfiával mérhető agyi aktivitás mértéke korrelál az elhangzott szónak az adott kontextusba illeszkedésével, azaz annak megjósolhatóságával. A kitérés mértéke kisebb, ha a hallott szó könnyen megjósolható volt, és jelentősen nagyobb, amennyiben az az adott kontextusban váratlan volt.

képes legyen az adott szerepekbe tartozó entitástípusok közötti tipikus kölcsönhatásokkal kapcsolatos eseménytudás felhasználására.

- Vizsgált jelenség: a mondatpárok esetében a megadott kiegészítés csak a főnevek adott sorrendje esetében helyes, így a modellnek érzékelnie kell ezt a különbséget.

- Példa:

„The restaurant owner forgot which **customer the waitress** had [MASK].”

„The restaurant owner forgot which **waitress the customer** had [MASK].”

Kiegészítés: served (helyes, majd helytelen).

Az elvárt kiegészítés csak az első mondat esetében lesz helytálló, ennek megítéléshez a modell különbséget kell, hogy tegyen a mondatok között a bennünk szereplő főnevek sorrendje (és a sorrendből következő szemantikai szerepek) alapján.

### NEG-136 (Fischer et al. 1983)

- Problémakör: negáció jelenlétének és jelentésének, valamint az egyes szavak szemantikai kategóriákhoz tartozásának érzékelése.
- Leírás: mondatpárok, ahol az igazságérték a negáció jelenléte szerint variábilis.
- Vizsgált jelenség: a negáció jelenlététől függően a lehetséges kitöltések az adott főnév egy hipernímája, vagy egy, a főnévhez nem kapcsolódó másik fogalom (negatív esetben).

- Példa:

A robin **is** a [MASK].

A robin **is not** a [MASK].

A vörösbegyőről szóló első mondat esetében az elvárt kiegészítés valamely hiperníma (pl. bird). Amennyiben a második mondat esetében is ezek lesznek a legvalószínűbb predikciók, az a modell érzéketlenségét mutatja a negáció jelenlétére.

Az egyes adathalmazok kiértékelése során Ettinger (2020: 39-45) megállapította, hogy a hétköznapi tudás, valamint a szemantikai szerepek érzékelésének tekintetében (CPRAG-102) a BERT modellek valóban figyelembe veszik mind a két mondat által jelentett kontextust. Ezt a szerzők az egyes mondatok szavainak véletlenszerű összekeverésével, illetve a második mondatból egy (a [MASK] tokenhez képest) 2 sugarú kontextust kivéve valamennyi szó eltávolításával mérték le; mindkét esetben a predikciós pontosság jelentősen csökkent. A hasonló szemantikai kategóriába eső lehetséges kitöltések esetében megfigyelhető volt, hogy a modellek nagyobb valószínűséget jósolnak a helyes kitöltés

javára. Elemezve azonban a helyes és helytelen kitöltésekhez a modell által hozzárendelt valószínűségek közötti különbségeket kiderült, hogy azok több esetben is csak marginálisak voltak, azaz az eredmények nem tükröztek olyan egyértelmű és biztos választást, mint az emberi válaszadók esetében. A legtipikusabb hiba ezen mondatpárok esetében az volt, hogy a modell magas valószínűségi értékeket rendelt olyan lehetséges kitöltésekhez is, amelyek egyetlen mondat kontextusában még helyesek lehettek volna, a kettőt összeolvasva azonban már nem.

A szemantikai szerepek, illetve az eseményekkel kapcsolatos tudás felhasználása szempontjából (ROLE-88) a BERT modellek legtöbbször helyesen következtettek a prediktálandó szóra a főneveknek a mondaton belül elfoglalt pozíciója alapján. Ettinger megvizsgálta azt is, hogy a mondaton belüli alany vagy tárgy eltávolítása mennyiben befolyásolta az egyes modellek predikciós pontosságát. Az általános tapasztalat szerint az alany vagy állítmány eliminálása külön-külön nemigen befolyásolta a pontosságot a BERT-base, azonban jelentősen rontotta azt a BERT-large esetében. Ez arra enged következtetni, hogy míg a BERT-base inkább a teljes A-T struktúrától teszi függővé a kapcsolódó ige predikcióját, addig a BERT-large mindkét komponensről külön-külön is jelentős függést mutat. A modellek szenzitivitása itt is hasonlóan alakult, mint a CPRAG-102 esetében, a helyes és helytelen predikciókhoz társított valószínűségek közötti eltérés sokkal kisebb volt, mint az emberi kitöltőre jellemző.

A tagadás kezelése kapcsán (NEG-136) a modellek többé-kevésbé érzéketlennek bizonyultak. Ennek oka leginkább abban volt keresendő, hogy pusztán a kiegészítendő szó elhagyása esetén az előtte elhelyezkedő pozícióban meghagyott névelő miatt a modell azonos predikciókat adott afirmatív és negatív esetben is (0% volt a helyes kitöltések aránya negatív esetben, míg 100% afirmatív esetben). Ez arra utal, hogy a szemantikai viszonyokat mindkét modell ignorálta, és minösszesen a szűk kontextus alapján választotta ki a mondatok végére illeszkedő szót.

Az eredményeket áttekintve az látszik, hogy mindkét BERT modell esetében a hétköznapi tudás, pragmatikai következtetések terén nagyjából az emberihez hasonló predikciók születtek (73%, illetve 79% pontossággal a BERT-base és BERT-large esetén). Ha azonban a helyes és helytelen kitöltések közötti valószínűségek elvárt különbségéhez egy minimális küszöbérték is meghatározásra került (0.01%, amely az emberi kitöltőkéhez képest

szignifikánsan alacsonyabb volt), akkor ez az eredmény jelentősen leromlott (44% illetve 58% -ra).

Szemantikai szerepek tekintetében a helyes predikciók aránya a BERT-large esetében volt a legmagasabb, a BERT-base esetén valamelyest alacsonyabb. Feltételezve, hogy a predikció csak akkor helyes, ha az elvárt kitöltés ténylegesen a legnagyobb valószínűségi értéket kapja a modell kimenetében, ez az érték mintegy 14,8% és 13,6% volt. Megengedőbb esetben, feltéve, hogy a modell kimenetében levő 5 legnagyobb valószínűségű elem között keressük a helyes kitöltést, ezek az értékek továbbra is csak 37,5%, illetve 27,7% szerint alakultak.

A dolgozat szempontjából legfontosabb következtetés szerint Ettinger (2020: 45-46) megállapítja, hogy a vizsgált modellek pragmatikai következtetéseket kizárólag egyszerűbb esetekben képesek figyelembe venni és a helyes-helytelen kitöltések valószínűségei közötti különbség az esetek többségében így is csak marginális, következésképpen a modell ilyen esetekben sem „magabiztos” a kiegészítendő szó prediktálása során.

Tekintettel a vizsgált modellek által elkövetett egyéb, gyakran triviális hibákra (például a negáció kezelése kapcsán) szembevetendő a különbség, amely a nyelvi modellek működési mechanizmusa, valamint az emberek nyelvi feldolgozási folyamatai között húzódik. Előbbiek alapvetően statisztikai alapon működnek, vagyis olyan kontextuális nyomokat fognak felhasználni a predikciók megalkotásához, amelyek a betanítási folyamat során a legvalószínűbben vezettek helyes eredményre. Ezzel szemben az emberek főként a jelentés „kiszámítását”, kontextusba helyezését, valamint az igazságérték megítélését végzik el hasonló interpretációs helyzetben. Ez a fajta strukturális eltérés alapjaiban kérdőjelezi meg a pragmatikai jellegű információk figyelembevételének képességét még egyszerűbb esetekben is, nem beszélve a befogadó lehetséges világtudásának reprezentálásáról, amely egy közérthetőségi szempontú intralingvális átfordítás esetében esszenciális jelentőséggel bír. Ez egyben jól rávilágít arra is, miért alkalmatlanok a jelenlegi NLP eszközök az olyan kreatív folyamatok teljes automatizálására, mint amilyen a jelen értekezés témájául választott felülvizsgálat a jogi / hivatalos szövegek esetében.

Ennek ellenére azonban az eredmények azt sugallják, hogy a nyelvmodellek által elsajátított információk (még ha azok mibenléte részben tisztázatlan is) alkalmasak a szakértői munka támogatására egy alacsonyabb szinten megvalósított rendszerben. Ennek a hatékonyságnak a mértékét a következő fejezet tárgyalja részletesen.

### 10.3 Kitekintés: generatív mesterséges intelligencia

Generatív mesterséges intelligencia alatt olyan gépi tanuló algoritmusokat értünk, amelyek képesek például szöveges tartalmakból, képekből, vagy hanganyagból tanulni annak érdekében, hogy önmaguk is képesek legyen teljesen új tartalmakat létrehozni. Ez élesen eltér a diszkriminatív mesterséges intelligenciától (mint például a korábban már bemutatott SVM), amelynek az a feladata, hogy a kapott példák osztályozása során felmerülő döntési határokat tanulja meg (Jiao et al. 2022: 212).

A ma létező LLM-ek, mint az OpenAI fejlesztésében álló GPT széria tagjai, például a cég legutóbbi, GPT-4-es modellje (OpenAI 2023) vagy a Google által épített Gemini modellcsalád (Google 2023) tagjai egy fejlődési ív legújabb elemei. Ennek első kiugróan sikeres eleme a Generative Adversarial Network – GAN neurális hálózati architektúra (Goodfellow et al. 2020) volt. Ez voltaképpen a mai legmodernebb megoldások (pl. ChatGPT) egyfajta előképe volt, noha jelentősen más felépítést követett, mint az utóbbiak.

A mai generatív modellek, köszönhetően a robosztus betanítási fázisnak és hatalmas adatmennyiségnek, amelyen az előtanítási (pretraining) folyamat zajlik, minden korábbinál inkább alkalmasak például az emberi nyelvhasználat szimulálására, ezáltal pedig természetesnyelvi szövegek generálására is. Nem véletlen, hogy a ChatGPT megjelenésével a technológia viharos gyorsasággal ismertette meg emberek tömegeit a mesterséges intelligenciával, és konkrétan annak generatív nyelvmodellekkel foglalkozó ágával.

A generatív modellekről tudományos alapokkal következtetéseket levonni azonban több szempontból is nehézkes. A probléma egyik oldalát az jelenti, hogy az ilyen eszközök létrehozása, de sok esetben még a használata is rendkívül erőforrásigényes. Összehasonlításképpen, egy bert-base modell (amely architektúráján a huBERT is alapszik) finomhangolása ma már megoldható például a Google Colab<sup>156</sup> felületén ingyenesen is. A hasonló modellek alkalmazása során az egyik legnagyobb korlátozó tényező a szükséges hardveres kapacitás, főként grafikus számítási kapacitás (GPU) megléte. Egy bert-base modell (implementáció függően) nagyjából 12-15 GB GPU RAM-mal már finomhangolható speciális feladatokra (downstream tasks), mint amilyen a korábban bemutatott közérthetőségi osztályozás is. Ezzel szemben egyes források szerint<sup>157</sup> a korábbi generációs, GPT-3-as modellnél mindössze annak betöltése és futtatása mintegy 700 GB memóriát

---

<sup>156</sup> <https://colab.google/>

<sup>157</sup> <https://neuroflash.com/blog/gpt-4-parameters-rumors-and-forecasts/>

igényel. A jelenlegi trendek szerint láthatjuk, hogy minden új generációs, piacvezető modell legalább egy nagyságrenddel költségesebb, mint az elődje.

Már ennyi információ alapján is könnyen látható, hogy miért terjedt el az a nézet, miszerint a jelenlegi SOTA eljárások fejlesztése szinte kizárólag a technológiai óriásvállalatok privilégiuma.

Ez számos reprodukálhatósági problémát is felvet. A hasonló modellekkel kapcsolatban számos fejlesztési részlet (sok esetben még a pretraining-hez használt tanítóadatok forrása is) üzleti titoknak minősül, illetve maga a modell sem publikus, amely gátolja a tudományos közösség hozzáférését és a módszer megismerhetőségét. Részben ezzel is magyarázható, hogy a témában mindeddig nem jelentek meg alapos összehasonlító tanulmányok; a leggyakoribb elemzések főként a használat empirikus tapasztalataira alapoznak.

A fentiek alapján látható, hogy a finomhangolás (egyrészt a modellek üzleti felhasználása, másrészt a szükséges kapacitás miatt) a gyakorlatban vagy nem, vagy csak nehezen megoldható<sup>158</sup>. Ennek hiányában jórészt csak a modellek pretraining során elsajátított képességeire alapozhatunk. Ennek a fajta tudásnak a hatékonyabb felhasználásra természetesen léteznek technikák, amelyek a jelen állás szerint sokszor jelentősen képesek is növelni a feladatok során elért pontosságot.

Egy ilyen megoldás például a prompt engineering<sup>159</sup>. Prompt alatt olyan utasítást értünk, amely tipikusan a kérdés-válasz szekvenciákon tanított („chatbotként optimalizált”) nyelvmodellek instruálására alkalmas. Egy prompt legfontosabb szerepe, hogy a nyelvmodellt kontextusba helyezze (pl. szerepekkel ruházza fel az elvárt válasz stílusára vonatkozóan), illetve természetesen, hogy benne az elvárt működésnek a lehető leginkább adekvát leírása szerepeljen. A prompt engineering olyan (főként gyakorlati tapasztalatokon alapuló) technikák alkalmazása, amelyek segítségével a LLM-ek képességei és korlátai felmérhetők, és működésük a konkrét feladatra szabva pontosabbá tehető.

---

<sup>158</sup> A szigorúan vállalati kézben lévő modelleken kívül számos nyílt forrású megoldás is rendelkezésre áll, mint például a Meta által fejlesztett Llama-2 modelles család (<https://www.hardware-corner.net/llm-database/Llama-2/>). Ezek bárki számára elérhetőek, azonban egyelőre (összehangolt vizsgálatok hiányában) nem tisztázott, hogy hatékonyságban hogyan viszonyulnak a piacvezető megoldásokhoz. Ennek az a fő oka, hogy a nagy nyelvmodellek széleskörű elterjedése még nem történt meg. Az ipari alkalmazási lehetőségek még éppen csak formálódóban vannak, a tudományos vizsgálatok pedig nem hoztak még egyértelmű eredményt az új generációs LLM-ek valós kapacitásaira vonatkozóan.

<sup>159</sup> <https://www.promptingguide.ai/>



A technika működését konkrét példákon keresztül a IV. Függelék mutatja be. A legfontosabb tapasztalat, amely ezek alapján leszűrhető, hogy a ChatGPT-től kapott válaszok természete alapvetően nem determinisztikus<sup>160</sup>, legalábbis a felhasználó szempontjából. Ez főleg a statisztikai nyelvmodellezésből következik. Ennek során a nyelvmodellek a nyelvgenerálás során valójában azt határozzák meg, hogy egy input szekvenciához mi a legvalószínűbb karaktersor, amely kimenetként tartozhat. Más szavakkal, a LLM-ek valójában nem helyes válaszokat, hanem olyan karakter sorozatokat adnak vissza, amelyek helyes válasznak **látszanak**.

Ez egyben a legnagyobb kockázat is a hasonló modellek kapcsán; a modell a legtöbb esetben nem fog olyan választ adni, amely expliciten eldönti, hogy nem tudta megmondani a helyes választ a kérdésre. Ehelyett (működéséből következőleg) legenerál egy karaktersort, amely a saját belső reprezentáció alapján a legközelebb áll a helyes válaszhoz. Ezt a jelenséget szokás hallucinációnak is nevezni, amelynek során a modell kimenete első ránézésre akár helyesnek is tűnhet, faktuálisan azonban a benne foglalt információk tévesek.

A prompt 1.-re adott válasz alapján feltételezhetjük, hogy a GPT-4 chatbotként finomhangolt változata esetében rendelkezésre áll az az alapvető tudás, hogy a PLM pontosan milyen elvek mentén javasolja szervezni a szöveget. Ez jó alapot adott néhány rövid, további vizsgálatnak.

A további prompt-ok esetében a szöveg, amelynek közérthetőre fogalmazását a chatbotnak meg kellett oldania, a Földhasználati törvény egy paragrafus<sup>161</sup> volt. Tekintve, hogy a jelen fejezet inkább csak egy betekintés abba, mire is képes ma a generatív mesterséges intelligencia, ezért a választott szöveg szándékosan egy összetett helyzet leírására szolgál. Ezen felül az idézett jogszabályhely megfogalmazásában is rendkívül bonyolult, például sokszor több bekezdés alkot egyetlen mondatot, számos összetett felsorolással.

A prompt 2 és 3 esetében egyenként több válasz kiértékelését is elvégeztem, jogi szakértő bevonásával. Több válasz generáltatása azért is volt szükséges, hogy a korábban már említett nem determinisztikus működés szemléltethetővé váljon. Tekintettel arra, hogy a chatbot minden, az adott beszélgetésben lezajlott interakciót figyelembe vesz egy-egy új válasz

---

<sup>160</sup> Az informatikában determinisztikusnak nevezünk egy algoritmust, amennyiben az azonos inputokra minden esetben azonos output-ot ad vissza. A generatív modellek esetében ezt nem célszerű úgy érteni, hogy azonos kérdésre a modell kimenete minden esetben azonos karaktersor lesz, azonban szemantikai azonosságot célszerű lehet elvárni az egyes válaszok között, vagyis, hogy azok jelentésükben megegyezők legyenek.

<sup>161</sup> 2013. évi CXXII. törvény a mező- és erdőgazdasági földek forgalmáról, 46§ (1)-(5) bek.

generálása során, ezért minden válasz generáltatása új chat ablakban történt, ezzel kizárva az esetleges interferenciákat.

A 2. prompt esetében a chatbot feladata az volt, hogy további instrukciók nélkül fogalmazza át a kapott szöveget közérthetőre. Mindhárom válasz esetén megfigyelhető volt, hogy a generált „közérthető” átfogalmazások inkább a hivatkozott törvényszöveg egyszerűbben érthető részének parafrázisait tartalmazták, mintsem egy a teljes normatív tartalmat könnyebben érthetőre fogalmazva.

Erre kiváló példa a 2-es prompt A válaszában a következő rész:

„Ezenkívül különleges esetekben, például ha valaki állattartáshoz akar földet bérelni, vagy bizonyos mezőgazdasági termékeket termelni, akkor előnyben részesülhetnek.”

A probléma az ilyen jellegű egyszerűsítésekkel az, hogy az „egyedülálló esetek” kategóriája pontosan a jogszabály azon elemeire utal, amelyek még a joghoz értők számára is nehezebben értelmezhetők. Hasonló a helyzet a C válasz esetében is:

„A törvény részletezi még, hogy bizonyos különleges célú haszonbérletek esetén (például állattartás, ökológiai gazdálkodás, kertészet, szaporítóanyag-előállítás) ki jogosult előnyben részesülni, figyelembe véve a föld művelési ágát és a bérlet célját.”

A B válasz esetében egy érdekes jelenségre lehetünk figyelmesek. A 2.-es pont szerint:

„2. Utána következnek azok a szomszédos földművesek, akik szintén a helyszínen élnek.”

Érdekes kérdés már az is, hogy a *helyszínen* fogalma pontosan mit takar, az azonban még inkább, hogy az átfogalmazott szövegrész vélhetőleg a jogszabály 46§ (1) b) és c) pontjára utal, amely alapján

„b) az olyan földművest, aki helyben lakó szomszédnak minősül;

c) az olyan földművest, aki helyben lakónak minősül;”

előhaszonbérleti jog illeti meg, mégpedig **ebben a sorrendben**. Ez utóbbi információ teljesen hiányzik az átfogalmazott verzióból.

A 3. prompt esetében az előhaszonbérleti jog elsőbbségéről a következőt olvashatjuk:

„Elsősorban azoknak a földműveseknek és mezőgazdasági termelőszervezeteknek ad előnyt, akik a földtől legfeljebb 20 km-re laknak vagy ott van a mezőgazdasági üzemközpontjuk, és legalább 3 éve ott vannak.”

Az állítás ez esetben igaz, azonban nem pontos. A kimaradó esetek lehet, hogy a gyakorlat szempontjából esetleg marginálisak, azonban a szöveg a továbbiakban nem tér ki arra, hogy mik a kimaradó lehetőségek. Az eredeti jogszabályból minden kategóriában érintett olvasó megtudhatja, hogy pontosan hol áll sorrendben, ami azonban ebből a részletből nem derül ki. Az átfogalmazott változat további része is érdekes.

A jogszabály külön kiköti, hogy az előhaszonbérlet jogosultság az ott meghatározott sorrendben lép érvénybe. Az átfogalmazott változatban (továbbra is *A válasz*) a következőt olvashatjuk:

„Elsőbbségi jogosultak

- Az előző haszonbérlők, ha a közelben laknak vagy gazdálkodnak.
- A szomszédos földművesek, különösen ha helyben laknak.
- Mezőgazdasági termelőszervezetek, ha helybeli illetőségűek vagy a közelben van az üzemközpontjuk.

Különleges esetek

- Ha a föld több tulajdonos között van, a tulajdonostársak előnyt élveznek.
- Bizonyos tevékenységeket végző gazdálkodók, mint állattartó telepek üzemeltetői vagy ökológiai gazdálkodást folytatók, szintén előnyben vannak.”

Egyrésztől nem egyértelmű, hogy az átfogalmazott verzióban a felsorolás sorrendje megfelel-e a jogosultság sorrendjének, mivel a chatbot erről nem tesz állítást. Amennyiben azonban feltesszük, hogy az átfogalmazott szöveg logikája követi a jogszabályét, akkor a fenti felsorolás téves. A jogszabály alapján az itt „különleges eset”-ként hivatkozott esetek minden esetben megelőzik az „elsőbbségi jogosultak” pontban felsoroltakat. Ugyanez érvényes a B válasz esetében felsorolt „Speciális előnyök” alatti lehetőségekre is.

Összegezve általános tapasztalat volt, hogy chatbot működése sokkal inkább egyfajta absztraktív összefoglalásra emlékeztetett, mintsem valódi intralingvális átfordításra. Formailag ugyan jellemzően követte azokat a jegyeket, amelyeket a prompt elvárt (pl. a 3-as prompt esetében összefoglaló a szöveg elején, jól látható tagolás stb.) ennek azonban a tartalom látta kárát. Jellemző volt az is, hogy a generált válaszok mintha egy „láthatatlan”

karakterlimithez próbáltak volna idomulni, amely szintén okozhatta azt a fajta információ tömörítést, amely miatt végül információ veszteség következett be.

Habár a fenti példák csak szemléltetési célokat szolgálnak, az jól látszik, hogy a statisztikai nyelvmodellezés, amely a LLM-ek működésének alapját jelenti, milyen veszélyeket rejt. Jogi szövegek esetében a legvalószínűbb karaktersor egyáltalán nem biztos, hogy a megfelelő válasz lesz egy kérdésre. Nem véletlen, hogy a legtöbb szakmai tartalmú prompt esetében több chatbot is figyelmeztető üzenetet jelenít meg, hogy mindenképpen érdemes egy szakértővel is konzultálni a kapott információk hitelességét illetően.

## 11. Szakértői kiértékelés

A kipróbált gépi tanulási modellek eredményeinek értékelése különösen nehéz feladat tekintetbe véve, hogy a szakirodalomban hasonló próbálkozás mindeddig (a szerző legjobb tudomása szerint) nem történt.

Annak érdekében, hogy ez áthidalható legyen, a modellek tesztelésére használt (összesen mintegy 1313 elemű) adathalmazból 250 eredeti és 250 átfogalmazott mondat került kiválasztásra kézi validáláshoz több lépcsőben.

- Első lépésként az adatokban jelenlevő zaj detektálása történt meg kézi ellenőrzéssel. Az előzetes ellenőrzés ellenére (vö. 6.3.1 fejezet) maradtak olyan szövegegységek a tanítóadatban, amelyek hibás mondatsegmentálás eredményeképpen jöttek létre<sup>162</sup>. Ezek kizárása 1097 elemű halmazt eredményezett (~16.46% veszteség).
- Ezt követően 250 eredeti és ugyanennyi átfogalmazott mondat került be a kézi validációs halmazba.
- A kézi validációs halmazt ezt követően két részre osztottam, mindkettőbe 125-125 mondat került címkékként.

A validációs halmaz első része (125 átfogalmazott és 125 eredeti mondat) annak ellenőrzését szolgálta, hogy pusztán a kontextus nélküli mondatok<sup>163</sup> megítélése mennyire jelent nehéz feladatot a szakértők számára. A feladat itt mindössze egy „közérthető” vagy „nem közérthető” címke kiosztása volt. A halmaz második fele arról volt hivatott adatot szolgáltatni, hogy az előzetes gépi címkézés mennyiben segíti, vagy akadályozza a szakértői döntést az előzőekhez képest. Ezen mondatok esetében a felkért szakértők nem csak a mondat szövegét, de a finomhangolt huBERT modell által prediktált címkét is megkapták. A feladat ez esetben szintén annak eldöntése volt, hogy az adott mondat közérthető-e a jelen formájában.

---

<sup>162</sup> A leggyakoribb probléma ezek esetében már nem a túlszegmentálás, hanem az egybe nem tartozó mondatok összeillesztése volt. Adódtak vitás esetek is, amelyekben például a felsorolást bevezető mondat után a felsorolás minden pontja nagybetűs kezdéssel szerepelt, illetve szerkezete alapján tekinthető akár önálló mondatnak, akár csak felsorolási elemnek is. Ezeket jellemzően mondatvégi írásjel sem zárta. A kézi címkézést megkönnyítendő minden ilyen esetben a „mondatjelöltet” kizártam a validálandó mondatok közül, ezáltal a szakértőkhöz már csak a vitán felül helyesen szegmentált esetek kerültek.

<sup>163</sup> Ez a megközelítés amiatt volt célszerű, hiszen így a szakértőknek is pontosan azonos információmennyiség állt rendelkezésére, mint a modellnek a betanulás során.

Az értékelésre felkért szakértők a SZTE-ÁJTK oktatói voltak, akik mindketten részt vettek korábban a jogi szövegek közérthetőségével kapcsolatos projekteken.

A szövegek annotálásához a szakértők minimális instrukciókat kaptak; a feladat megfogalmazása szerint elsődlegesen a tapasztalataik alapján kialakított intuíció alapján kellett a szövegeket osztályozniuk. Ennek az volt a szerepe, hogy az eredmények által modellezni lehessen azt a prototipikus forgatókönyvet, amikor is a nyelvészeti ismeretekkel nem (feltétlenül) rendelkező, de a közérthető kommunikációban már valamelyest jártas jogászok „elsőként” találkoznak a problémával. Az így nyert eredmények betekintést engedhetnek egyrészt arra vonatkozóan, hogy a szaknyelv ismerete mennyiben nehezíti az átlagos befogadó nézőpontjában való belehelyezkedést, másrészt pedig, hogy a gépi javaslatétel mennyiben befolyásolhatja az eredményeket.

Az annotálás interpretációjánál különösen fontos figyelembe venni, hogy a gold standard ez esetben a NAV nyelvész szakértőinek ítélete volt. A későbbiek során érdemes lehet a validációs halmazt laikusok általi címkézésre is közreadni, hogy az ő nézőpontjuk is leképeződhessen a kérdés megítélésében.

|             |               | Önálló      |             |            | Géppel támogatott |             |             |
|-------------|---------------|-------------|-------------|------------|-------------------|-------------|-------------|
|             |               | P           | R           | F1         | P                 | R           | F1          |
| huBERTPlain | Átfogalmazott | 0.76        | 0.86        | <b>0.8</b> | 0.67              | 0.79        | <b>0.73</b> |
|             | Eredeti       | <b>0.83</b> | 0.73        | 0.78       | <b>0.75</b>       | 0.62        | 0.68        |
| A1          | Átfogalmazott | 0.52        | 0.89        | 0.66       | 0.58              | 0.76        | 0.66        |
|             | Eredeti       | 0.63        | 0.19        | 0.29       | 0.66              | 0.46        | 0.54        |
| A2          | Átfogalmazott | 0.53        | <b>0.91</b> | 0.67       | 0.49              | <b>0.91</b> | 0.64        |
|             | Eredeti       | 0.69        | 0.19        | 0.3        | 0.31              | 0.04        | 0.07        |

16. táblázat: Eredmények a kézi validációs halmazon a szokásos metrikák (pontosság – P, fedés – R és F1 tekintetében. A1, A2: első és második annotátor).

A 16. táblázat az elért eredményeket ismerteti. A huBERTPlain modell által kiosztott címkék esetében jól látható az az aszimmetria, hogy a szakértők jelentősen több mondatot tartanak közérthetőnek, mint amennyi azok valós aránya a kiválasztott mondatok között. Ezt jól szemlélteti az általuk elért 0.89, illetve 0.91-es fedés érték a közérthetőnek ítélt mondatok esetében, amellyel párhuzamos az átfogalmazandó szövegosztályt ugyanezen metrika szerint 0.19 jellemezte mindkét szakértő esetében. Ennek egyrészt oka lehet az a fajta inherensen jelenlevő domén tudás, amellyel oktatói, valamint jogi ismereteiknek

köszönhetően rendelkeznek, és amely segítheti őket abban, hogy például a jogi szakszókincset mélyebben használó mondatokat is könnyen tudjanak értelmezni. Másrészt a validációs adatok között sok olyan mondat is helyet kapott, amelyek esetében a döntés rendkívül nehéznek bizonyulhatott. A 18) példa két mondatot tartalmaz, amelyek közül az a) a NAV szakértői által átfogalmazandónak ítélt, a b) pedig egy már átfogalmazott mondat volt.

- 18) a) „Ezzel a módszerrel kimutatható hogy melyik címzetti körnél ér leginkább célba a tájékoztatás.”  
b) „Az ”eSZJA” felületen a „Gyakori mezők” opciót választva egy 15 lehetőséget tartalmazó felület válik láthatóvá, ahol kiválasztható az „Egyéni vállalkozók” lehetőség.”

Az első mondat esetében ismét csak előtérbe kerül az a körülmény, hogy a szakértők nem csak közérthetőségi, de helyesírási természetű javításokat is végrehajtottak. A „hogy” utáni vessző betoldása lényegileg nem befolyásolta a mondat jelentését, azonban (tekintettel arra, hogy a két részkorpusz válogatása során a szegmentált mondatok közötti eltéréseket karakter alapon vettem figyelembe) az így az átfogalmazott részkorpuszba került. A hasonló jelenségek automatikus szűrése például a MS Word helyesírás ellenőrző moduljának kiértékelése által lehetne megoldható, azonban megjegyzendő, hogy a feladat ez esetben sem triviális (a program nem tartalmaz olyan lehetőséget, amely ezt natívan biztosítaná), azonban még ezek után is maradnának olyan esetek a korpuszban, amelyekre a MS Word jelen pillanatban nem biztosít implementált szabályt. Erre a teljeskörű megoldást egy saját helyesírás ellenőrző implementálása jelentené, amely azonban mind terjedelmében, mint összetettségében messze meghaladja a jelen értekezés kereteit.

A b) esetben ismertetett mondat valójában két másik összevonásával keletkezett: „Az ”eSZJA” felületen a „Gyakori mezők” opciót választva egy 15 lehetőséget tartalmazó felület válik láthatóvá. Itt a kiválasztható az „Egyéni vállalkozók” lehetőség. Ez utóbbi egyértelműen stilisztikai javítás eredménye, amelynek célja feltehetően a kifejezésre kerülő gondolat egy mondaton belül tartása volt, azonban kérdéses, hogy ez mennyiben befolyásolta a mondat érthetőségét.

A feladatot ezen felül az annotátorok elmondása alapján tovább nehezíthette a mondatok (korábban megindokolt) kontextusnélkülisége.

Abban az esetben, amikor a huBERTPlain által kiosztott címkék metaadatként szintén jelen voltak az annotálás során, a tapasztalatok vegyesek. A1 esetében jelentős javulás figyelhető meg az átfogalmazandó mondatok felismerésében (a fedés mintegy 0.27-tel javult), A2 esetében azonban az eredmény romlott az önállóan kiosztott címkékkel összevetésben (R: - 0.15). Elmondásuk alapján A1 többségében elfogadta a modell által adott javaslatot, míg A2 továbbra a saját intuícióját követte. Realisztikusan nézve, azaz eltekintve a magas R értékektől, amelyek a közérthető mondatok „túljelölése” által keletkeztek, a legjobb eredményeket a gépi tanult modell érte el mindkét validációs halmazon.

Fontos azonban megjegyezni, hogy a NAV szakértői által hozott döntések (közérthető vagy átfogalmazandó-e az adott mondat) rekonstruálása nem feltétlenül feleltethető meg teljes mértékben a mondat könnyen érthető természetéről alkotott általános döntéssel. Ennek hátterében részben a fent említett stilisztikai vagy helyesírási természetű javítások is állnak. Tekintettel arra, hogy a dolgozat elkészültekor egyéb tanítóadat nem állt rendelkezésre, ennek összevetése más tanítóadatokon mért eredményekkel egyelőre nem lehetséges.

A problémát tovább árnyalja, miszerint visszajelzéseik alapján a szakértők több esetben is olyan tényezők alapján döntöttek, amelyek nem képezik részét a hagyományos közérthetőségi iránymutatásoknak, azonban véleményük szerint az egyes mondatok érthetőségét érdemben befolyásolták.

- 19) a) „Fel kell-e bármilyen módon tüntetni ezt a tényt az XML fájlban és ha igen, hol?”
- b) „Ha az ONYA felületen a gyermekek adatai és a kiutaláshoz szükséges adat is feltüntetésre került, a nyilatkozatot nem szükséges beadni.”
- c) „Év elején elhunyt a gépjármű tulajdonosa, hagyatéki tárgyalás már a tavasz folyamán lezajlott.”

A 19) a) esetében az annotációs fájlban feltüntetett komment szerint kérdéses lehet magának az XML kiterjesztés fogalmának közérthetősége. A NAV szakértői a mondatot közérthetőnek ítélték, A1 azonban átfogalmazandónak. Ez jól illusztrálja a jelenséget, amikor a jogi szövegek inkorporálják más szaknyelvek terminusait, így a szöveg érthetősége valójában több szaknyelv szempontjából is kérdéses lehet. A 19) b) esetében a mozaikszó (ONYA – Online Nyomtatványkitöltő Alkalmazás) közérthetőségét kérdőjelezte meg, ugyanakkor a mondatban például funkciógés szerkezet is felbukkan. Mind a NAV szakértői, mind A1 átfogalmazandónak ítélte a mondatot, azonban, mivel a NAV szakértői esetében a



mozaikszavak feloldása nem volt jellemző gyakorlat az átfogalmazások során, kérdéses, hogy ezt a döntést azonos motiváció támasztotta-e alá? A mondatok kontextusnélkülisége itt is nehezítő tényező lehetett, hiszen, amennyiben a kérdéses dokumentumban ez korábban feloldásra került, az A1 döntését is befolyásolhatta volna. A 20) c) esetében A1 a „hagyatéki tárgyalás” terminust, mint szakkifejezést jelölte meg annak indokaként, hogy a mondatot nem közérthetőnek jelölte. Az átfogalmazás során azonban nyelvtani hiba került javításra a határozott névelő betoldása által a 20) szerint:

20) „Év elején elhunyt a gépjármű tulajdonosa, **a** hagyatéki tárgyalás már a tavasz folyamán lezajlott.”.

A1-nek az egyes osztályozandó mondatokhoz fűzött megjegyzései alapján mindhárom eset olyan volt, amelyet tipikusan problémaként jelölt meg a mondatok címkézése során.

A fentiek jól rávilágítanak a *közérthetőség* fogalmának erősen értelmezésfüggő és komplex természetére, egyben arra is, hogy akár a jogi szakmán belül is jelentős eltérések mutatkozhatnak annak értelmezésében attól függően, hogy a szakértő milyen területen mozog. Mindemellett nem hanyagolhatjuk el a NAV szakértőinek tapasztalatait a hivatal ügyfeleivel lefolytatott kommunikáció terén, amely szintén nem állhat egységesen rendelkezésre minden joghoz értő számára, azonban az egyes szövegek megítélését jelentősen képes befolyásolni.

Az egyes szakértők, illetve a huBERTPlain modell között mért egyetértések feltehetőleg a fent ismertetett, esetleg egyéb, a konkrét példák kapcsán fel nem tárt divergenciákat is tükröznek a „közérthető” fogalmának megítélése kapcsán. A vonatkozó értékeket a 17. táblázat ismerteti.

| Önálló |             | huBERTPlain | A1      | A2 |
|--------|-------------|-------------|---------|----|
|        | huBERTPlain |             |         |    |
|        | A1          | -0.0011     |         |    |
|        | A2          | -0.0452     | -0.0205 |    |

| Géppel támogatott |             | huBERTPlain | A1      | A2 |
|-------------------|-------------|-------------|---------|----|
|                   | huBERTPlain |             |         |    |
|                   | A1          | 0.3020      |         |    |
|                   | A2          | 0.0235      | -0.0229 |    |

17. táblázat: : Mért egyetértések az egyes címkézők között (mérés szám: Cohen  $\kappa$ ).

Az egyes címkézők közötti egyetértés (Inter-Annotator Agreement - IAA) kiértékelése Cohen  $\kappa$  alapján történt, amely olyan esetekben alkalmazható, amikor is a címkézőknek minden adathoz egyetlen címkét kell kiválasztania (Cohen 1960)<sup>164</sup>. Ennek előnye többek között, hogy egyetértésként a véletlen egyezésen felüli azonos címkézéseket érti. A mérőszám a  $[-1, 1]$  intervallumon belülre esik, ahol  $-1$  az inverz egyetértést,  $1$  pedig a tökéletes (véletlenül felüli) egyetértést jelenti. Ennek megfelelően szokásos értékelése szerint  $0$  körüli érték esetén nem beszélhetünk egyetértésről,  $\kappa < 0.2$  esetén minimális- (slight),  $0.21 < \kappa < 0.4$  esetén elfogadható- (fair),  $0.41 < \kappa < 0.6$  esetén mérsékelt- (moderate),  $0.61 < \kappa < 0.8$  között jelentős- (substantial), míg  $0.81 < \kappa$  esetén tökéletes (near perfect / perfect) egyetértésről beszélhetünk.

A 17. táblázat kiválóan reflektál a fentebb ismertetett problémákra; értékei jellemzően mind az önálló, mind pedig a géppel támogatott esetben  $0$  körül konvergálnak. Kivétel ez alól a jelenős növekedés A1 és a huBERTPlain esetében géppel támogatott címkézés mellett, ahol a mért érték  $\sim 0.3$  szerint alakult.

Az 1.1 fejezetben megfogalmazott iii. hipotézis szerint a megfelelő tanítóadatok birtokában alkotható olyan gépi tanult modell, amely képes egymástól hatékonyan elkülöníteni közérthető, valamint a még (további) átalakításra szoruló szövegegységeket. A fentiek tükrében azonban annyi jelenthető csak ki biztosan, hogy a finomhangolt modell hatékonyan volt képes rekonstruálni a NAV szakértőinek közérthetőségről alkotott intuícióját. A kiinduló hipotézis szempontjából ennek nincs jelentősége, de megjegyzendő, hogy nem egyértelmű, a modell ezt pontosan milyen nyelvi kompetenciák elsajátítása alapján érte el (vö. 10.1 fejezet).

Az eredmények ismét rávilágítanak a kézzel készített szabálykészlet szükségességére (azok „univerzális” alkalmazhatósága miatt), valamint a közérthető terminusnak a jogterületekhez kötöttségére, és a kérdés megítélése kapcsán jelentkező egyéni variabilitásra annak értelmezése kapcsán.

---

<sup>164</sup> Szokásos számítása:  $\kappa = \frac{p_0 - p_e}{(1 - p_e)}$ , ahol  $p_0$  az annotátorok között megfigyelt relatív egyetértés,  $p_e$  pedig a véletlen egyetértés feltételezhető valószínűsége.

## 12. Szabálykészlet

A szoftver tervezett felépítését a 6. fejezet bevezetőjében már röviden felvázoltaknak megfelelően a gépi tanulással készített modellek kimenete egy halmaz, amely a potenciálisan problémásnak ítélt mondatokat tartalmazza.

Ahhoz azonban, hogy ebből a kimenetből a felhasználó számára is hasznos visszajelzéseket tudjunk létrehozni, szükséges egy „kézzel írt” szabályrendszer kidolgozása is, amely az egyes, a közérthetőséggel foglalkozó kutatási irányok javaslatait kísérli meg implementálni azon esetekben, amikor a javaslat a magyar nyelvre is alkalmazható, valamint nyelvtechnológiai / NLP eszközökkel megoldható ennek automatizálása.

Ennek megfelelően a jelen fejezet összegzést nyújt a szakirodalomban felmerült közérthetőségi javaslatokról, továbbá azok implementációs lehetőségeiről a magyar nyelvre tekintettel.

### 12.1 Áttekintés

Az egyes kutatási irányok részéről a szakirodalomban felmerülő javaslattételi lehetőségeket a 18. táblázat foglalja össze a pszicholingvisztika (3.1.3 fejezet), az olvashatósági formulák készítői (3.2.2), a PLM (3.3.4 fejezet; az FPLG javaslatai), valamint a hazai vonatkozású Stíluskönyv (4.3.1 fejezet) javaslatai alapján. A táblázat nem tartalmazza valamennyi, a dolgozatban ismertett kutatási irányzatot, sem pedig azok valamennyi javaslatát. Itt kizárólag azok jelennek meg, amelyek szoftveres implementáció szempontjából szóba jöhetnek, azaz kellően konkrét nyelvi jelenség(ek)re vonatkoznak.

A PLM ISO sztenderdje által hivatkozott ajánlások az FPLG ajánlásrendszerén keresztül reprezentálódnak. Ennek oka, hogy a sztenderd alapvetően a dokumentumok fogalmazóinak készült, vagyis nem volt célja, hogy szoftveres megoldásokat támogasson. Éppen ezért a benne leírtak döntő többsége olyan elveket fogalmaz meg, amelyek megkerülhetetlenek az a kreatív munkavégzés során, azonban programozási szinten megvalósításuk nem elképzelhető.

|  | Szempont  | PLM | Stíluskönyv | Pszicholing-<br>viztika | Olvashatósági<br>formulák |
|--|---|-----|-------------|-------------------------|---------------------------|
| Szószint   | Elhagyott kopulák, vonatkozó névmások visszaillesztése  |     |             |                         |                           |
|  | Funkcióigék ("terpeszkedő kifejezések") kerülése  | X   | X           | X                       |                           |
|  | Személytelen szerkezetek - nominalizáció kerülése; igésítés   | X   | X           | X                       |                           |
|  | Tartalmas / funkciószavak (kötőszók, névutók, segédigék) arányának helyes megválasztása   |     |             |                         |                           |
|  | Lehetőleg közismert és rövid tartalmas szavak alkalmazása   |     |             |                         | X                         |
|  | Absztrakt vonatkozású szavak helyett konkrét, egyszerűbb jelenésűek használata (főleg igék esetén)  | X   |             |                         | X                         |
|  | Modalitások explicit jelölése (segédigékkel)  |     |             |                         |                           |
|  | Összevonások megengedése  |     |             |                         |                           |
|  | Rövidítések kerülése, inkább azok feloldása   | X   |             |                         |                           |
|  | Köznyelvben is szereplő, de a jog kontextusában más értelmet nyerő kifejezések kerülése   | X   |             |                         |                           |
|  | Konzisztens terminushasználat   | X   | X           |                         |                           |
|  | Technikai és szakzsargon kerülése   | X   |             |                         |                           |
|  | Archaizmusok kerülése, pl.: "akként", "ilyetén"   | X   | X           |                         |                           |
|  | Igeidők egységes használata   |     | X           |                         |                           |
|  | "illetve", "illetőleg" használatának kerülése, ha lehet vagy - vagy alkalmazása a kirekesztő választás érzékeltetésére a "vagy" egyszeri használata helyett |     | X           |                         |                           |
|  | "mégpedig" túlhasználata  |     | X           |                         |                           |
| "ugyanakkor" használata "azonban" értelemben           |   | X   |             |                         |                           |
| "követően" használata az "után", "nyomán" helyett      |   | X   |             |                         |                           |
| Mondat szint   | Cselekvő (~nem nominalizált) igei konstrukciók használata - passzív mondat szerkesztés elkerülése   | X   |             | X                       | X                         |
|  | Többszörös tagadás kerülése   | X   |             | X                       |                           |
|  | Prepozíciók (előljárószók) egyértelműsítése   |     |             |                         |                           |
|  | Közbevetett, hátravetett tagmondatok kerülése   |     |             | X                       |                           |
|  | Rövid (tag)mondatok   |     |             |                         | X                         |
|  | Inkább affirmatív (állító), mint negatív mondatok   |     |             |                         | X                         |
|  | Többszörösen összetett konstrukciók (pl.: beágyazott tagmondatok, többszörös alárendelések, birtokos szerkezetláncok) kerülése                              | X   | X           |                         | X                         |
|  | Kivételek alóli kivételek - típusú szerkesztés kerülése   | X   |             |                         |                           |
| Legfontosabb normatív tartalom a mondat élére kerüljön | X   |     |             |                         |                           |
| Bekezdés   | Tételmondat alkalmazása   | X   |             |                         |                           |
|  | Kötőszavak alkalmazása (tag)mondatok határán (ezért, tehát, vagyis, ezzel ellentétben)  | X   |             |                         |                           |
|  | Rövid bekezdések, témánkénti szövegtagolás  | X   |             |                         |                           |
|  | Soktagú (mondatközi) felsorolások felbontása listákká   | X   |             | X                       |                           |
| Szöveg   | Szemléletes példák alkalmazása  | X   |             |                         |                           |
|  | Kereszthivatkozások minimalizálása  | X   |             |                         |                           |

18. táblázat: Közérthető fogalmazás jellemzői keretek szerint („X”: jelen van, „-----”: a magyarban nem értelmezhető)

Fontos megemlíteni, hogy ezek között is fellelhetők első olvasatra igen absztraktnak tűnő, de valójában könnyen implementálható elvárások, és fordítva – olyanok is, amelyek triviálisnak tűnnek, de a gyakorlatban mégis csak nehezen, vagy egyáltalán nem megvalósíthatók a nyelvtechnológia mai eszközeivel. Az egyes megkötésekhez optimális implementációs stratégia megválasztásához szükséges tehát tisztázni, mit jelentenek az itt felsoroltak, amikor a sor a konkrét kódba ültetésre kerül. A jelen fejezet sorra végighalad a 18. táblázatban felvázolt pontokon, azonban megállapítást csak annak kapcsán tesz, hogy az adott közérthetőségi ajánlás megvalósíthatónak bizonyult-e. Az implementáció részleteit a 13.2 fejezet tartalmazza, az itteni felsorolással azonos sorrendben.

## 12.2 Lexikai szabályok

A fejezet bemutatja azon lehetséges szabályokat, amelyek a lexikai szinthez köthetők. Fontos megjegyezni, hogy gyakorlati szempontból az, hogy az adott irányzat maga mely nyelvi szinthez sorolta az ajánlást, nem minden esetben egyezik meg az implementáció során alkalmazható praktikus szempontokkal. A jelen összefoglalásban minden esetben az utóbbi dominált. Ahol emiatt a 18. táblázathoz képest inkonzisztencia jelentkezett, ott ezt külön kiemelésre került.

Minden tartalmi egység elején a 18. táblázat egy, vagy több vonatkozó sora mutatja, hogy az pontosan mely nyelvi jelenséggel foglalkozik. Az egyes nyelvi jelenségek értékelésének szempontja egyrészt, hogy az a magyar nyelv esetében releváns-e (lévén a legtöbb idézett ajánlás eredetileg az angol nyelvvel kapcsolatosan lett megfogalmazva). Másrészt az egyes nyelvi jelenségek kapcsán minden esetben rövid összefoglaló is szerepel az implementáció lehetséges irányáról, vagy amennyiben az nem megoldható, ennek okairól.

*„Elhagyott kopulák, vonatkozó névmások visszaillesztése”*

*„Modalitások explicit jelölése (segédigékkel)”*

A magyarban Keszler (2000: 254-257) a mondatrésztteremtő segédigéket, a segédigeneveket, valamint a segédszókat sorolja a kopulák közé. A segédige kategóriáját olyan viszonyzóként határozza meg, amely az ige grammatikai jelentését hordozza, illetve mondatrésztteremtő és szóalakteremtő jelleg szerint osztályozza őket. Példaként a „volt”, a „marad” és a „fog” alakokat hozza. Segédigenevek alatt a segédigékből képzők segítségével

előállított igenévi alakokat érti, mint például a „lenni”, vagy a „lévén”. Segédszóként elsődlegesen a jövő idejű igealakok segédigéjét, a „fog” alakot, valamint a feltételes mód jelentését hordozó „volna” alakot említi.

A magyarban ezek közös jellemzője, hogy elhagyásuk a mondatból legtöbbször agrammatikus szerkezetet eredményez, ilyen módon tehát a kopulák elhagyásának detektálása nem sorolható világosan egy közérthető megfogalmazást támogató szoftver által megoldandó feladatai közé<sup>165</sup>.

Hasonló a helyzet a modális (aspektuális és pragmatikai) segédigékkel is, mint például a „szokott”, a „talál” és a „tetszik”. Ezek elmaradása esetén a mondat alapvető jelentése módosul, vagy ismét csak agrammatikussá válik.

#### *„Funkcióigék (‘terpeszkedő kifejezések’) kerülése”*

A funkcióigés szerkezetek (light verb constructions - LVC) legtágabban értelmezve olyan névszói-igei szerkezetként írhatók le, amelyekben „egy határozóragos, tárgyragos vagy zéró ragos, olykor névutós névszó és egy rendszerint kisebb-nagyobb mértékben alulspecifikált jelentésű ige, ún. funkcióige szerepel, pl. *aláírásra kerül, különbséget tesz, elintézését nyer, követelményt támaszt, hivatkozás történik, vizsgálat alá vesz; kivételt képez, magyarázatként szolgál, gondot jelent, türelmet tanúsít*. A szerkezet fogalmi jelentésének magvát a névszó, nyelvtani jelentését az ige, stílusértékét pedig a kifejezés egésze hordozza, amely eltérhet mindkét tag stílusértékétől (pl. az *aláírás* és a *kerül* közömbös stílusértékű szavak, az *aláírásra kerül* szerkezet viszont formális stílusértékű)” (Lanstyák 2020: 63).

Ezen szerkezetek olyan alakjait, ahol a mondatban a névszói tag szerepel első helyen, majd ezt követi az igei tag, az alkalmazás lista alapon szűri és javaslatot tesz a megfelelő (kvázi) szinonim alakra<sup>166</sup>. A kapcsolódó kifejezések listáját, és a helyettük javasolt alakokat a II. Függelék tartalmazza tételiesen.

---

<sup>165</sup> Az angol esetében is csak marginálisan, például az angolt, mint idegennyelvet, esetleg Lingua Francát beszélők szóhasználata kapcsán merülhetnek olyan esetek, amikor a „be” egyes alakjai nem jelennek meg a mondatban az alany és az állítmány között (vö. Leuckert – Neumaier, 2016).

<sup>166</sup> Az értekezés keretében elkészítendő szoftver esetén a csere csak az alanyesetű alakokat érinti, mivel a megfelelő inflexiók áttemelése a helyettesítő alakra nyelvgenerálási feladatokat is magában hordozna.

*„Személytelen szerkezetek - nominalizáció kerülése; igésítés”*

*„Cselekvő (~nem nominalizált) igei konstrukciók használata - passzív mondat szerkesztés elkerülése”*

A fenti két kritérium a 18. táblázatban a lexikai, illetve a mondat szint alatt szerepeltek, aszerint, hogy az adott irányzat hová sorolta be őket. A gyakorlatban azonban ezek inkább a szószinten oldhatók meg, és többé kevésbé egymással szinonimák, ezért itt, együttesen kerülnek tárgyalásra.

A nominális stílus természetét a magyar nyelvre vonatkozóan főként irodalmi kontextusban szokás vizsgálni (pl. Danyi 2010, Kemény 2020). Ezek a munkák az ilyen fogalmazásmód legfontosabb jegyeként a névszói elemek dominanciáját emeli ki, konkrétan a névszói állítmányok túlsúlyát, valamint a nyelvi kapcsolóelemek hiányát nevezik meg fő ismérvként. A problémakör felmerült a MJK elemzése kapcsán is (vö. 3.5.2 fejezet, 5. táblázat), amely során Vincze (2018) kimutatta, hogy a jelenség a köznyelvhez képest szignifikánsan nagyobb arányban jellemzi a jogi domén szövegeit, így automatikus vizsgálata e helyütt is feltétlenül releváns.

Az egyik legbiztosabb heurisztika az ilyen mondatok megtalálására az ige hiányának figyelése a mondatok morfológiai elemzésének felhasználásával, így tehát az alkalmazásban is ez a konkrét implementáció kapott helyet.

*Tartalmas / funkciószavak (kötőszók, névutók, segédigék) arányának helyes megválasztása*

Pinker (2013: 116-117) a funkciószavakhoz a segédigéket, a névelőket, a névmásokat, a ragokat és névutókat, a mellékmondatokat bevezető szavakat, továbbá a kötőszavakat sorolja. Megfogalmazásában ezek tekinthetők „a kristályosodott grammatika egy-egy darabjának”, amelyek előrerajzolják minden összetettebb kifejezés szerkezetét, és ezáltal egyfajta állványzatot hoznak létre a teljes mondat számára.

Az ezek arányára vonatkozó megállapítás (megszorítás) az olvashatósági formulák időszakából származik, és a későbbi, a közérthetőséget tárgyaló egyéb irányzatokban nem is bukkan fel. Az előirányzott ideális arányt illetően a vonatkozó munkák nem tartalmaznak konkrét, számszerű ajánlást. Ezen felül a magyarban a funkciószavak alkalmazása meglehetősen kötött. Az ilyen kifejezések elhagyása valamilyen ideálisnak tartott arány elérése érdekében inkább lenne káros, mint hasznos; azáltal, hogy a mondatbeli logikai,

grammatikai viszonyokat adott esetben kevésbé expliciten jelöljük, veszélyeztetjük a szándékolt jelentés helyes és teljes átadását, plusz funkciószavak hozzáadása viszont plusz jelentések megjelenését von(hat)ja maga után, ami ismét csak nemkívánatos cél<sup>167</sup>.

Az említett érvek miatt magyar nyelven írt szövegek esetében ezt az elvet sem tartom alkalmazhatónak.

*„Lehetőleg közismert és rövid tartalmas szavak alkalmazása”*

*„Absztrakt vonatkozású szavak helyett konkrét, egyszerűbb jelenésűek használata (főleg igék esetén)”*

A fenti két kritérium közül az első inkább a mindennapi nyelvhasználatban mérhető gyakoriság, illetve karakterszám, vagy szótagszám szerint kívánja osztályozni a főnevek, igék, jelzők és határozók csoportját, míg a második esetében a distinkció alapját egy szemantikai kritérium adja. A konkrét jelentésű szavak ilyen értelemben jellemzően a fizikai világ egy-egy entitására utalnak, olyanra, amelyet érzékszerveink, vagy például belső pszichológiai állapotaink megfigyelése segítségével képesek lehetünk megtapasztalni (asztal, szék, szomorúság stb.). Ezzel szemben az absztrakt kifejezések olyanok, amik nem rendelkeznek ezzel a tulajdonsággal, jellemzően (elvont) fogalmak (pl. zsarnokság, politika stb.).

Az első esetén a problémát az okozza, hogy a „közismert” fogalma egy rendkívül homályos kifejezés, amely nem feltétlenül rendelkezik számszerűsíthető referenciával. Egy szótár szavainak lemmáira rákeresve valamely korpuszban (például az MNSZ2 -ben) természetesen meg tudjuk határozni a szövegek leggyakoribb elemeit, egy szófaji szűrés után pedig könnyen kinyerhető például a főnevek, igék stb. listája gyakoriság szerint rendezve.

---

<sup>167</sup> Az angol nyelvvel kapcsolatban Pinker a telegrafikus nyelvet hozza példaként, amelyben a funkciószavak elhagyása egy viszonylag gyakori jelenségnek számított az elküldött karakterek magas költsége miatt. Kiemeli, hogy hogy ilyen fajta fogalmazás mindig egyfajta hazardjáték, amikor is a feladó sosem lehet teljesen biztos abban, hogy a címzett pontosan a szándékolt jelentéssel bíró üzenetet fogja az értelmezés során dekódolni. A beszélt nyelv szempontjából a telegrafikus nyelvet a páciensek nyelvhasználatára is hatást gyakorló mentális betegségekkel foglalkozó szakirodalom is vizsgálati tárgyának tekinti (pl. Gabig 2013). Ebben az esetben a fogalom olyan mondatok (szinte) kizárólagos alkalmazását jelöli, amelyekben csakis tartalmas szavak szerepelnek, és amelyek ezáltal a funkciószavakat teljesen mellőzik. Ezen két példa, habár szélsőségeket ábrázol, mégis jól érzékelteti, hogy a funkciószavak esetén akár csak azok arányának mesterséges megváltoztatása milyen komoly következményekkel járhat a jelentés rekonstruálására vonatkozóan.



Közérthető megfogalmazás esetében azonban a nézőpont, amely alapján meg kell ítélnünk egy kifejezés vagy szerkezet milyenségét, az átlagos befogadó nézőpontja. Könnyű olyan kifejezést találni, ami egyes csoportok számára közismert és jól bejártott (elég csak például pontosan a „lemma” szóra gondolni, amely sokszor alkalmazott és „közismert” terminus a nyelvészetben, de a szakmán kívül jóformán ismeretlen). A közismertségre valójában nem létezik egzakt meghatározás, ha pusztán gyakoriság alapján közelítjük meg a kérdést. Például, ha egy konkrét (reprezentatívnak ítélt) korpusz szavainak leggyakoribb x % -át soroljuk ide, az valójában csak a korpusz egy jellemzőjét fogja tükrözni, nem ad feltétlenül számot egy beszélői közösség általános ismereteiről.

Az absztrakt – konkrét elhatárolás ezzel szemben implementálható hatékonyan lista alapon. Ennek megoldásához támpontot nyújt Vincze et al. (2006), amelyben a szerzők a MetaMorpho fordítócsalád különböző fordítási feladataihoz készített adatbázisok építése során a magyar alapszókincshez tartozó szavakból építettek adatbázist. Ez ugyan nem tartalmazza a szakzsargonba tartozó szavakat, azonban az „ABSTRACT” szemantikai jellemzővel ellátott szavak leválogatása így is 35481 elemű listát eredményez. Az annotálási útmutató alapján ilyen címkével azon bejegyzéseket látták el, amelyek:

- mentális jelenségre,
- elvont fogalomra, vagy
- valamilyen absztrakcióra utalnak.

Ennek megfelelően egy szöveg szókészletének lemma alapú összevetése (a főnévi POS-tag-et kapó szavak tekintetében) a lista elemeivel jó támpontot adhat a teljes szöveg szókészletének konkrét / absztrakt vonatkozásáról is.

*„Összevonások megengedése”*

A javaslat az FPLG-nek egy olyan elve, amely a hivatalos szövegeket hivatott írásképében közelíteni a mindennapok során alkalmazott nyelvhasználathoz, pl. az angol *I am* → *I'm* átírás megengedésével. A magyar esetében ez nem alkalmazható.

„Rövidítések kerülése, inkább azok feloldása”

A rövidítések kezelése kapcsán alapvetően két megközelítés tűnt célravezetőnek. Tekintettel arra, hogy az érintett tokenek viszonylag konzisztens karaktermintázatot követnek, a szűrésük elvben megoldható reguláris kifejezéssel (regex), vagy listából való kereséssel (amennyiben rendelkezésre áll a lehetséges rövidítéseket tartalmazó teljes felsorolás).

A hivatalos tájékoztatókban előforduló rövidítések referenciájukat tekintve nem pusztán az adott doménre nézve specifikus rendeleteket, törvényeket, szervezeteket rövidíthetnek (pl. Vht. – 1994. évi LIII. törvény a bírósági végrehajtásról), de egyéb, doménfüggetlen esetek is nagy számban előfordulhatnak (pl.: [Dd]r.). Emiatt a regex-alapú megközelítésnél magas annak a kockázata, hogy a rövidítésekre írt kifejezés más, a mindennapi életben is gyakori rövidítésekre is illeszkedni fog, ezáltal a kifejezés erősen túlgenerálhat.

Tipikus mintázat az egyes jogszabályok, törvények rövidítései esetén például a következő<sup>168</sup>:

(1) ([A-ZÖÜÓÓÚÉÁÚÍ]{1}[a-zöüóóúéáúí]+)(\.)

amely a nagy kezdőbetűvel induló, majd tetszőleges számú kisbetűvel folytatódó, és ponttal záródó esetekre illeszkedik, pl.: Tft. – 1994. évi LV. törvény a termőföldről.

Egy másik eset, amikor a rövidítés végződése utal az adott jogszabálynak a jogszabályi hierarchiában (Petrétei – Tilk 2014) elfoglalt helyére is. Ezt jellemzően a „tv” végződés, vagy a rövidítés első tagja után szóközzel hozzáfűzött „tvr.”, „tv.”, „vhr.” vagy „r.” utótag jelzi (ebben a sorrendben: törvényerejű rendelet, törvény, kormányrendelet, rendelet). A rövidítés végén elhelyezkedő „tv.” utótag, valamint a rövidítés végén, esetleg szóközzel elválasztva szereplő másik három lehetséges utótag szintén megkereshető például a következő kifejezéssel:

(2) (\w+tv\b|([A-ZÖÜÓÓÚÉÁÚÍ]{1}[a-zöüóóúéáúí]+)(\.)([ ]?vhr\.|[ ]?tv\.|[ ]?tvr\.)?)|([^\r\n\t]+([ ]?tv\.|[ ]?r\.|[ ]?vhr\.|[ ]?tvr\.)

Ezek után hozzáfűzve például egy Negative Lookbehind kifejezést kiszűrhetők a gyakori rövidítés mintázatok, amelyek nem csak az adott domain-t jellemzik (a példában a már említett „[Dd]r.” mintázat kizárása történik meg):

(3) (?<!( [Dd]r )\.)

<sup>168</sup> Megjegyzendő, hogy az alap regex kifejezések csak az ASCII karakterekre fognak illeszkedni, ezért az UTF-8-as, magyar ékezetes karaktereket külön szükséges a mintában felsorolni.

A Stíluskönyv mellékleteként elérhető egy felsorolás a magyar jogszabályok rövidítéseiről<sup>169</sup>. Az ezen szereplő 209 entitás rövidítései közül a (2) és (3) kifejezés összefűzésével keletkező regex 192 esetében ad találatot; ez tehát jelzi például a „GET” (2008. évi XL. törvény a földgázellátásról) tokent, amely a fentiekől eltérő mintázatba illeszkedik. Ha az összefűzött kifejezést ezek alapján bővíteni próbáljuk, akkor az azonban jelezni fog például a római számok esetén is. Ezek kizárásához újabb komplex mintát kell definiálni kivételként, valamint számítani kell rá, hogy például az „Ftv.” (2004. évi XXIV. törvény a lőfegyverekről és lőszeréről) és az Szja. azonos mintázatba fognak esni, ezzel fals pozitív találatokat generálva.

Mindezek figyelembevételével a szoftverben a lista alapú szűrést választottam a Stíluskönyv mellékletében elérhető rövidítésekre. Ezzel ugyan a program veszít a később használatba vett rövidítésekkel kapcsolatos „általánosítási képességéből”, azonban pontossága jelentősen nő. A szűrt rövidítéseket a III. függelék tartalmazza részletesen<sup>170</sup>.

*„Köznyelvben is szereplő, de a jog kontextusában más értelmet nyerő kifejezések kerülése”*

A kritérium részben a 2.4.2 fejezetben már tárgyalt jelenségre reflektál, amely során köznapi értelemben is használt fogalmak a jogi terminológiában speciális jelentést kaphatnak. Az ilyen fogalmak megjelenése valamely hivatalos szövegben magában hordozza egyfajta „kvázi-megértés” kockázatát a laikus olvasó részéről. Ennek során az adott szakterületen nem jártas olvasó a fogalom köznapi értelmezését helyettesíti be a szöveg megfelelő pontjára, amennyiben nincs tudatában a speciális jelentéstöbbletnak.

A jelenség azért kockázatos, mert a megértés illúzióján túl ilyen esetben érdemi új tudás nem jön létre, csak egy hiányos információkon alapuló meg (nem) értés. Emiatt sérül az értelmezés során a szakszavakhoz társítandó azon elváráshalmaz, amelyet a szaknyelvek állítanak a saját terminológiájukkal szemben, például, miszerint a szakmai terminusoknak:

- „...Jól meghatározott jelentésük van,

---

<sup>169</sup> [https://kuria-birosag.hu/sites/default/files/joggyak/3\\_melleklet.pdf](https://kuria-birosag.hu/sites/default/files/joggyak/3_melleklet.pdf) (ellenőrizve: 2022.08.22.)

<sup>170</sup> Megjegyzendő, hogy az eredeti lista tartalmazott ismétlődő értékeket is. Ilyen például a Felnőttképzési tv., amely utalhat a 2001. évi CI. törvényre és a 2013. évi LXXVII. törvényre is. Az ilyen esetekben a vonatkozó rövidítés nem került duplikálásra a végleges listában. Ennek egyrészt az az oka, hogy a rövidítések feloldása nem célja az alkalmazásnak, valamint, hogy az (például a fenti esetben) nem is lenne feltétlenül megvalósítható automatikus eszközökkel, hiszen a megfelelő feloldás kiválasztásához szükséges a területet jól ismerő szakértők tudása is.

- Jelentésük nem tágítható vagy csökkenthető, tehát szövegkörnyezettől, pragmatikai tényezőktől függetlenek;
- Mindig ugyanabban az értelemben használjuk őket;
- Nem jellemző rájuk a konnotáció (másodlagos jelentés); ...” (Imre 2014: 15)

Ahhoz, hogy az érintett kifejezésekre a szerkesztő figyelmét fel lehessen hívni, szükséges egy nyílt felhasználású jogi terminológiai adatbázis (TB). Habár kereshető változatban a magyar esetében létezik hasonló<sup>171</sup>, ez „csak” keresőfelületet biztosít, az alapjául szolgáló adatok nem érhetőek el nyilvánosan.

A hasonló adatbázisok közös jellemzője, hogy valamely téma vagy jogterület köré csoportosítható szakszavakat gyűjtenek össze, vagyis hosszú távon szükséges lenne beszerezni az összes releváns szakterület szakszókészletét bemutató forrást. Ideális esetben az adatbázis metaadatként kellene, hogy tartalmazzon olyan tag-et is, amely a kifejezés vagy szó köznapi értelemben is megjelenő mivoltára tartalmazna utalást. Mindezek után a kifejezés kijelölésével a szövegben a fogalmazó jelzést kaphatna a fogalomnak a laikus értelmezés szempontjából vett problémás mivoltáról.

A jelen dolgozat elkészületkor a fentieknek megfelelő TB szabadon hozzáférhető verzióban egyelőre nem áll rendelkezésre, emiatt pedig az ehhez kapcsolódó funkció implementációja sem valósulhat meg<sup>172</sup>.

*„Konzisztens terminushasználat”*

*„Technikai és szaksargon kerülése”*

A kritérium implementációja ismét csak visszavezethető nyíltan elérhető terminológiai adatbázisok elérhetőségére. Bár egyes specifikus jogterületek legfontosabb terminusai külön-külön összegyűjthetőek<sup>173</sup>, a jog egészére, és különösen az adózással kapcsolatos jogi

<sup>171</sup> pl. az Igazságügyi Minisztérium gondozásában: <https://termin.im.gov.hu/> (ellenőrizve: 2022.08.22.)

<sup>172</sup> Hasonló adatbázis elkészítése csak a megfelelő jogi szaktudás birtokában, jelentős befektetéssel valósulhatna meg, emiatt a szerző nem tett kísérletet saját TB létrehozására. Megemlítendő, hogy az Országos Fordító és Fordításhitelesítő Iroda (OFFI) gondozásában például jelenleg is kidolgozás alatt áll egy adatbázis-alapú, integrált fordítói környezetben is használható TB, azonban a dolgozat írásakor ez egyelőre kizárólag belső használatra érhető el (vö. Klenk et al. 2020, Tamás 2019).

<sup>173</sup> Hasonló gyűjtemény elérhető például a közjog területére vonatkozóan (Márki – Szaniszló 2020), amelyben a szerzők a legfontosabb közjogi fogalmak kontextusba ágyazott magyarázatát adják meg azon terminusok esetében, amelyekkel a laikusok a legnagyobb valószínűséggel találkozhatnak a mindennapokban.

fogalmakra vonatkozóan megfelelően teljes jegyzék a vonatkozó terminus technicus-okról (a szerző tudomása szerint) nem áll rendelkezésre.

Amennyiben rendelkezésre állna hasonló, megvalósítható lenne például a terminushasználat dokumentumon belüli konzisztenciájának felmérése (az automatizmusok korlátainak figyelembevételével).

*„Archaizmusok kerülése, pl.: 'akként', 'ilyetén'” – „'illetve', 'illetőleg' használatának kerülése, ha lehet” – „'mégpedig' túlhasználata” – „'ugyanakkor' használata 'azonban' értelemben” – „'követően' használata az 'után', 'nyomán' helyett”*

A fenti öt javaslatot egyben tárgyalom, mivel ezek kezelése egységesen megoldható a figyelembe vett lexikális elem feltűnése esetén a szövegben, egy-egy figyelmeztető jelzés beiktatásával, amely a fogalmazót a Stíluskönyvben szereplő ajánlásról tájékoztatja.

Az 'akként', 'ilyetén', 'illetve', és az 'illetőleg' szavak külön listában kerülnek, amely az archaizmusok jelenlétét indikálja az adott mondaton belül. Az 'illetőleg', 'illetve', 'mégpedig', 'ugyanakkor', illetve a 'követően' jelenléte szintén külön megjelenik a szabályok kimenetében.

*„Igeidők egységes használata”*

A Stíluskönyvben ez a megszorítás főként a pertörténet, valamint a határozat rendelkező részének elkülönítése kapcsán kerül előtérbe. Előbbi esetében ajánlott a múlt idő egységes használata, míg utóbbi esetében a jelenidejűség preferált.

Nem szabad figyelmen kívül hagyni, hogy a dokumentum eredeti javaslatai kúriai határozatok szövegezésére vonatkoznak, így általános esetben, vagy például tájékoztató szövegek felülvizsgálata során nem minden ajánlás alkalmazható közvetlenül. Az igeidők fent leírt kezelése is egyértelműen csak a határozatok szövegére vonatkoztatható<sup>174</sup>.

---

<sup>174</sup> Megjegyzendő továbbá, hogy ahhoz, hogy ezt az ajánlást a gyakorlatban automatizált eszközökkel megoldhassuk, elsőként szükséges lenne ennek a két tartalmi egységnek a szintén automatikus elkülönítése. A gyakorlatban azonban ezeket nem választják el szükségszerűen nyilvánvaló szövegszerkezési elemek, mint például címek, vagy egyéb, a tartalmat kontextusba helyező indikátorok (például az egyes tartalmi elemeket jellemző fejezetcímek).

### *„Többszörös tagadás kerülése”*

A tagadás a gyakorlatban leginkább akkor nehezítheti az adott mondat feldolgozását, amennyiben a mondat logikai struktúrájában valamely kifejezés két negáció hatókörében is szerepel (pl. „nem igaz, hogy ne lett volna ott” -> ott volt értelemben). Ilyen esetekben a két tagadást a feldolgozás során „érvényteleníteni” kell, a kijelentés igazságértéke továbbra is igaz lesz. A magyarban tagadást a „nem”, illetve a „ne” partikulákkal fejezzük ki, amelyek jelenléte egyszerűen meghatározható egy-egy tagmondatban lexikális alapon.

Éppen ezért, habár a pszicholingvisztika és a PLM is a mondat szintjén kezeli a javaslatot, az implementációs megoldás miatt a jelen felsorolásban az a lexikai szinthez sorolandó.

### *„Prepozíciók (előjárószók) egyértelműsítése”*

A magyarra nem jellemző az előjárószók használata. Az egyetlen kivétel ez alól a „mint” kötőszó, amely ritka esetben viselkedhet a prepozíciókra emlékeztető módon, pl. a *bátyámat mint tanút idézték be* szerkezet esetében. Keszler (2000: 275) szerint ebben az esetben a „mint” kötőszó funkcióját tekintve névutóként viselkedik; nemcsak összeköti a két accusativusban álló főnevet, de a másodikkal morfológiai természetű szerkezetet is alkot, ezáltal alkalmassá válik a mondatrészi szerep jelölésére is.

A „mint” ilyen használata esetén az értelmezést könnyítő javaslat lehet az essivusi alakkal történő helyettesítés (*tanúként*), ezt a konkrét esetet leszámítva azonban a megkötés a magyarra (annak sajátosságai miatt) nem alkalmazható.

## **12.3 Szintaktikai szint**

### *„'vagy – vagy' alkalmazása a kirekesztő választás érzékeltetésére a 'vagy' egyszeri használata helyett”*

A kirekesztő értelmű vagylagos kapcsolatok detektálására egy egyszerű, de hatékony heurisztika lehet a tagmondat határán belül előforduló egyetlen, „vagy” -gyal összekapcsolt szerkezet jelenlétének figyelése.

Tekintettel arra, hogy itt is felmerül a fogalmazói szándék rekonstruálásának kérdése, az ilyen esetekben egy figyelmeztető üzenet jelenhet meg a „vagy – vagy” kapcsolat esetleges

szükségességére vonatkozóan. Ugyanezért a javaslat sokkal inkább értelmezhető a teljes mondat jelentésének tükrében, vagyis nem tartozik egyértelműen kizárólag a lexikai szinthez, ahová a Stíluskönyv eredetileg sorolta.

*„Rövid (tag)mondatok”*

Ahhoz, hogy a „rövid” mondatokkal kapcsolatos elvárást számszerűsíteni tudjuk, egy lehetséges megközelítés annak felmérése, hogy mi számít a nyelvhasználatban átlagos mondathossznak?

A válaszhoz legegyszerűbben különböző korpuszok segítségével juthatunk el. A felmérést nehezíti, hogy sok magyar nyelvű korpusz esetében a leíró adatokon belül csak a teljes szövegállomány tokenzáma érhető el, a mondatoké nem, így a korpusz készítőjével való direkt összeköttetés vagy az adathalmaz újraelemzése<sup>175</sup> nélkül nem becsülhető az adott korpuszban jellemző átlagos mondathossz. Akad azonban több olyan korpusz is, amelyek esetében ismert ez a leíró statisztikai jellemző is, így ezek segítségével valamelyest képet alkothatunk róla, milyen érték tekinthető átlagosnak a mondatok tokenzáma vonatkozóan. Ezek közül néhányat a 18. táblázat ismertet.

| <b>Korpusz</b>                                       | <b>Domén</b>     | <b>Típus</b>         | <b>Méret</b> | <b>Átlagos mondathossz</b> |
|--|------------------|----------------------|--------------|----------------------------|
| Magyar Wikipedia szűrések nélkül (Nemeskey 2020: 96) | köznyelvi        | írott                | ~169M szó    | 12,15 szó                  |
| SzT  | köznyelvi        | írott                | ~1.2M token  | 17,14 token                |
| MJK  | jogi             | írott + beszélnyelvi | ~2.4M token  | 16,21 token                |
| NYTK-NerKor (Simon – Vadász 2021)                    | köznyelvi + jogi | írott                | ~1M token    | 15,22 token                |

18. táblázat: „Átlagos mondathossz” négy kiválasztott korpusz esetében.

<sup>175</sup> Az ilyen fajta elemzés sok esetben egyáltalán nem triviális feladat; tekintve például a huBERT tanításához készített Hungarian Webcorpus 2.0 mintegy 9 milliárd szavas szövegállományát, amelynek elemzési ideje (még ha csak a mondatsegmentálásra fókuszálunk) jelentősen meghaladja (legalábbis nem professzionális környezetben) az erre fordítható időt.

A bemutatott néhány korpusz a bennük szereplő szövegek forrására tekintettel rendkívül heterogén, őket alapul véve nagyjából azt a következtetést vonhatjuk le, hogy írott szövegek esetében az átlagosnak ítélnélhető mondathossz nagyjából 12 – 17 token között mozog.

A jelen dolgozat keretében elkészített szoftver azonban a legjobban akkor működhet, ha figyelembe veszi azokat a konkrét jellemzőket, amelyek az alkalmazási területeként kijelölt domént jellemzik; ez esetben hivatalok tájékoztató szövegeit. Ez a későbbi felhasználás szempontjából nem rontja az általánosítási képességet – más doménre alkalmazva azonos vizsgálatok elvégzése után ugyanolyan hatékony megoldást kaphatunk, mint a rendelkezésre álló konkrét szövegek esetében. Ennek érdekében a konkrét küszöbértéket, amely tokenszám felett az alkalmazás jelzést küld a felhasználó felé, a modelltanításhoz is használt korpusz elemzésével határozható meg.

*„Inkább afirmatív (állító), mint negatív mondatok”*

A tagadás jelenlétének detektálása egy viszonylag triviális heurisztika mentén nagyjából a megfelelő szavak (pl. *nem, ne, se, sem*) jelenléte által megvalósítható. Azonban az újabb szakirodalom nem tartalmaz meggyőző bizonyítékot arra vonatkozóan, hogy a nem állító mondatok feldolgozása lényegesen nagyobb befektetést kívánna meg a befogadó részéről, mint az egyszerű állításoké.

*„Többszörösen összetett konstrukciók (pl.: beágyazott tagmondatok, többszörös alárendelések, birtokos szerkezetláncok) kerülése”*

A mondatok összetettségének meghatározására egy könnyen megvalósítható heurisztika lehet a tagmondatok számának nyomon követése. Bár ez a megoldás komplexitásában elmarad a javaslatban megfogalmazottaktól, a gyakorlati implementáció során szükséges egyensúlyt kialakítani a kód átláthatósága, erőforrásigénye és az alkalmazott megoldás gyakorlati előnyei között.

Tagmondatok detektálásra a közelmúltban kísérletet tett például Dömötör – Nemeskey (2023). Módszerük az egyes természetesnyelvi elemzőkben megtalálható szintaktikai elemző modul kimenetére épít, amelyből néhány egyszerű szabállyal kísérik meg a mondaton belüli egyes tagmondatok határainak feltérképezését az általuk definiált struktúrákkal izomorf részgráfok keresése által. A módszer előnye, hogy (amennyiben a



szintaktikai elemzés rendelkezésre áll és helyes), képes megtalálni az olyan eseteket is, amikor a tagmondatba egy másik tagmondat ékelődik be (pl.: „Az út mentén, ahol sétáltunk, ült kint néhány öreg ember.”).

A módszerrel kapcsolatos kísérletezés során nyilvánvalóvá vált, hogy a fejlesztéshez választott HuSpaCy szintaktikai elemzőmodulja nem alkalmas maradéktalanul a szükséges szintaktikai információk kinyerésére. A szintaktikai információkat is tároló *token* objektum több olyan attribútuma hiányzik, amely más nyelvek (például az angol esetén) elérhető, mint például a dependencia fában a csomópont jobb és bal szomszédai, a csomópont „gyekereinek” listája stb. A rendelkezésre álló adatokból a teljes szintaktikai fa, amelynek bejárásával a tagmondatok a fent ismertetett módon megadhatók lennének, csak külön erre a célra fejlesztett eszközök implementálása után lenne lehetséges. Tekintettel arra, hogy a HuSpaCy egy jelenleg is fejlesztés alatt álló eszköz, ez a későbbiekben változhat. Jelenleg azonban a tagmondat határok meghatározása egyszerű regex alapon valósult meg az ismertetett nehézségek okán.

#### *„Kivételek alóli kivételek - típusú szerkesztés kerülése”*

A feltétel megvalósulása főleg olyan esetekben érhető tetten, amikor (például egy jogszabály esetén) annak érvényesülése alól a jogalkotó kivételeket határoz meg, majd pedig valamilyen speciális rendelkezésben ezeket a gyakran általános eseteket tovább szűkíti, ezáltal olyan eseteket meghatározva, amikor a kivételek valamely típusa mégsem érvényesül.

A gyakorlatban ennek megtalálása inherensen szemantikai természetű probléma, annak ellenére, hogy az érintett szerkezet detektálása több esetben megoldható a mondat szintaktikai elemzésének felhasználásával. A jelzett szemantikai jelleg miatt azonban a rendelkezésre álló eszközökkel automatizált formában mégsem kezelhető hatékonyan.

#### **12.4 Bekezdésekre vagy a teljes szövegre vonatkozó ajánlások**

##### *„Legfontosabb normatív tartalom a mondat élére kerüljön” – „Tételmondat alkalmazása”*

A fenti két ajánlás összefügg, azonban önmagában nézve mindkettőre igaz, hogy ismét rendkívül absztrakt elvárást támaszt egy dokumentum megszövegezése elé. Ennek ellenére,

ha az elvárás tágabb kontextusban értelmezzük, akkor adható olyan NLP megoldás, amely jól közelíti annak célját.

A tartalomkivonatolás (Text Summarization) a nyelvtechnológiának egy olyan területe, amely egy adott szöveg legfontosabb tartalmát hivatott azonosítani, és célja ez alapján automatikus eszközökkel egy összefoglaló elkészítése. Az ilyen összefoglalók alkalmasak például arra, hogy az olvasó a lehető leggyorsabban megérthesse a szöveg témáját, és eldönthesse, hogy az az ő szempontjából releváns-e.

A feladatot leggyakrabban valamilyen felügyelet nélküli gépi tanuló algoritmussal szokás megoldani az elvárt témakörök számosságának előre történő definiálása mellett.

A két fő irányvonal, ahogyan a feladat megoldható, az extraktív (Extractive Summarization – ES) illetve az absztraktív (Abstractive Summarization – AS) összefoglalók készítése (Hahn – Mani 2000). Előbbi esetében a szövegből szó szerint kinyert részletekből készülhet el egy tartalmi összefoglaló, például a leginkább releváns mondatok kiemelésével, míg utóbbi esetében a kinyert információkat (például nyelvgenerálással) új mondatokba szervezve lehet visszaadni.

Míg az ES előnye, hogy garantáltan értelmes mondatot ad eredményül, addig az AS esetén a keletkezett eredmény közelebb állhat ahhoz az intuícióhoz, ahogyan az emberek a konkrét szövegtől elvonatkoztatva összegzik annak lényegi mondanivalóját. Ugyanakkor a megvalósítás korlátai közé sorolható, hogy míg az ES esetében kevésbé erőforrásigényes algoritmusokkal is megoldható a szöveg tartalmilag legjellemzőbb mondatainak kiemelése, addig AS esetén ahhoz, hogy értelmezhető eredményt kapjunk, szükséges lehet nyelvi modellek alkalmazása is.

Mindezekén túl fontos azt is kiemelni, hogy az AS-re jellemző megoldások, habár gyakran lényegesen fejlettebbek, mindeddig nem bizonyultak szignifikánsan jobbnak a „hagyományosabb” ES-alapú megoldásoknál.

A dolgozat keretében elkészített megoldás éppen ezért ES módszerek alkalmazásán alapul. A kiválasztott algoritmus az sklearn KMeans<sup>176</sup> (magyarul: K-közép) volt, amelynek bemeneteként a huBERT modellel vektorizált adatok szolgáltak.

---

<sup>176</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

### *„Rövid bekezdések, témánkénti szövegtagolás”*

A bekezdéshosszok limitálása olyan szövegek esetében, amelyek legfőbb célja a teljeskörű tájékoztatás, nem feltétlenül lenne célravezető. Tekintettel arra, hogy a szövegek felé támasztott egyik legfontosabb elvárás az átlagos befogadó szempontjából vett érthetőség javulása a közérthetőségi szempontú felülvizsgálat során, ezért praktikusabb lehet, ha az egyes témák egy-egy bekezdésen belül maradnak, mintsem, hogy egy „mesterséges” korlát miatt azok (tartalmilag indokolatlan) tagolása váljon szükségessé.

Az egyes témák, valamint a szövegben helyet kapó bekezdések viszonyának meghatározása például topic-modellezéssel lehetséges. Elvi szinten elképzelhető olyan megoldás, amely meghatározza a szövegben megjelenő témákat, valamint az ezekre jellemző szó együtt előfordulásokat, majd pedig ezt összeveti az egyes bekezdések határaival. A gyakorlati probléma a megoldással főként az, hogy a hagyományosan hosszabb szövegekre fejlesztett topic modellező algoritmusok (mint például a pLSA, vagy az LDA<sup>177</sup>) tipikusan rosszul működnek rövid szövegek esetében, hiszen ezek csak korlátozott mértékben tartalmaznak ilyen együtt előfordulási információkat. A probléma megoldása jelenleg is intenzív kutatások tárgya, azonban mindezidáig nem született konszenzus arra vonatkozóan, hogy milyen algoritmus család lehet a legalkalmasabb a probléma megnyugtató kezelésére (Quiang et al. 2022).

### *„Soktagú (mondatközi) felsorolások felbontása listákká”*

Az ilyen jellegű szerkesztési sajátosságok meghatározása automatikusan ismét csak a szintaktikai szerkezet, valamint a szófaji információk együttes figyelembevételével lenne biztosan megoldható. Tekintettel a használt elemző eszközök már tárgyalt korlátaira, a dolgot keretében a feladat nem került implementálásra.

### *„Szemléletes példák alkalmazása”*

A példák elsődlegesen értelmezéskönnyítő szereppel bírnak, szerepük például, hogy a (gyakran elvont) mondanivalót kontextusba helyezték, vagy az absztrakt fogalmakat mindennapi (a befogadó számára is ismerős) kontextusba helyezték.

---

<sup>177</sup> Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation

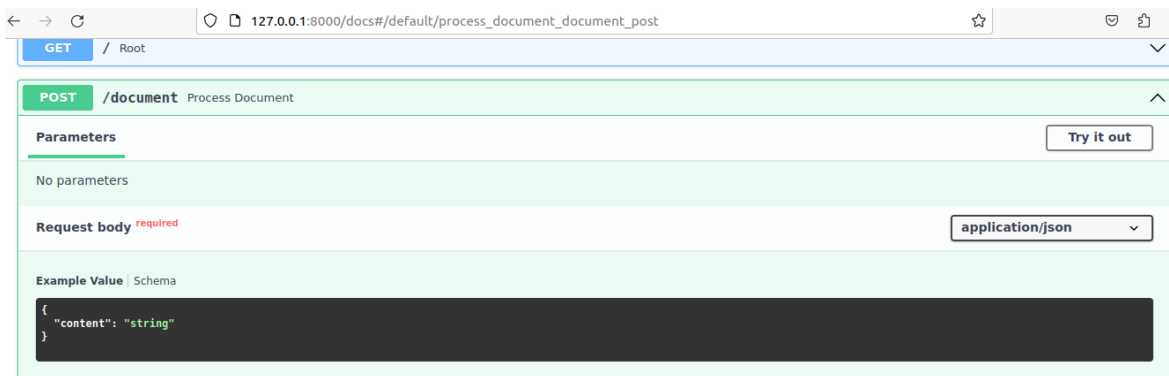
A „példa” lemma, vagy a „pl.” rövidítés jelenléte a szövegben egyértelműen jelezheti, hogy az tartalmaz-e efféle magyarázó részt. Annak megítélése azonban, hogy egy példa mennyiben „szemléletes” inkább szubjektív vélemény kérdése. Kérdéses az is, hogy mi pontosan a szövegnek azon része, amely további magyarázatra, illusztrációra szorul. Mindezen tulajdonságok a szöveg szemantikai tulajdonságaiból következtethetők ki, azonban ez a következtetési folyamat (a szerző véleménye szerint) határozottan emberi intuíciót igényel, ezért például a fent vázolt egyszerű lexikai alapú szabály funkciójában azt nem helyettesíthetné megfelelően.

#### *„Kereszthivatkozások minimalizálása”*

A jogi szövegekben és konkrétan a korpuszban előforduló leggyakoribb kereszthivatkozási típus a törvényekre, valamint kormányrendeletekre történő utalás. Az alkalmazásban ezek szűrése regex alapon lehetséges. Az előforduló hivatkozások kigyűjthetők (akár mondatonként), ezt követően pedig lehetséges a felhasználó figyelmét felhívni ezek lehetőség szerinti kerülésére, vagy az alkalmazott kereszthivatkozások számának minimalizálására.

### 13. A felület és működés bemutatása

Az értekezéshez készített alkalmazás implementálása Python nyelven történt. Az alkalmazással történő kommunikáció a FastAPI<sup>178</sup> keretrendszer segítségével megvalósított, /document endpointon keresztül lehetséges. Lokálisan futtatva a program a uvicorn<sup>179</sup> (szintén) Python nyelvű web szerver implementációt használja.



14. ábra: Az alkalmazás tesztelési felülete. A „content” szöveges mezőt kitöltve az API feldolgozza a kapott szöveget és a választ (az egyes szabályok kimenetét egy-egy kulcs alá rendezve) JSON formátumban adja vissza.

Az alkalmazás fejlesztésekor alapvető elvárás volt, hogy az a lehetőségekhez mérten rugalmas legyen, illetve támogassa tetszőleges frontend hozzáillesztésének lehetőségét. A gyakorlatban ez azt jelentette, hogy a kifejlesztett back end-nek<sup>180</sup> támogatnia kellett a széleskörű integrálhatóságot, valamint, hogy a felhasználónak ne legyen szükséges adott esetben szenzitív adatokat külső szerverre mozgatni.

A FastAPI-ban elkészített alkalmazások futtathatók lokális szerverként, azaz nem szükséges hozzájuk az interneten történő adatmozgatás, ezen felül, mivel önálló API-ként működnek, tetszőleges felhasználói felület (User Interface – UI) illeszthető föléjük. A visszakapott adatok esetleges vizualizálása, az egyes másodlagos funkciók köre ennek megfelelően az alkalmazási helyhez mérten testreszabható, az API működés biztosítja tehát az ehhez szükséges mögöttes absztrakt réteget.

<sup>178</sup> <https://fastapi.tiangolo.com/>

<sup>179</sup> <https://www.uvicorn.org/>

<sup>180</sup> A back end a fejlesztés során a háttérben futó folyamatokra utal, pl. szerveroldali programozás, úrlapon beküldött adatok feldolgozásával, statisztika készítéssel stb. Az itt tárgyalt esetben ide tartozik a teljes alkalmazás, amely fölé tetszőleges honlapot lehet építeni, vagy amely tetszőleges, már létező felhasználói felület mögé integrálható. Szokásos szembeállítás a front end-el történik, amely voltaképpen minden, amit egy weboldalon látunk: funkciók, dizájnelemek, felépítés.

A megoldás előnye, hogy az API-k (iparágilag) standardizált működéssel rendelkeznek, valamint olyan absztrakciós réteget képeznek, amely elrejtí a program implementációja által végbemenő működést, fő hangsúlyuk pedig a kommunikáció támogatása a kód, illetve tetszőleges más kódok között. A gyakorlatban ez két ok miatt jelentős. Egyrészt a már említetteknek megfelelően az elkészült API futtatható lokálisan, vagy bármilyen linux alapú szerver környezetben is. Ez hasznos, amennyiben a feldolgozott dokumentumok szenzitív adatokat tartalmaznak és emiatt azok továbbítása külső szerver felé nem megoldható, vagy csak bonyolult titkosítási protokollok közbeiktatása által lenne lehetséges. Másrészt az API-ként létrehozott backend fölé a felhasználási helytől és igényektől függően saját grafikus felhasználói interfész (GUI), esetleg webalkalmazás telepíthető, annak konkrét működésére vonatkozó ismeretek nélkül is. Mindez olyan fajta flexibilitást biztosít, amely más módon nehezen lett volna csak megvalósítható. Az említett adatbiztonsági megfontolások különösen felértékelődnek például egy hivatali környezetben, ahol az adatok külső helyre történő mozgatása sok esetben szigorúan tiltott is.

Az alkalmazás indításhoz a projekt<sup>181</sup> fő mappájából (`application`) terminálban a `uvicorn application:main:app` parancsot szükséges kiadni, ezt követően a működés a `127.0.0.1:8000/docs` lokális IP-címet egy tetszőleges böngészőbe beírva tesztelhető (a továbbiakban: főoldal).

A főoldalon a `POST / document` fület lenyitva, majd a *Try it out* gombra kattintva érhető el az éles működés. Az alkalmazás a webes környezetben (főleg a böngészőkben napjainkban legelterjedtebb JavaScript alapú kódok alatt) szabványos JSON formátumban várja a feldolgozandó adatot, és a feldolgozás eredményét is JSON-ként adja vissza. Az adatok beküldéséhez a főoldalon mindössze a `content` kulcs utáni szöveg (alapértelmezetten „string”) helyére szükséges a dokumentum szövegét bemásolni, ügyelve arra, hogy sortörés (`\n`) karakterek ne szerepeljenek benne. Ezután az *Execute* gombra kattintani. A szöveg feldolgozását a program automatikusan elvégzi, az eredményben pedig az egyes szabályok kimenete látható.

A fejezet további részében az egyes funkciókhoz tartozó fontosabb implementációs részletek kerülnek ismertetésre.

---

<sup>181</sup> <https://github.com/uveges/Comprehensibility>

### 13.1 Extraktív kivonatolás

Az elkészített alkalmazásban a szövegek tartalmi összefoglalását K-közép (kmeans) algoritmussal valósítottam meg. Az algoritmus a szöveges adatok választott formájú (TF-IDF, fastText, BERT-CLS token stb.) vektorrepresentációján dolgozik felügyelet nélküli módon, a célja pedig, hogy az adatban természetesen jelenlevő csoportosulásokat felmérje, és az egyes vektorokat ez alapján csoportokba sorolja, ezzel partícionálva az eredeti vektorteret.

A kmeans algoritmusnál az egyik legfontosabb probléma az optimális klaszterszám megtalálása, amely legtöbbször csak próbálkozással valósítható meg optimálisan. Az első lépés ebben az esetben a minimális, és maximális megengedett klaszterek számának intervallumos meghatározása, amelyet a konkrét esetben az alábbi függvény segítségével határoztam meg:

$$f(t) = \begin{cases} 1, & \text{ha } t \leq 5 \\ [3,5], & \text{ha } 6 \leq t \leq 100 \\ \left[ 3, c + \left( \frac{d-c}{b-a} \right) (t - a) \right] & \text{ha } t > 100 \end{cases}$$

Itt  $t$  jelenti az adott szöveg szegmentálást követő mondatszámát,  $d$  a kivonat lehetséges maximális mondatszámát,  $c$  a minimális elvárt mondatszámot az összefoglalóban,  $a$  értéke 100-ra lett beállítva (ez a az előző két feltétel nemteljesülése esetén a szöveg lehetséges legkisebb mondatszáma),  $b$  pedig a korpuszban előforduló leghosszabb szöveg mondatszáma (641). A függvény a kiugróan rövid szövegekhez ( $t \leq 5$ ) egy mondatos összefoglalót enged meg, a közepes hosszúságúakhoz ( $6 \leq t \leq 100$  esetben) a kivonat lehetséges hosszát a  $[3,5]$  intervallumon korlátozza. A fennmaradó, hosszú szövegek esetében a maximális mondatszámot egy lineáris görbe  $t$ -beli értéke adja meg, amelynek maximuma a korpuszban mért leghosszabb szöveg (641 mondat) esetében 11; ez jelenti a kivonatok hosszának elméleti maximumát. Ebben az esetben a kapott (valós) értéket a konkrét implementáció minden esetben a legközelebbi egész értékre kerekíti.

A függvény jellemzője, hogy amennyiben ennél a 641 mondatnyi hosszánál nagyobb  $t$  paramétert kap, a lineáris egyenes tulajdonságai miatt a kivonat maximális mondatszámát továbbra is arányosan korlátozza (1000 mondatos szöveg esetében például 14 lenne ez az érték). Az ilyen esetek azonban feltehetőleg rendkívül ritkák a gyakorlati alkalmazás során.

Az értékek fenti beállítását indokló mögöttes intuíció az, hogy rövid szövegek esetében azok feltehetőleg kevesebb témából állnak, így kevesebb releváns mondat választható ki a hozzájuk rendelt kivonatba is, mint hosszabb szövegeknél.  $t$  értékének egyes limitációit a feltételekben a szövegek hosszának gyakorlati elemzése alapján állapítottam meg.

Tekintve, hogy például a BERT modellek esetében a vektorokba kódolt szemantikai információ egyben utal az adott mondatok témájára is, a meghatározott klasztereket tekinthetjük úgy, hogy a szövegben jelenlevő témákat is körülhatárolják.

A lehetséges klaszterek intervallumának megadása után az optimális klaszterek tényleges kiszámítása volt a következő feladat. Ezt a rés-statisztika (gap statistic) algoritmus valósítja meg (Tibshirani et al. 2001). Ennek a megközelítésnek a lényege, hogy módot találjunk a lehetséges klaszterek tömörségének összehasonlítására az adatok null-referenciaeloszlásával, azaz egy olyan eloszlással, amely nem tartalmaz nyilvánvaló klaszteresedést. A klaszterek optimális számának becslése az az érték, amelynél a klaszterkompaktság (azaz az adatpontok sűrűsödése a választott klaszter középpontok – centroidok körül) az eredeti adatokon a legmesszebb esik ezen referencia-görbe alá. Az algoritmus a megadott intervallumon belül minden egész számra végrehajt egy-egy klaszterezést, eltávolítja a referencia-értéktől vett távolságot, majd végül az a klaszterszámot adja vissza, amely optimálisnak bizonyult.

Az ilyen módon meghatározott klaszterekbe tartozó mondatok közül végső lépésben kiválasztásra kerül a centroidhoz legközelebbi mondat (vektor), és ez fog az összefoglalóba kerülni. Az összefoglalót a visszaadott JSON 'summary' kulcsa tartalmazza.

## 13.2 További implementációs részletek

### *Funkcióigék szűrése*

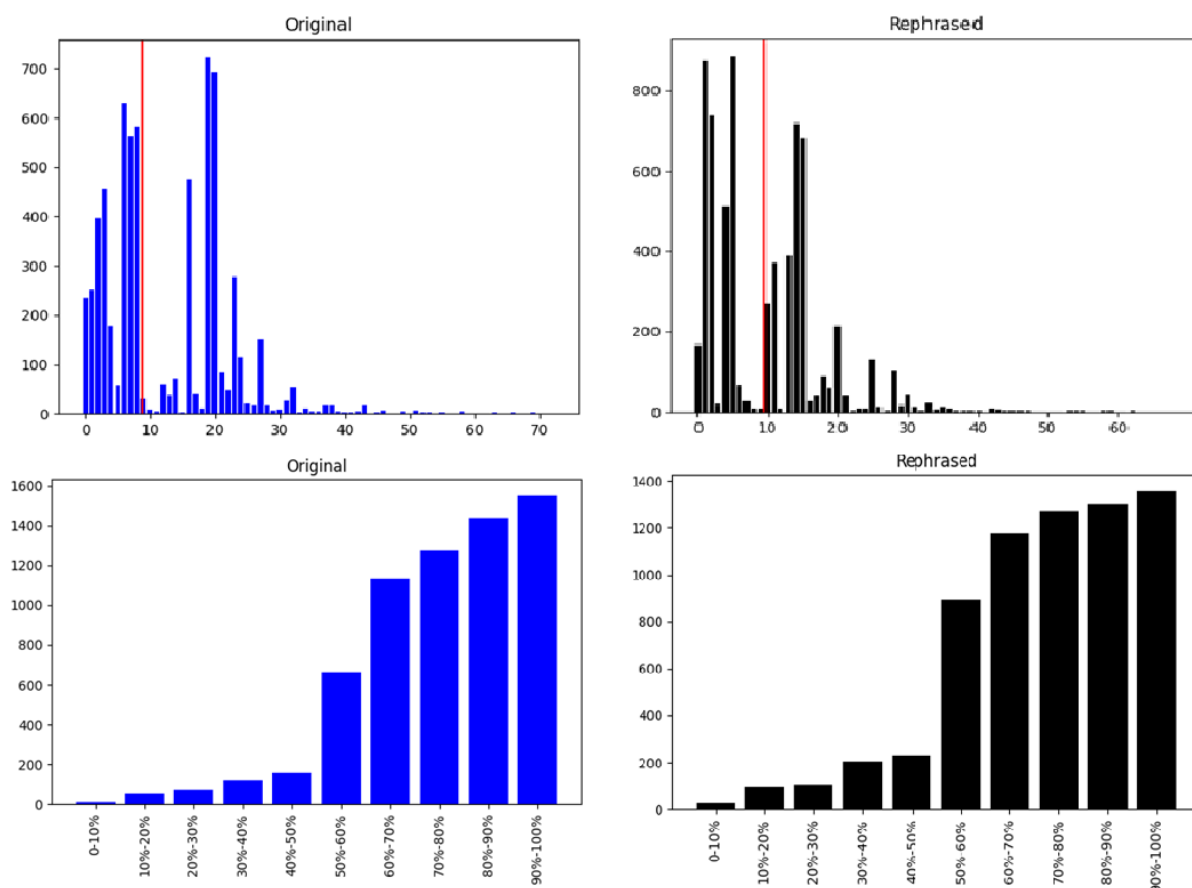
A program a II. Függelékben bemutatott csere lehetőségekre utaló javaslatokat generál a bemenetként kapott szöveghez. A jelenleg keresett alakok között kizárólag jelen idejű, E/3 személyű és kijelentő módú alakok szerepelnek. Ennek oka, hogy az összes lehetőség felsorolásszerű megadása nem hatékony (illetve az emberi tényező miatt feltehetőleg nem is teljes) megoldás, a mód, idő, szám, személy tekintetében minden esetben helyes alakok beillesztése pedig (korszerű módon) már a nyelvmodellekkel történő nyelvgenerálás



feladatkörébe tartozik. Tekintettel arra, hogy a program tervezésekor a lehetséges erőforrásigények minimalizálása fontos cél volt, így ennek megvalósítása (és a szükséges generatív nyelvmodellek projektbe integrálása) nem történt meg.

### *Konkrét / absztrakt vonatkozás*

A szöveg jellemzéséhez ezen a téren a 11. fejezetben említett szótári adatbázis nyújtott segítséget. Az implementáció előtt érdemes volt megvizsgálni, hogyan viszonyul egymáshoz az absztrakt jelentésű főnevek megjelenési gyakorisága az eredeti és a már átfogalmazott mondatokra tekintettel.



15. ábra: A két részkorpuszban szereplő absztrakt vonatkozású főnevek eloszlásának jellemzői

A 15. ábra a két részkorpuszban (eredeti / átfogalmazott mondatok) előforduló absztrakt jelentésű főnevek eloszlását szemlélteti két szempont alapján. Az ábra első oszlopa az eredeti mondatok jellemzőit ismerteti, a második pedig az átfogalmazottakra vonatkozik.

Az első sor ábráin az x-tengely az egy mondaton belül előforduló absztrakt jelentésű főnevek darabszámát jelenti, az y-tengely pedig, hogy ez a gyakoriság hány mondatot jellemez. A függőleges vörös vonal az átlagértéket jelenti; ez az eredeti mondatok esetében 8,85 absztrakt főnév / mondat körül alakult, az átfogalmazottak esetében pedig mintegy 9,65 -re tehető. Ez a tendencia némileg ellentétes a várakozásokkal, hiszen, amennyiben az absztrakt jelentés önmagában nehezíti a szöveg feldolgozását, akkor az átfogalmazás során az ilyen főnevek arányának csökkenése lenne várható.

Az ábra második sora az absztrakt főneveknek a mondaton belüli összes főnévhez viszonyított arányát mutatja. Az eredeti részkorpusz esetében ez az arány átlagosan 75.25% volt, míg az átfogalmazott mondatoknál mintegy 71.93% -ot tett ki, itt tehát már tetten érhető bizonyos mértékű csökkenés.

Jól látható, hogy az átalakítások hatására a közérthetőbbre fogalmazott mondatok részkorpusza esetében számosságot tekintve kevesebb olyan mondat fordul elő, amelyben 80%-nál nagyobb az absztrakt jelentésű főnevek aránya az összeshez képest, majd ez a hatás a 70-80% közötti aránnyal rendelkező mondatok esetében már nem érvényesül. Jól kivehető az is, hogy az ebből az intervallumból kikerült mondatok darabszáma eloszlik a 60% alatti arányt mutató mondatok között, az eltérés jelentős mértékben pedig ennek az intervallumnak a felső tartományában csapódik le.

A két grafikont összeolvasva látható, hogy ugyan az egyes mondatokon belüli absztrakt jelentésű főnevek száma összességében emelkedett, arányuk azonban az összes főnévhez viszonyítva valamelyest csökkent. Fontos lehet továbbá, hogy az átfogalmazott részkorpuszban jelentősen kevesebb olyan mondat szerepel, amelyben 80% feletti lenne az absztrakt főnevek aránya az összes mondatbeli főnévhez képest. Elfogadva a konkrét vonatkozás értelmezéskönnyítő szerepét tehát a gyakorlatban ez azt jelenti, hogy a szinte teljesen absztrakt főnevekkel operáló mondatok számának csökkenése pozitívan hathat vissza az átfogalmazott szövegek befogadhatóságára.

A konkrét implementációhoz szükséges volt valamilyen küszöbértéket megadni, amely felett a rendszer jelezheti az absztrakt főnevek „túlzott” arányát. Tekintetbe kell azonban venni, hogy a jogi, és általában véve a hivatalos szövegek témája maga is sok esetben rendkívül absztrakt. Ha meg akarjuk őrizni a szöveg eredeti jelentésének sértetlenségét, akkor könnyen lehet, hogy ezen a téren a mozgástér rendkívül szűkös, illetve olyan eset is elképzelhető (például egy ügymenet leírása), amikor a mondat főneveinek helyettesítése egyáltalán nem

megoldható. Éppen ezért a program az absztrakt főnevek arányára vonatkozóan 90%-os küszöbértéke elérése esetén figyelmeztet. Ez nagyjából egybevág a szakértők által végrehajtott módosításokkal, ahol is szintén ebben a tartományban következett be a leginkább érdemi csökkenés.

### *Igék – nominalizált természet – igétlen mondatok*

A rendelkezésre álló korpuszban az eredeti mondatok között 785, az átfogalmazottak között pedig 907 igétlen mondat fordult elő<sup>182</sup>. A részkorpuszok számosságára tekintettel ez 12,14%-ot jelent az előbbiekből, illetve 13,61%-ot az utóbbiak esetében. A mondatbeli szavak szófajait tekintve az igék aránya azonban növekedett, az eredeti mondatok esetében átlagosan mintegy 7,36%-ot tett ki, az átfogalmazott mondatok esetében pedig 8,07%-ot, vagyis mondhatjuk, hogy több ige oszlik meg kevesebb mondat között. Ezzel párhuzamosan a szövegekben a főnevek aránya a mondatok átalakításának hatására 30,59%-ról 28,98%-ra csökkent. A fentiek együtt arra utalnak, hogy a korpuszban az igétlen mondatok gyakoriságának növekedése ellenére a főnevek aránya az egyéb szófajok (például kötőszavak, jelzők, határozószók) javára csökkent, amelyek jó eséllyel értelmezéskönnyítő szereppel bírnak. Ez igazolni látszik a nominalizáció csökkentésének hatását a mondatok könnyebb befogadhatóságára, de legalábbis azt, hogy a szakértői revízió figyelembe vette ezt az alapelvet is.

Az elkészült szoftverben a mondatok POS-taggelése után, amennyiben a mondatban nem szerepelt „VERB” címkével ellátott token, akkor az adott mondat a túlzott nominalizációt jelző kulcs alá ('verbless') került a kimenetben.

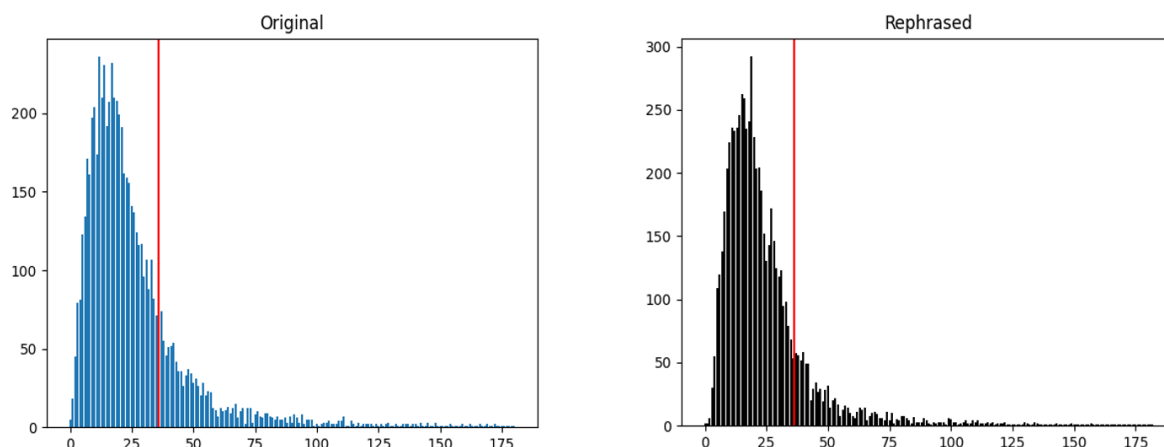
### *Rövidítések*

A 12. fejezetben írtaknak megfelelően a program jelzést küld a felhasználó felé, amennyiben a vizsgálatra küldött szöveg tartalmazza valamely tételt a III. Függelékben szereplő listából. Az érintett rövidítések listája az 'abbreviations' kulcs alatt érhető el.

---

<sup>182</sup> Az arányok meghatározásához a huSpaCy-ben elérhető transformer-alapú nyelvmódellet használva a kiosztott szófajcímkéket vettem alapul.

## Rövid mondatok



16. ábra: Mondathosszok (tokenben mért) alakulása az eredeti és átfogalmazott részkorpuszokban. (Függőleges vonal: átlagos érték.)

Ahhoz, hogy a program futása során valamilyen empirikus küszöbértéket adhassunk meg, amely jelzi a túlzottan hosszú mondatokat, szintén szükséges volt az adatok előzetes értékelése. A mondatok hosszát a bennük megjelenő tokenek száma szerint értelmeztem, amelybe nem számított bele a jelentéssel nem bíró PUNCT és SYM POS-taggal ellátott tokenek száma, valamint azoké, amelyeket a huSpaCy transformer alapú nyelvmodellje X, azaz nem meghatározható szófaji címkével látott el. Az ilyen módon számolt átlagérték  $\sim 35,7$  token volt mondatonként az eredeti, és  $\sim 39,39$  az átfogalmazott mondatok esetében<sup>183</sup>.

Megvizsgálva az adatsorokat jellemző kvartiliseket az látható, hogy míg az eredeti mondatok esetében a Q1, Q2 és Q3 ebben a sorrendben 16, 24 és 36 token szerint alakult, addig az átfogalmazott mondatokra a 15, 22 és 35 értékek voltak jellemzők. Ebből arra következtethetünk, hogy az átlag emelkedését arányaiban kevés számú, viszont a többihez képest kiugróan magas tokenszámmal rendelkező mondat okozhatta. Tendenciaszerűen azonban mégis a mondatok rövidülése volt jellemző a szakértői átalakítás során.

A program a futása során az átfogalmazott korpusz Q3 kvartilisének értéke, azaz 35 token feletti hosszúságú mondatokat jelöl meg rövidítendőként (nem számolva a már említett központozásra vagy szimbólumra utaló POS-taggal, illetve a nem meghatározható szófaji címkével ellátott tokeneket).

<sup>183</sup> A maximum érték 571 token volt az eredeti részkorpuszban, illetve 535 az átfogalmazott mondatok esetében. Az ábra csak a [0-175] tartományt ábrázolja gyakorisági szempontok figyelembevétele miatt.

Az ilyen mondatok esetében a 'too\_long' kulcs értéke az alapértelmezett *False* helyett *True* logikai értéket kap.

### *Kereszthivatkozások*

A kereszthivatkozások szűrése az alkalmazásban a 22) példában ismertetett regex kifejezés segítségével valósul meg.

```
21) (?#törvény)(([0-9]{4}\.\s?évi\s?M{0,3}(CM|CD|D?C{0,3})
    ?(XC|XL|L?X{0,3})?(IX|IV|V?I{0,3})?\.\s? (törvény)?)
    (?#kormányrendelet)|[0-9]{,3}/[12][0-9]{3}\.\s?\
    ((CM|CD|D?C{0,3})?(XC|XL|L?X{0,3})?(IX|IV|V?I{0,3})
    \.\s?[0-9]{1,2}\.\.\s?(Korm(ány|\.)\s?rendelet)
```

A kiválogatott kereszthivatkozások ezután a kimenetben mondatonként egy külön kulcs alá ('abbreviations') kerülnek összegyűjtésre.

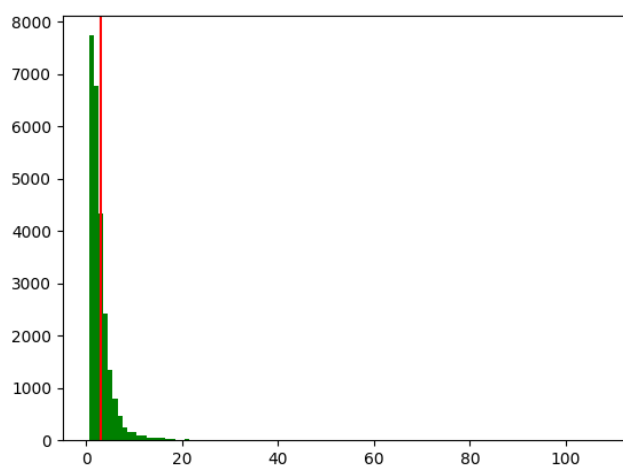
### *Tagmondatokkal kapcsolatos javaslatok*

Az alkalmazásban a tagmondathatárok felismerését egyszerű heurisztika segíti, amely kihasználja, hogy a magyar helyesírás szabályai szerint azokat valamilyen mondatközi írásjel választja el egymástól. Ennek megfelelően az egyes mondatok a (,;) reguláris kifejezés mentén kerültek szétválasztásra.

Ez nem ad teljesen helyes eredményt, tekintve, hogy nem kezeli például az egyes tagmondatok közé beékelődő egyéb tagmondatok kérdését (pl.: „A parkban, ahol tegnap jártam, szép napos idő volt”), azonban az ilyen esetek biztos kezeléséhez a már említett dependencia nyelvtani elemzés helyes és teljes kimenete lenne szükséges, amely a fejlesztés során használt HuSpaCy természetesnyelvi elemzőben jelen pillanatban áll rendelkezésre.

A 17. ábra szemlélteti, hogy a korpusz mondatainak jelentős többségét ezen heurisztika alapján szegmentálva tipikusan a [0,5] intervallumba eső tagmondat számokat kapunk. A teljes szöveg állományban mindösszesen 74633 tagmondat volt fellelhető, amelyek közül

50746 öt, vagy annál kevesebb tagmondatból álló mondat részeként fordult elő<sup>184</sup>. A tagmondatok átlagos száma megközelítőleg 2,97 szerint alakult mondatonként.



17. ábra: A korpuszt jellemző eloszlás az egyes mondatokon belüli tagmondatokra nézve (x: tagmondatok száma, y: ahány mondatban az adott számosság előfordult, függőleges vonal: átlagos érték).

Az alkalmazásban mesterséges korlátként 10 tagmondat került beállításra, ennek túllépése esetén a kimeneti JSON-ben a `'too_much_clauses'` kulcs alá `True` logikai érték kerül, ellenkező esetben `False` (alapértelmezés).

A többszörös tagadás jelenlétét szintén egyszerű megoldással akkor jelzi az alkalmazás, ha bármely két egymást követő tagmondat mindegyikében szerepel a „nem”, „sem” vagy a „ne” tagadószó. Azon mondatok esetében, ahol az algoritmus jelez, a mondathoz tartozó kulcsok közül a `'multiple_negations'` ismét csak `True` értéket vesz fel.

#### *Archaizmusok, kerülendő kifejezések*

Végül az alkalmazás jelzést küld a felhasználó felé, amennyiben a Stíluskönyvben szereplő diszpreferált kifejezések valamelyike (pl. „akként”, „ilyetén”) előfordul a szövegben. A kimeneti JSON `'archaic'` kulcsa alá ez esetben ismét `True` érték kerül.

---

<sup>184</sup> 1 tagmondat: 7741 mondat, 2: 13568, 3: 13029, 4: 9683, 5: 6725 szerint.

### 13.3 Az alkalmazás kimenetének áttekintése

A jelen fejezet röviden összefoglalja az alkalmazás kimeneteként szolgáló JSON főbb strukturális elemeit, egyúttal (a könnyebb áttekinthetőség kedvéért) röviden ismét kitér azok értelmezésére.

A kimenet első szintjén négy kulcsot láthatunk:

- 'content': Az eredeti, feldolgozásra beadott szöveg változatlanul.
- 'problematic': Azon mondatok listája, amelyek a huBERTPlain modell szerint javításra szorulnak.
- 'suggested\_changes': a szöveg mondatainak listája. Amennyiben a mondat nem szerepel a 'problematic' kulcs alatt, akkor utána az „OK!” string szerepel. Ha a modell szerint javítandó, akkor önálló JSON objektumban itt található a javasolt módosítások (lásd lentebb).
- 'Summary': A szövegből készített tartalmi összefoglaló (a 13.1 fejezetnek megfelelően).

A 'suggested\_changes' kulcs alatt, amennyiben tehát a mondatot huBERTPlain átfogalmazandónak ítélte, az alábbi kulcsok jelennek meg a mondat szövege mellé rendezve:

- "light\_verbs": Lista a szövegben talált funkcióiágés szerkezetekkel, valamint az azokhoz javasolt cserékkal (II. Függelék szerint).
- "too\_abstract": true; amennyiben a mondat túl sok absztrakt jelentésű főnevet tartalmaz, false; egyébként.
- "verbless": true; Amennyiben a mondat nem tartalmaz igét, false; egyébként.
- "abbreviations": A mondatban fellelhető rövidítések a III. Függelék szerint.
- "too\_long": true; Amennyiben a mondat túl sok tokent tartalmaz, false; egyébként.
- "references": Jogszabályhivatkozások listája a mondatban.
- "too\_much\_clauses": true; a mondatban túl sok tagmondat szerepel, javasolt a annak szétválasztása több különálló mondattá, false; egyébként.
- "multiple\_negations": true; többszörös tagadás a mondatban, false; egyébként.
- "archaic": true; archaikus kifejezések jelenléte a mondatban, false; egyébként.

Az API fölé megírt tetszőleges GUI esetében a javaslatok a helyzetnek megfelelően felhasználhatók, például pirossal kiemelhetők a problémás mondatok, a funkcióiágés szerkezetek cseréje kattintással automatizálható a visszaadott javaslat alapján stb.

### 13.4 A fejlesztéssel kapcsolatos összefoglaló

Az alkalmazás elkészítésével elsődlegesen a dolgozat bevezetésében felvetett iv. hipotézis igazolása volt a cél, amely szerint a kialakított gépi tanult modellek integrálhatók olyan feldolgozási láncba, amely képes lehet a hivatalos / jogi dokumentumok fogalmazását végzők munkáját segíteni.

Az alkalmazás elkészítése során felmerült nehézségek egy része a rendelkezésre álló NLP eszközökben fennálló hiányosságok miatt merült fel (pl.: a SpaCy magyar nyelvre elérhető transformer-alapú modellje a fejlesztés jelen fázisában nem támogatja a kellő összetettségű szintaktikai elemzést néhány részletezett feladat megoldásához), más esetekben a szakirodalomban fellelhető, a közérthető megfogalmazást segíteni kívánó javaslatok természete nem teszi őket alkalmassá automatizmusokká szervezésre. Ilyen például az olyan, erős pragmatikai és szemantikai szakértői tudás meglétét feltételező elvárás, amely szerint az érintett szövegekben törekedni kell a közismert szókészlet alkalmazására. Annak eldöntése azonban, hogy mi számít a címzettek számára közismertnek, nem feltétlenül egzaktan számszerűsíthető vagy mérhető restriktió. Az ilyen és ehhez hasonló problémák miatt a hipotézis csak részben tekinthető igazoltnak.

Emellett fontos ismét hivatkozni arra a korlátozottságra, amely az elkészített alkalmazást a felhasználási lehetőségek tekintetében jellemzi. A gépi tanult modellek készítéséhez csak egy rendkívül szűk doménből állt rendelkezésre tanítóadat. A Nemzeti Adó- és Vámhivatal Közérthetőségi Munkacsoportja által elkészített dokumentumokon tanított modellek esetében nem várható el, hogy azok más, általános értelemben véve jogi szövegek esetében is hasonló hatásokkal legyenek képesek operálni. Éppen ezért az elkészített alkalmazás felhasználhatósága is (legalábbis annak jelen formájában) erősen limitált marad.

Tekintettel azonban arra, hogy a jelenleg alkalmazott modell könnyen cserélhető, a kézi szabályok kiválasztása során pedig fontos szempont volt azok lehetőség szerinti doménfüggetlensége (a jogi, hivatalos szövegek rétegzettségére tekintettel), így az alkalmazás mégis egyfajta könnyen alakítható prototípust jelent, amely kellő plaszticitással bír a más típusú szövegekre történő könnyű adaptáláshoz is.



## 14. A hipotézisek értékelése és a kutatási kérdésekre adható válaszok összefoglalása

A kutatás kezdetén négy kiinduló hipotézist fogalmaztam meg, amelyek nagyrészt egymásra épülve voltak hivatottak feltérképezni azokat a legfontosabb kérdéseket, amelyek a közérthető fogalmazás automatizálási lehetőségei kapcsán felmerülnek. Ezek közül az első így szólt:

- i. A hazai és/vagy a nemzetközi szakirodalom alapján a közérthetőségnek nemcsak definíciója létezik, de meghatározható nyelvi jellemzők konkrét csoportja is, amely rontja egy szöveg könnyű érthetőségét.

Az i. hipotézis voltaképpen arra vonatkozott, hogy a közérthető fogalmazásnak nem pusztán definícióit / definíciós kísérleteit találhatjuk meg a vonatkozó szakirodalom tanulmányozása során, de olyan konkrét nyelvi / nyelvhasználati jellemzőket is, amelyek a szöveg érthetőségére negatívan hatnak. A vonatkozó kutatási irányzatokat az 1920-as évektől áttekintve (3. fejezet) **az i. hipotézis világosan igazolódott.**

Az olyan általános érvényű meghatározásokon felül, mint például a közérthetőség általánosan elfogadott, nemzetközi definíciója (3. fejezet – bevezetés) az érintett kutatások konkrét kritérium rendszereket is megfogalmaznak. Számos olyan nyelvi jelenség, fogalmazásmódbeli választás és preferált, diszpreferált nyelvhasználati jellegzetesség mutatkozott meg, amelyek segítségével konkrétan azonosítani lehet a szöveg azon pontjait, amelyek nehezítik annak megértését.

- ii. A jogi nyelvnek létezik olyan rétege, amelyben a közérthetőség, mint célkitűzés a gyakorlatban is érvényesíthető.

A ii. hipotézis a gyakorlati alkalmazhatóság alapköve volt. Amennyiben nem találunk a jogi doménen belül olyan szövegtípust, amelyik (főként a hazai kontextusban) alkalmas közérthetőségi javaslatok befogadására, úgy a további vizsgálódás sem léphetett volna tovább az alapkutatási jellegen. Az 5.5 fejezetben megállapítottak szerint ez a szövegtípus a funkcionális szövegek csoportja, amelyek dedikáltan a laikusok felé vannak címezve. Ennek kapcsán konkrét intézményi program is létezik a Nemzeti Adó- és Vámhivatal égisze alatt, amely a jelzett szövegeket közérthetőségi „átvilágítás” után teszi csak elérhetővé a hivatal honlapján. Az ilyen szövegekből konkrét korpusz építése is lehetségessé vált, köszönhetően

a hivatal Közérthetőségi Programjában dolgozó nyelvész szakértőkkel kialakított együttműködésnek (6.2 fejezet). A fentiek alapján **a ii. hipotézis** is egyértelműen **igazoltnak tekinthető**.

- iii. Megfelelő tanítóadatok birtokában lehetséges gépi tanult modellel közérthető, és átalakításra szoruló szövegek automatikus szétválogatása (klasszifikáció).

A iii. hipotézis számos buktatót rejtett magában. Attól ugyanis, hogy egy feladat esetében lehetséges tanítóadatokat gyűjteni, közel sem biztos, hogy a feladat maga olyan, ami gépi tanulási úton is megvalósítható. Számos probléma merülhet fel például a rendelkezésre álló adatok kapcsán, amennyiben azok ellentmondásosak, vagy a probléma túlmutat a kiválasztott modell képességein.

A 7-10 fejezetekben látottak alapján kijelenthető, hogy közérthetőségi osztályozásra képes modell létrehozása lehetséges volt, amit a kipróbált modell család eredményei is igazoltak. A gyakorlati alkalmazhatóság szempontjából azonban kritikus, hogy az elkészített modell milyen hatékonysággal képes operálni. Főként a huBERTPlain esetében volt megfigyelhető, hogy a modell teljesítménye (0.73 macro átlag F1) jelentősen elmaradt a bináris klasszifikációs feladatokban szokásos (jellemzően > 0.9 macro átlag F1) eredményektől. Az elért pontosság és fedés azonban kellően jó, hogy a szakértői munka támogatására alkalmas legyen, amelyet a 11. fejezetben bemutatott, kézi validáció is alátámasztott.

A bináris osztályozási feladatokban megszokottól jelentősen elmaradó teljesítmény tehát együtt járt azzal, hogy a jelenleg elérhető legjobb modell is képes lehet a szakértői munka hatékony támogatására. Emiatt a kettősség miatt **a iii. hipotézis részben tekinthető igazoltnak**. A tanítóadatok tisztítása, esetleg később megjelenő, erőforrásigényesebb modellek alkalmazása ezen változtathat.

- iv. Egy ilyen modell köré lehetséges olyan szoftvert építeni, amely a szöveg fogalmazóját (egyfajta speciális helyesírás-ellenőrzőként) közérthetőségi javaslatokkal tudja támogatni.

A iv. hipotézis főként a i. által feltételezett szabályrendszer, illetve a iii. által elvárt gépi tanuló modell összekapcsolhatóságára reflektál. Egyben konkretizálja azt az elvárást is, hogy a feltárt, közérthetőségi szabályok programozott formában is megvalósíthatók.

Ennek vizsgálatát behatóan a 12-13 fejezetek ismertették. Látható volt, hogy a közérthetőséget támogató szabályok jelentős része végül nem bizonyult kód szinten

megvalósíthatónak. Ennek részben az elégtelen mennyiségű, nyilvánosan elérhető adatbázis (pl. terminológiai adatbázisok), részben a mai NLP eszközök kapacitásának korlátossága volt az oka. Azonban még ezekkel a limitációkkal is megoldhatónak bizonyult olyan rendszer építése, amely a gépi tanult modellel előszűri a kapott szövegeket, majd a modell által problémásnak ítélt esetekben szabályalapon ad javaslatokat azok érthetőbbé tételére. Ilyen szabályok voltak:

- személytelen szerkezetek / nominalizáció szűrése (*szófaji elemzéssel*)
- absztrakt vonatkozású szavak arányának monitorozása (*lexikon alapon*)
- (jogi) rövidítések jelzése – javaslat feloldásra (*lexikon alapon*)
- Jogszabályi hivatkozások keresése – kiszervezés lábjegyzetbe (*regex alapon*)
- archaizmusok, többszörös tagadás, és funkcióigék szűrése (*lexikai alapon*)
- kiugróan hosszú mondatok, és kiugróan sok tagmondat jelzése (*nyelvmodell segítségével, empirikus küszöbértékkel*)
- tartalom összefoglalása a szöveg elején: extraktív kivonat (*felügyelet nélküli gépi tanuló modell segítségével*).

Fontos azt is megemlíteni, hogy a legjobbnak választott gépi tanult (csakúgy, mint minden ML modell) korlátozott felhasználási lehetőségekkel rendelkezik. Ennek a legfontosabb oka az a fajta domén függés, ami a tanítóadatok jellegéből következik. Tekintettel arra, hogy azok között kizárólagosan a NAV közérthetőségi programjából származó funkcionális szövegek szerepelhettek (más adat elérhetetlensége okán), így a modell is csak azonos, vagy nagyon hasonló szövegeken lehet képes hatékonyan működni. Nem várható el tőle például, hogy bírósági határozatok indoklásait is azonos pontossággal legyen képes közérthetőségi szempontból osztályozni.

Részben visszautalva a iii. hipotézis részleges igazoltságának okára, részben pedig hozzávéve az itt ismertetett korlátokat a szabályok implementálása és a gépi tanult modell kapcsán, **a iv. hipotézis ismét csak részben tekinthető igazoltnak.**

Ugyancsak a dolgozat nyitó gondolatai között mindösszesen hét kutatási kérdés merült fel, amelyek mind a kutatás menetét, mind pedig az értekezés felépítését meghatározták. Ezek közül az első a következők szerint hangzott:

- I. A szaknyelvek szoros elvárásrendszerben működnek, a szakma művelői joggal követelik meg tőlük a pontosságot, és a szakmai korrektséget. A közérthetőség ezzel látszólag ellentétesen hat; a szöveget a befogadó (laikus) nézőpontjához igazítja.

Hogyan egyeztethető össze ez a két szempont, ráadásul az utóbbi prioritizálása mellett a jogi doménben?

A két nézőpont közül fontos volt meghatározni, hogy a jogi domén rétegzettségére tekintettel melyik pontosan milyen szövegtípusok esetében tekinthető elsődlegesnek. A pontos megfogalmazás követelménye, vagy éppen a normavilágosság (a büntetőjog esetében) a jogalkalmazó szempontjából értékeli a szövegeket, a közérthetőség pedig az átlagos befogadó szempontjából. Azon szövegtípus meghatározásával (funkcionális szövegek), amely konkrétan a (jog szempontjából) laikus célközönségnek szól, a látszólagos ellentmondás feloldhatónak bizonyult.

II. Ebből következőleg mi a jogtudomány álláspontja egy ilyen intralingvális átfordításról hazai és nemzetközi kontextusban?

A kérdésre erős befolyással bír, hogy azt milyen kulturális közegben tesszük fel. Az Egyesült Államok példáján jól látjuk, hogy az ottani hivatalok minden kommunikációs csatornája, amelyekkel az állampolgárokhoz szólnak, kötelező érvénnyel átesik közérthetőségi felülvizsgálaton. A skandináv államokban nem ritka, hogy konkrét törvénytöredékek is közérthető megfogalmazásban íródnak. Magyarországon ezzel szemben azt láttuk, hogy a hasonló törekvések szórványosak, céljuk általában valamilyen emberi jogi vagy uniós normának való megfelelés.

III. Melyek a jogi domén azon rétegei, amelyekben a közérthető fogalmazás primátusa legitimálható és megvalósítható?

A megfelelő szövegtípus a korábban már említett funkcionális szövegek csoportja, amelyek dedikált célja a témában járatlan, jól meghatározott (például adózói) célközönség tájékoztatása, vagy nekik iránymutatás valamilyen tevékenység végrehajtása kapcsán.

IV. Mely preferált és diszpreferált nyelvi jellemzők jelennek meg a közérthetőséget tárgyaló hazai és nemzetközi szakirodalomban?

A vizsgált szakirodalmi irányzatok megközelítési módjukban (pl. kognitív irányzatok vs. gyakorlati alapokon kiinduló kezdeményezések, mint a PLM) széttagoltnak bizonyultak. Ennek ellenére számos közös nyelvi jellemzőt felsoroltak (pl. többszörösen összetett, hosszú mondatok, szakmára jellemző rövidítések stb.) amelyek az érthetőségre általában, és konkrétan a közérthetőségre is negatívan hatnak. A 12. fejezet táblázatos formában is

összefoglalta ezeket a tekintetben, hogy melyek bizonyultak a leginkább programozási szinten is megvalósíthatónak.

V. Az egyes (felügyelt) gépi tanulási algoritmusok közül melyik működik kellően megbízhatóan, hogy a szakértők munkáját érdemben támogathassa?

A 7-10. fejezet részletesen értékelte a gépi tanulási algoritmusok három jellemző típusába tartozó modellek teljesítményét. Az itt megfogalmazottak alapján a legmegbízhatóbbnak a kontextusfüggő beágyazásokat alkalmazó, a célra finomhangolt BERT modell bizonyult.

VI. Milyen tervezési elvek mentén implementálható egy közérthetőségi ellenőrző / javaslattevő szoftver?

A kérdésre főként a 6., illetve a 13. fejezet volt hivatott választ adni. Az elkészült alkalmazás API jellegét a széles körben való használhatóság biztosítása indokolta, hiszen az így tetszőleges felhasználói interfész (pl. egy webalkalmazás) mögé beilleszthető. Annak érdekében, hogy a „kézzel írt” szabályok nyújtotta átláthatóságot, és a gépi tanult modellekre jellemző általánosítási képességet is kihasználhassa, az alkalmazásban mindkét megközelítés helyet kapott. Az, hogy a gépi tanult modell kimentétől függ a szabályok lefutása, biztosítja a konzisztenciát a két módszer között.

VII. Ha a szakirodalomban léteznek az egyes nyelvi szintekhez kötődő, konkrét javaslatok a közérthetőségre vonatkozóan, akkor ezek közül melyeket, és hogyan lehet algoritmizálni?

A 3. fejezet feltárta az egyes kutatási irányok által az érthetőséghez rendelt konkrét nyelvi jellegzetességeket, a 12. fejezet pedig részletesen ismertette azokat a szabályszerűségeket, amelyek implementálhatónak bizonyultak (utóbbira példát a jelen fejezetben, a iv. hipotézisre adott összefoglaló válasz is tartalmazott).

## 15. Összefoglalás

A dolgozat célja annak vizsgálata volt, hogy egy olyan, elsősorban általánosnak és talán nehezen megfogható tünő fogalom, mint a „közérthető fogalmazás”, megragadható-e valamiképpen konkrét nyelvi jellemzők segítségével. Ezen felül megkerülhetetlen volt annak megállapítása is, hogy ezen támpontok miképpen segíthetnek (egyéb NLP eszközökkel együtt) egy olyan program létrehozásában, amely a közérthetőséget célzó szakértői fogalmazási vagy utólagos felülvizsgálati feladatokat hatékonyan képes támogatni.

A témakör kiindulását a határozott megfogalmazás követelménye jelentette, avagy annak a kérdésnek a felderítése, hogy a jogalkotó szempontjából milyen szerepe van a normaszövegek érthetőségének, a bennük alkalmazott fogalmazásmód koherens és konzisztens voltának. Ezt a kérdéskört a második fejezet járta körül, csakúgy, mint a normavilágosság és a közérthetőség fogalmi párosának relációját, amely elvezetett ahhoz az értelmezéshez, miszerint ami a jogalkotónak a normavilágos szöveg, az az átlagember számára a közérthető fogalmazás.

Tekintettel arra, hogy az egyes tudományterületek jobbára szeparáltan tárgyalták a szövegek érthetőségét befolyásolni képes tényezőket, illetve korábban a szakirodalomban nem valósult meg a témával foglalkozó irányzatok álláspontjainak integrálása, indokolt volt a harmadik fejezetben összegyűjteni mindazon nagyobb álláspontokat, amelyek valamilyen formában állításokat fogalmaztak meg a közérthető kommunikáció mibenléte kapcsán. A tapasztalatok alapján ezen irányzatok nagyjából hasonló lexikai, szintaktikai, illetve szövegszervezési elemeket azonosítanak, mint az értelmezést nehezítő tényezőket.

Annak felméréséhez, hogy az érthető fogalmazást a hivatali kommunikáció középpontjába állító törekvések miképpen juthatnak érvényre a hazai kontextusban, leginkább más országok hasonló kezdeményezéseinek áttekintése jelenthetett támpontot. Ezt a fajta kitekintést a negyedik fejezet volt hivatott megvalósítani. A közérthető fogalmazás szerepét a hivatalos érában az angolszász világban főként a kommunikáció hatékonyságának, valamint a „szolgáltató állam” szemléletmódjának előtérbe kerülése biztosítja, míg a skandináv országok és jellemzően a kontinentális jogszerhez tartozó országok esetében a fő hangsúly inkább annak az állampolgárok demokratikus érdekérvényesítő képességét támogató jellegére tevődik. A könnyen érthető hivatalos / jogi dokumentumokkal kapcsolatos szabályozás jelentős variációt mutat a kérdést szabályozó konkrét törvényekig

(amilyen például Svédországban és az Egyesült Államokban hatályban van) egészen az olyan szórványos kezdeményezésekig, mint amelyek a magyar helyzetet jellemzik (különösen a 2010-es évek lezárulta óta). Ez kiválóan rávilágít egyrészt a jogtudatosság fontosságára az állampolgárok részéről, másrészt az állam cselekvő szerepvállalásának megkerülhetetlen voltára, amennyiben a jog, és általában a hivatalos közlések érthetőségét demokratikus jogként tekintjük az állampolgárok javára.

A szakértői munka támogatása a témában hatékonyan csakis a rendelkezésre álló NLP eszközök minél szélesebb körű felhasználása által valósulhat meg. Éppen ezért a közérthetőségi javaslatok kritikáinak áttekintését követően (ötödik fejezet) a problémát gépi tanulási, és szabályalapú megoldások szempontjából is megvizsgáltam. Előbbit a hatodiktól a tizedik fejezetig tartó rész tárgyalja, utóbbit pedig a tizenkettedik fejezet.

A gépi tanulási modellek terén a klasszikus gépi tanuló algoritmusoktól indulva teszteltem mind a TF-IDF, mind pedig a kontextusfüggetlen és kontextusfüggő beágyazásokat használó modelleket. Utóbbiak közül a huBERT modell finomhangolásával elkészített huBERTPlain jelenti azt a neuronháló alapú modellt, amely a leginkább ígéretes eredményeket volt képes elérni a Nemzeti Adó- és Vámhivatal Közérthetőségi Munkacsoportjától kapott tanítóadatok szétválogatásában, ahol a cél az átfogalmazásra szoruló, valamint a közérthetőségi szempontból már megfelelő mondatok automatikus elkülönítése volt. Ezen elkülönítés után, a problémásnak ítélt mondatok jelentették a bemenetet a kézi szabályrendszer számára, melynek célja a fogalmazási munka támogatása automatikus javaslatokkal.

A tizenegyedik fejezetben a korábban leghatékonyabbnak ítélt gépi tanulási modell kvalitatív értékelésén keresztül mutattam be a felmerülő problémákat a konkrét személyekhez kötődő közérthetőségi intuíció, valamint az általános közérthetőségi szempontú értékítélet rekonstrukciójának tekintetében.

A kézi szabályok implementálása során fontos szempont volt, hogy azok az egyes kutatási irányok (pszicholingvisztika, PLM, jog és nyelv, korpusznyelvészet) által alátámasztottak legyenek. Emellett több esetben gyakorlati limitet jelentettek a jelenleg elérhető, vagy az alkalmazás fejlesztéséhez választott NLP eszközök jelenlegi korlátai. Utóbbira jó példa a spaCy-hez elérhető transformer-alapú nyelvmodell szintaktikai elemzőmodulja. A modell amiatt került kiválasztásra az implementációs fázis korai szakaszában, mivel ez jelenti a magyar nyelvre elérhető elemzők közül az egyik SOTA megoldást, azonban a befejezetlen fejlesztésből adódó következményekkel csak a munka előrehaladott fázisában

szembesültem. Ennek ellenére az elkészített kézi szabályrendszer, amelyet a tizenkettedik fejezet ismertet, alkalmas lehet arra, hogy a huBERTPlain-nel együtt alkalmazva konkrétan adózási kérdésekkel foglalkozó tájékoztató anyagok közérthetőségi szempontú felülvizsgálatát támogassa, a modell cseréje esetén pedig olyan általános eszköz válhat belőle, amely doménfüggetlenül is alkalmazható hasonló célra.

Az elkészített alkalmazást a konkrét felhasználói felülettől független API-ként implementáltam, amely segít elkerülni az (esetleg szenzitív) adatok kényszerű mozgását, mivel lokális szerverként, esetleg docker containerbe csomagolva is futtatható. Az alkalmazás működését, valamint a kimenet értelmezését a tizenharmadik fejezet ismerteti.

A dolgozat a közérthető fogalmazás mibenlétét, automatizálhatóságát olyan új aspektusból tárgyalja, amely korábban a vonatkozó szakirodalomban nem jelent meg. Emellett a disszertáció eredményei számos további kérdést is felvetnek a digitalizáció következtében egyre inkább terjedő automatizáció, valamint az olyan komplex kérdések viszonyáról, mint amilyen a szakértői munka gépi támogatásának lehetőségei a jogi doménben, annak eldöntetlensége, hogy a neuronháló alapú nyelvmodellek milyen nyelvi kompetenciát képesek reprezentálni, valamint, hogy hol húzódik a határ az emberi intuíciót nem nélkülözhető, és az NLP eszközökkel is hatékonyan közelíthető feladatok között.

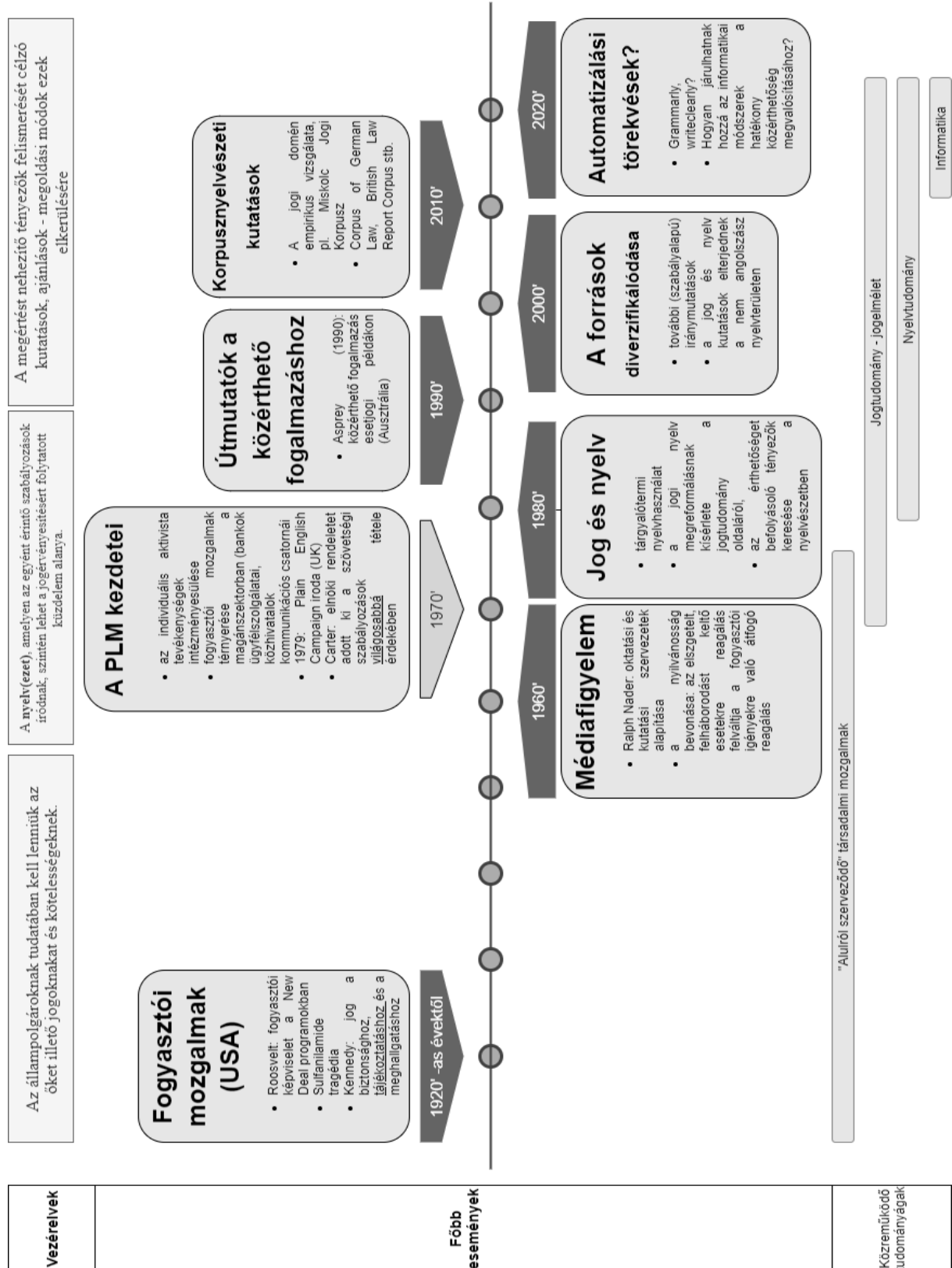
Összességében a dolgozat megkísérelt a jogtudomány, a nyelvtudomány és az informatika interdiszciplináris keretében mozogva egy, a hazai szakirodalomból eddig hiányzó megoldási módszert és szemléletmódot előállítani. Ennek fő eszköze a nyelvtudományi és jogtudományi álláspontok integrálása, illetve az automatizálási, informatikai perspektíva beemelése volt. Az eredmények hasznosak lehetnek akár a hatékonyságnövelésben és a jogállam követelményeinek előmozdításában érdekelt állami szereplőknek, akár a közérthető kommunikációt kutató elméleti szakembereknek. Ehhez főként a téma egy eddig kidolgozatlan megközelítésének bemutatása, illetve a módszer ismert korlátainak felmérése nyújthat segítséget.

Bízom benne, hogy a jelen kutatás ezen felül megkönnyítheti mindazok munkáját, akik a közérthető kommunikáció előmozdítását a jövőben hasonló eszköztárral kívánják elősegíteni.



# Függelék

I.: A PLM egy lehetséges idővonalával a főbb állomásokkal és vezérelvekkel.



## II. Funkcióigés szerkezetek és helyettesítésük

| Funkcióigés szerkezet   | Helyettesítés     | Funkcióigés szerkezet  | Helyettesítés |
|-------------------------|-------------------|------------------------|---------------|
| ad jogot                | feljogosít        | minden alapot nélkülöz | alaptalan     |
| ad számot               | beszámol          | misét tart             | misézik       |
| ad választ              | válaszol          | módosítást eszközöl    | módosít       |
| adja hozzájárulását     | hozzájárul        | mutat csökkenést       | csökken       |
| annyit tesz             | annyit jelent     | mutat eltérést         | tér el        |
| azonos fontossággal bír | azonos fontosságú | mutat emelkedést       | emelkedik     |
| beszámolót ad           | beszámol          | mutat érdeklődést      | érdeklődik    |
| beszámolót nyújt        | beszámol          | mutat fejlődést        | fejlődik      |
| beszélgetést folytat    | beszélget         | mutat fejlődést        | fejlődik      |
| beszélgetést folytat    | beszélget         | mutat növekedést       | növekszik     |
| bír jelentőséggel       | jelentős          | mutat növekményt       | növekszik     |
| bizonyítást nyer        | bebizonyosodik    | mutat változást        | változik      |
| búcsút vesz             | búcsúzik          | nyilatkozatot tesz     | nyilatkozik   |
| célzást tesz            | céloz             | nyomozást folytat      | nyomoz        |
| csökkenést mutat        | csökken           | nyújt beszámolót       | beszámol      |
| döntésre jut            | dönt              | nyújt segítséget       | segít         |
| egyezségre jut          | megegyezik        | pillantást vet         | pillant       |
| előkészületet tesz      | előkészül         | részét képezi          | része         |
| eltérést mutat          | eltér             | ritkaságszámba megy    | ritka         |
| érdeklődést mutat       | érdeklődik        | segítséget nyújt       | segít         |
| eszközöl módosítást     | módosít           | sétát tesz             | sétál         |
| eszközöl változtatást   | változtat         | szabaddá tesz          | kiszabadít    |
| ezt megelőzően          | ez előtt          | számot ad              | beszámol      |
| fejlődést mutat         | fejlődik          | tárgyalást tart        | tárgyal       |
| fogadást köt            | fogad             | tárgyát képezi         | tárgya        |
| folytat beszélgetést    | beszélget         | társalgást folytat     | társalog      |
| folytat beszélgetést    | beszélget         | tart előadást          | előad         |
| folytat nyomozást       | nyomoz            | tart misét             | misézik       |
| folytat társalgást      | társalog          | tart tárgyalást        | tárgyal       |
| fürdőt vesz             | fürdik            | tart titokban          | titkol        |
| háborút visel           | háborúzik         | tart ülést             | ülésezik      |
| haszonnal jár           | hasznos           | tesz célzást           | céloz         |
| hatással bír            | hatású            | tesz előkészületet     | előkészül     |
| hibát vét               | hibázik           | tesz esküt             | esküszik      |
| hoz döntést             | dönt              | tesz jelentést         | jelent        |
| hoz eredményt           | eredményes        | tesz kirándulást       | kirándul      |
| hozzájárulását adja     | hozzájárul        | tesz kivételt          | kivételez     |
| jár eredménnyel         | eredményes        | tesz kizárttá          | kizár         |
| jár haszonnal           | hasznos           | tesz kockára           | kockáztat     |
| jelentést tesz          | jelent            | tesz sétát             | sétál         |
| jogot ad                | feljogosít        | tesz szabaddá          | kiszabadít    |

|                            |              |                             |                   |
|----------------------------|--------------|-----------------------------|-------------------|
| jut döntésre               | dönt         | tesz utazást                | utazik            |
| jut egyezségre             | megegyezik   | tesz vizsgát                | vizsgál           |
| jut megállapodásra         | megállapodik | titokban tart               | titkol            |
| jut megegyezésre           | megegyezik   | ugyanolyan fontossággal bír | ugyanolyan fontos |
| kialakítással bír          | kialakítású  | utazást tesz                | utazik            |
| kidolgozással bír          | kidolgozású  | ülést tart                  | ülésezik          |
| kifogást emel              | kifogásol    | választ ad                  | válaszol          |
| kirándulást tesz           | kirándul     | változtatást eszközöl       | változtat         |
| kísértésbe visz            | megkísért    | vesz búcsút                 | búcsúzik          |
| kivételt tesz              | kivételez    | vesz fürdőt                 | fürdik            |
| kötelező erővel bír        | kötelező     | vesz zuhanyt                | zuhanyozik        |
| kritikát gyakorol          | kritizál     | vét hibát                   | hibázik           |
| külsővel bír               | külsejű      | vet pillantást              | pillant           |
| leltárt felvesz            | leltározik   | videóra vesz                | levideóz          |
| majdhogynem                | majdnem      | visel háborút               | háborúzik         |
| megállapodásra jut         | megállapodik | visz kísértésbe             | megkísért         |
| megegyezésre jut           | megegyezik   | vizsgálódás tárgyává teszi  | megvizsgál        |
| megfontolás tárgyává teszi | megfontolja  | vizsgát tesz                | vizsgál           |
| méretekkel bír             | méretű       | zuhanyt vesz                | zuhanyozik        |

### III. A keresett doménspecifikus rövidítések listája

|            |                    |                  |                       |                  |
|------------|--------------------|------------------|-----------------------|------------------|
| Art.       | Ekho.              | Hjt.             | KÜSZ                  | Szabs. r.        |
| ÁSZ tv.    | EVA                | Hpt.             | Kvt.                  | Szabs. tv.       |
| Ávr.       | Evectv.            | Inytv. vhr.      | Lakás tv.             | Szakhoz tv.      |
| Ávt.       | Evt.               | Itv.             | MRP                   | Szakhoz. tv.     |
| Avtv.      | Evt. vhr.          | Jat.             | Mttv.                 | Szoctv.          |
| Be.        | Felnőttképzési tv. | Jht.             | Munka tv.             | Szövetkezeti tv. |
| Bit.       | Felsőokt. tv.      | Jöt.             | Munkavéd. tv.         | Tao tv.          |
| Bjt.       | Fétv.              | Jstv.            | Munkavéd. vhr.        | Társasházi tv.   |
| Bnytv.     | Fgy. tv.           | Jszbt.           | NAV.                  | Tbj.             |
| Bszi.      | Filmtv.            | Jszr.            | Nbjt.                 | Tbj. vhr.        |
| Bszt.      | Flt.               | Kat.             | Nbtv.                 | Tft.             |
| Btk.       | Fmhtv.             | Kbftv.           | Nek tv.               | Tftv.            |
| Btk.       | Fnyt.              | Kertv.           | Netatv.               | Tny.             |
| Btká.      | Földforgalmi tv.   | Ket.             | Nsztv.                | Tny. vhr.        |
| Btké.      | Fttv.              | Kjnp.            | Nyvtv.                | Tpt.             |
| Büsz.      | Ftv.               | Kjt.             | Nyvtv. vhr.           | Tpvt.            |
| Bv.tv.     | Ftv. vhr           | Kknyt.           | Ob tv.                | Útdij r.         |
| Bv.tvr.    | Géntv.             | Kkt.             | OTÉK                  | Útdij tv.        |
| Civil tv.  | GET                | Kkt. vhr.        | OTSZ                  | Üsztv.           |
| Ctv.       | Gfbt.              | Koncessziós tv.  | Öpt.                  | Ütv.             |
| Csjt.      | Gjt.               | Kötv.            | Ötv.                  | Üvegseb tv.      |
| Csjté.     | Gt.                | Közbesz. tv.     | Pénzmos. tv.          | Vagyontv.        |
| Csódtv.    | Gyer.              | Közbesz. tv.(új) | Postatv.              | Vámtv.           |
| Cst.       | Gyftv.             | Közjegyzői tv.   | Pp.                   | Vbit.            |
| Cst. vhr.  | Gyógyszer tv.      | Közjegyzői vhr.  | Praxistv.             | Vbt.             |
| Eat        | Gyülekezési tv.    | Köznev. tv.      | Ptk.                  | Ve.              |
| Ebktv      | Gyvt.              | Közokt. tv.      | Ptké.                 | VET              |
| Ebtv.      | Harmtv.            | Kpt.             | Regisztrációs adó tv. | Vht.             |
| Ebtv. vhr. | Házaszály          | KRESZ            | Reklámadó tv.         | Vjt.             |
| Efo. tv.   | Helyi adó tv.      | Krt.             | Reklámtv.             | Vtv.             |
| Eho.       | Her.               | Ktdt.            | Rtv.                  |                  |
| Eht.       | Hetv.              | Ktjv.            | Smtv.                 |                  |
| Ehtv.      | Hghvtv.            | Kttv.            | Stabilitási tv.       |                  |
| Einfo tv.  | Hgt.               | Ktv.             | Státusz tv.           |                  |

IV. Rövid kísérletek a ChatGPT chatbottal a közérthető fogalmazás kapcsán (GPT-4 a modell 2024.02.01.-ei állapota szerint)

**Leírás:**

Prompt 1: annak előzetes felmérése, hogy egyedi utasítások nélkül milyen információkkal rendelkezik a nyelvmodell a közérthető fogalmazás (Plain Language) elveit illetően.

**Prompt 1.**

*Mik a legfontosabb elvek, amelyeket egy szöveg újraírása során követni kell, amennyiben a cél a szöveg Plain Language-en történő újraírása?*

**Válasz** (prompt 1.)

A szöveg Plain Language (Egyszerű Nyelv) használatával történő újraírásakor az alábbi alapelveket érdemes követni, hogy a tartalom mindenki számára könnyen érthető és hozzáférhető legyen:

1. **Célközönség megértése:** Ismerd meg, hogy kinek írod a szöveget. Milyen előismeretekkel rendelkeznek? Milyen információra van szükségük? Az írás stílusát és nyelvezetét a célközönséghez kell igazítani.
2. **Egyszerűség:** Használj egyszerű, köznyelvi kifejezéseket és kerülj a szakzsargont, hacsak nem teljesen elkerülhetetlen. Ha muszáj szakmai kifejezéseket használni, mindig magyarázd el őket.
3. **Rövid mondatok és bekezdések:** Írj rövid, lényegre törő mondatokat és bekezdéseket, hogy a szöveg könnyen olvasható és érthető legyen.
4. **Aktív hang:** Az aktív hang használata általában világosabb és közvetlenebb, mint a passzív. Például "Az igazgató jóváhagyta a jelentést" helyett inkább "A jelentést az igazgató hagyta jóvá".
5. **Logikus szerkezet:** A szövegnek logikus és könnyen követhető struktúrával kell rendelkeznie. Használj címsorokat, felsorolásokat és alpontokat, hogy segítsd az olvasót az információk könnyű feldolgozásában.

6. **Konkrétság:** Adjon konkrét példákat és világos utasításokat, hogy az olvasó pontosan tudja, mit kell tennie vagy mire kell figyelnie.
7. **Személyesség:** Az olvasóhoz szólva használhatod a "te" vagy "mi" személyes névmást, hogy barátságosabb és közvetlenebb hangvételt teremts.
8. **Kulcsfontosságú információk kiemelése:** Fontos, hogy a legfontosabb információk könnyen észrevehetőek legyenek. Használj kiemeléseket, félkövért vagy dőlt betűtípust, hogy felhívjad a figyelmet a legfontosabb pontokra.
9. **Konzisztencia:** Tartsd egységesen a terminológiát és a formázást az egész szövegben, hogy az olvasó könnyebben követhesse az előadás menetét.
10. **Tesztelés és visszajelzés:** Ha lehetséges, teszteld a szöveget a célközönségen, és kérj visszajelzést arról, hogy mennyire találták érthetőnek és hasznosnak a tartalmat. Az így kapott információk segíthetnek a szöveg további finomításában.

Ezek az elvek segítenek abban, hogy a szöveged világos, közvetlen és mindenki számára hozzáférhető legyen, függetlenül az olvasók háttértudásától vagy képességeitől.

---

**Leírás:** A Földforgalmi törvény<sup>185</sup> 46§ (1)-(5) bek. egy részletének átírási kísérlete közérthető megfogalmazására, kizárólag a modell előzetes „ismereteire” támaszkodva. Az A., B. és C. válaszokhoz a promptot minden alkalommal új chat ablakban kapta meg a modell, hogy a korábbi kontextust ne vegye figyelembe.

A törvény idézett szövege, amelyek a chatbot minden prompt esetén megkapott:

46. § (1) Az erdőnek nem minősülő föld haszonbérbe adása esetén az alábbi sorrendben előhaszonbérleti jog illeti meg:

a) \* a volt haszonbérelő olyan földművest, illetve mezőgazdasági termelőszervezetet, aki helyben lakónak minősül, illetve amely helybeli illetőségűnek minősül, vagy akinek az életvitelszerű lakáshasználata helye, illetve akinek, vagy amelynek a mezőgazdasági üzemközpontja legalább 3 éve olyan településen van, amelynek közigazgatási határa a haszonbérlet tárgyát képező föld fekvése szerinti település közigazgatási határától közúton vagy közforgalom elől el nem zárt magánúton legfeljebb 20 km távolságra van;

b) az olyan földművest, aki helyben lakó szomszédnak minősül;

c) az olyan földművest, aki helyben lakónak minősül;

d) \* az olyan földművest, akinek az életvitelszerű lakáshasználata helye vagy a mezőgazdasági üzemközpontja legalább 3 éve olyan településen van, amelynek közigazgatási határa a haszonbérlet tárgyát képező föld fekvése szerinti település közigazgatási határától közúton vagy közforgalom elől el nem zárt magánúton legfeljebb 20 km távolságra van;

e) az olyan mezőgazdasági termelőszervezetet, amely helybeli illetőségű szomszédnak minősül;

f) az olyan mezőgazdasági termelőszervezetet, amely helybeli illetőségűnek minősül;

g) \* az olyan mezőgazdasági termelőszervezetet, amelynek a mezőgazdasági üzemközpontja legalább 3 éve olyan településen van, amelynek közigazgatási határa a haszonbérlet tárgyát képező föld fekvése szerinti település közigazgatási határától közúton vagy közforgalom elől el nem zárt magánúton legfeljebb 20 km távolságra van.

(2) \* A közös tulajdonban álló föld esetében a tulajdonostárs tulajdoni hányada egészének vagy egy részének megfelelő terület harmadik személy javára történő haszonbérbe adása esetén az (1) bekezdés b)–g) pontjában meghatározott előhaszonbérletre jogosultakat – az előhaszonbérletre jogosultak sorrendjében – megelőzi a földműves tulajdonostárs.

(3) \* Az (1) bekezdés a) pontjával megegyező ranghelyen illeti meg előhaszonbérleti jog

a) \* azt a helyben lakó földművest vagy helybeli illetőségű mezőgazdasági termelőszervezetet, aki, illetve amely az előhaszonbérleti joga gyakorlását megelőzően legalább 3 éve a föld fekvése szerinti település közigazgatási határától közúton vagy közforgalom elől el nem zárt magánúton legfeljebb 20 km távolságra fekvő településen állattartó telepet üzemeltet, és a haszonbérletének a célja az állattartáshoz szükséges és azzal arányban álló takarmányszükséglet biztosítása és rendelkezik a Fétv. végrehajtására kiadott rendeletben meghatározott állatsűrűséggel

aa) rét, legelő (gyep) művelési ágban nyilvántartott föld haszonbérbe adása esetén szarvasmarhafélék, ló, szamár, öszvér, juh, kecske vagy méh állatfajok vonatkozásában,

ab) szántó művelési ágban nyilvántartott föld haszonbérbe adása esetén az aa) alpontban fel nem sorolt állatfajok, valamint szarvasmarha vonatkozásában, továbbá engedély vagy bejelentés alapján nyilvántartásba vett takarmány-vállalkozásnak minősül;

b) \* a szántó, kert, szőlő, gyümölcsös művelési ágban nyilvántartott föld haszonbérbe adása esetén azt a helyben lakó földművest, vagy helybeli illetőségű mezőgazdasági termelőszervezetet, aki vagy amely számára a haszonbérlet célja oltalom alatt álló földrajzi árujelzővel ellátott termék előállítására vagy ökológiai gazdálkodás folytatására;

<sup>185</sup> 2013. évi CXXII. törvény a mező- és erdőgazdasági földek forgalmáról

c) kert, szőlő, gyümölcsös művelési ágban nyilvántartott föld hasznóbérbe adása esetén azt a helyben lakó földművest, vagy helybeli illetőségű mezőgazdasági termelőszervezetet, aki vagy amely számára a hasznóbérlet célja kertészeti tevékenység folytatásához szükséges terület biztosítása;

d) szántó művelési ágban nyilvántartott föld hasznóbérbe adása esetén azt a helyben lakó földművest, vagy helybeli illetőségű mezőgazdasági termelőszervezetet, aki vagy amely számára a hasznóbérlet célja szaporítóanyag-előállításához szükséges terület biztosítása;

e) \* szántó, szőlő, gyümölcsös, kert művelési ágban nyilvántartott föld hasznóbérbe adása esetén a föld területe legalább fele részének öntözhetőségét biztosító és a földdel alkotórész kapcsolatba kerülő öntözésfejlesztési beruházást végrehajtó földművest vagy mezőgazdasági termelőszervezetet, ha a hasznóbérleti szerződés időtartamának legalább fele részében a végrehajtott öntözésfejlesztési beruházás a számvitelről szóló törvény rendelkezései szerint még értékkel bír.

(4) \* Az (1) bekezdés b)–d) pontjában, valamint a (2)–(3) bekezdésben meghatározott földműves jogosulti csoportokon belül az előhasznóbérletre jogosultak sorrendje a következő:

a) családi mezőgazdasági társaság tagja vagy őstermelők családi gazdaságának tagja,

b) fiatal földműves,

c) pályakezdő gazdálkodó.

(5) \* Földre előhasznóbérleti jog szerződéssel nem alapítható. Az e rendelkezés megsértésével megkötött szerződés semmis.

A fenti törvényszöveget a további példákban minden esetben „§§§” karaktorsor jelöli a prompt szövegében, helytakarékosági okokból, valamint azért, hogy a folyamat átlátható maradjon.



**Prompt 2.**

*Fogalmazd át közérthetőre az alábbi szöveget:*

§§§

**Válasz A. (prompt 2.)**

Ez a szöveg egy jogi szabályozást ír le arról, hogy milyen sorrendben jogosultak az emberek vagy szervezetek az erdőnek nem minősülő földterületek haszonbérleti jogára. Egyszerűbben fogalmazva, ha egy földterületet ki akarnak bérbe adni, vannak bizonyos szabályok, hogy ki kaphatja meg először a lehetőséget:

1. Először az a volt bérlő, aki a környéken él, vagy a környéken működő mezőgazdasági szervezet. Ennek az a feltétele, hogy legalább 3 éve ott éljenek vagy ott legyen a mezőgazdasági központjuk, és a földtől legfeljebb 20 km-re lakjanak vagy dolgozzanak.
2. Azután a helyi lakosok, akik szomszédok is egyben, jogosultak a bérletre.
3. Azok a helyi lakosok, akik nem feltétlenül szomszédok.
4. Azok a mezőgazdasággal foglalkozók, akik legalább 3 éve a földtől legfeljebb 20 km-re laknak vagy dolgoznak.
5. A környéken működő mezőgazdasági szervezetek, amennyiben szomszédok.
6. A környéken működő bármely mezőgazdasági szervezet.
7. Azok a mezőgazdasági szervezetek, amelyek központja legalább 3 éve a földtől legfeljebb 20 km-re van.

Ha a föld több tulajdonosa van, akkor a tulajdonosok előnyt élveznek másokkal szemben, ha bérbe akarják adni a földet.

Ezenkívül különleges esetekben, például ha valaki állattartáshoz akar földet bérelni, vagy bizonyos mezőgazdasági termékeket termelni, akkor előnyben részesülhetnek.

Vannak további szabályok arra vonatkozóan, hogy bizonyos csoportok hogyan részesülhetnek előnyben, mint például családi gazdálkodók, fiatal vagy kezdő gazdálkodók.

Végül, a föld előhaszonbérleti joga nem alapítható szerződéssel, és ha valaki megszegi ezt a szabályt, az adott szerződés érvénytelennek tekinthető.

### Válasz B. (prompt 2.)

Ez a szöveg arról szól, hogy ha egy olyan földet, ami nem erdő, haszonbérbe szeretnének adni, akkor vannak bizonyos szabályok, hogy ki jogosult először a bérletre:

1. Először is, azok a korábbi bérlők, akik:
  - helyi lakosok,
  - helyi mezőgazdasági termelők,
  - vagy legalább 3 éve olyan helyen élnek vagy dolgoznak, ami maximum 20 km-re van a bérbe adni kívánt földtől.
2. Utána következnek azok a szomszédos földművesek, akik szintén a helyszínen élnek.
3. Ha a föld több tulajdonosé, és valamelyik tulajdonos bérbe szeretné adni a saját részét, akkor először a többi tulajdonostársnak van joga a bérletre.
4. Van egy különleges jog is azok számára, akik legalább 3 éve állattartó telepet üzemeltetnek a terület közelében, ha a bérbeadás célja az állattartáshoz szükséges takarmány biztosítása.
5. Különbféle célokra, mint például ökológiai gazdálkodás vagy kertészet, bizonyos földművesek vagy mezőgazdasági termelőszervezetek előnyben részesülnek a bérletre való jogosultságban, ha helyben élnek vagy helyi szervezetek.
6. A földművesek és mezőgazdasági termelőszervezetek között is van egy sorrend, ahol először a családi gazdaságok tagjai, aztán a fiatal földművesek, végül a pályakezdő gazdálkodók jönnek.
7. Végezetül, a földre vonatkozó előhaszonbérleti jogok nem alapíthatók meg szerződéssel, és ha mégis így történik, az ilyen szerződések érvénytelenek.

### Válasz C. (prompt 2.)

Ez a szövegrészlet egy törvényi rendelkezést ír le arról, hogy ha egy erdőnek nem minősülő földterületet haszonbérbe adnak, akkor ki jogosult előnyben részesülni a bérletre. Egyszerűen fogalmazva, ha egy földet bérbe akarnak adni, van egy sorrend, hogy ki kaphatja meg először a lehetőséget a bérletre:

1. Elsőként azok a korábbi bérlők jönnek, akik a terület közelében élnek vagy ott van a mezőgazdasági munkájuk központja, és már legalább 3 éve ott vannak.
2. Ezután azok a helyi lakosok jöhetnek, akik szomszédosak a bérbe adni kívánt földdel.
3. Azután azok, akik szintén a helyben élnek.

4. Következnek azok, akik nem feltétlenül élnek a közelben, de mezőgazdasági központjuk vagy lakóhelyük a terület közelében van, és már legalább 3 éve ott tevékenykednek.
5. Utána azok a mezőgazdasági szervezetek jönnek, amelyek szintén a terület közelében vannak, és szomszédosak vagy helyiek.
6. Végül azok a szervezetek, amelyek legalább 3 éve helyiek.

Ha a föld több tulajdonosé, akkor a tulajdonostársak jogot élveznek először a haszonbérletre, még mielőtt bárki más sorra kerülne.

A törvény részletezi még, hogy bizonyos különleges célú haszonbérletek esetén (például állattartás, ökológiai gazdálkodás, kertészet, szaporítóanyag-előállítás) ki jogosult előnyben részesülni, figyelembe véve a föld művelési ágát és a bérlet célját.

Az előhaszonbérleti jogot illetően a családi mezőgazdasági társaságok tagjai, fiatal földművesek és pályakezdő gazdálkodók élveznek előnyt.

Végül, a törvény kimondja, hogy a földre előhaszonbérleti jogot szerződéssel nem lehet létrehozni, és ha mégis megtörténik, az ilyen szerződés érvénytelen.

**Leírás:** A Földforgalmi törvény<sup>186</sup> 46§ (3) bek. egy részletének átírási kísérlete közérthető megfogalmazására, speciális instrukciók megadása mellett. A chatbot ezúttal külön kontextust és szerepet is kapott a prompt szövegében, valamint külön kiemelve a szöveg néhány elvárt tulajdonságát is.

### Prompt 3.

*Nyelvész szakértő vagy, akinek az a feladata, hogy az elé kerülő szövegeket Plain Language-ben újrafogalmazza. Fontos, hogy a szöveg által közvetített normatív tartalom ne változzon meg az általa előállított szövegben. Éppen ezért az átírást kizárólag a szöveg nyelvtani, lexikai jegyeinek változtatásával végezheted el. Az átírás során törekedj főként a következőkre:*

- *A szöveg elején röviden foglald össze a lényegét!*
- *A szöveget, ha lehetőség van rá, tagold alcímek mentén!*
- *Igyekezz közismert szavakat használni, kerülj a szakzsargont!*
- *Kerüld a passzív mondat szerkesztést, és a túlságosan hosszú, esetleg többszörösen összetett mondatokat!*

*Az átfogalmazandó szöveg a következő:*

§§§

### Válasz A. (prompt 3.)

#### Összefoglalás

Ez a szöveg az erdőnek nem minősülő föld haszonbérbe adásának szabályait írja le. Elsősorban azoknak a földműveseknek és mezőgazdasági termelőszervezeteknek ad előnyt, akik a földtől legfeljebb 20 km-re laknak vagy ott van a mezőgazdasági üzemközpontjuk, és legalább 3 éve ott vannak. A tulajdonostársak és különféle gazdálkodó csoportok, mint családi társaságok tagjai, fiatal és pályakezdő gazdálkodók is előnyben részesülnek bizonyos feltételek mellett.

#### Szabályok részletezése

Elsőbbségi jogosultak

---

<sup>186</sup> 2013. évi CXXII. törvény a mező- és erdőgazdasági földek forgalmáról

- Az előző haszonbérlők, ha a közelben laknak vagy gazdálkodnak.
- A szomszédos földművesek, különösen ha helyben laknak.
- Mezőgazdasági termelőszervezetek, ha helybeli illetőségűek vagy a közelben van az üzemközpontjuk.

#### Különleges esetek

- Ha a föld több tulajdonos között van, a tulajdonostársak előnyt élveznek.
- Bizonyos tevékenységeket végző gazdálkodók, mint állattartó telepek üzemeltetői vagy ökológiai gazdálkodást folytatók, szintén előnyben vannak.

#### Rangsor a jogosultak között

- Családi gazdaságok tagjai.
- Fiatal földművesek.
- Pályakezdő gazdálkodók.

#### További megjegyzések

- Az előhaszonbérleti jogot nem lehet szerződéssel alapítani, és a szabályok megsértésével kötött szerződések érvénytelenek.

Válasz B. (prompt 3.)

## ChatGPT

### Összefoglalás

Ez a szöveg azt írja le, hogy kik kaphatnak elsőbbséget, ha mezőgazdasági területeket szeretnének bérelni. Elsőbbséget élveznek a helyi gazdák és mezőgazdasági szervezetek, különösen azok, akik már hosszabb ideje élnek vagy működnek a terület közelében. Továbbá, speciális előnyben részesülnek az állattartó telepek, bizonyos termékek előállítói, ökológiai gazdálkodók és az öntözési fejlesztéseket végrehajtók.

### Elsőbbséget élvezők a bérleti jogoknál

#### 1. Helyi Gazdálkodók és Szervezetek

- Volt bérlők, akik a közelben élnek vagy működnek, legalább 3 éve.
- Helyben lakó szomszédos gazdálkodók.
- Olyan mezőgazdasági szervezetek, amelyek szintén a közelben vannak.

#### 2. Speciális Előnyök

- Állattartó telepek tulajdonosai, akik a bérlet céljából takarmányt szeretnének termelni.

- Azok, akik védett földrajzi árujelzővel ellátott terméket, vagy ökológiai gazdálkodást szeretnének folytatni.
- Kertészeti tevékenységhez szükséges területet kereső gazdálkodók.
- Szaporítóanyag-előállításához területet keresők.
- Öntözésfejlesztési beruházást tervező gazdálkodók és szervezetek.

### **3. Prioritási Sorrend**

- Családi mezőgazdasági társaságok tagjai és családi gazdaságok tagjai.
- Fiatal gazdálkodók.
- Pályakezdők.

#### **Megjegyzés**

A föld előhasznábérleti joggal való biztosítása nem lehetséges szerződéssel, és ha ezt megsértik, a szerződés érvénytelen.

## Hivatkozott esetek és jogszabályok

2010. évi CXXX. törvény a jogalkotásról (Jat.)

3258/2015. (XII. 22.) AB határozat

1263/B/1993. AB határozat ABH 1994, 672, 673–674.;

534/E/2001. AB határozat,

ABH 2002, 1283, 1288–1289.; megerősítve: 4/2013. (II. 21.) AB határozat, Indokolás [59]

9/1992. (I. 30.) AB határozat

1989. évi III. törvény a gyülekezési jogról

2011. évi CLXI. törvény (Bszi.)

Administrative Procedure Act (434/2003)

Consumer Leasing Act of 1976

Electronic Fund Transfer Act of 1978

Fair Packaging and Labelling Act of 1966

Hazardous Substances Labelling Act of 1960

Kefauver-Harris Amendment (1962)

Magnuson-Moss Warranty-Federal Trade Commission Act of 1973

National Traffic and Motor Vehicle Safety Act of 1966

## Irodalomjegyzék

- Anderson, Ty – Hart-Davis, Guy. *Beginning Microsoft Word 2010*. Springer, 2010.
- Asprey, Michele. *Plain language for lawyers*. Federation Press: Sydney, Australia, 2003.
- Assy, Rabeea. „Can the Law Directly tot he Subjects? The Limitation of Plain Language.” *Journal of Law and Society*, Volume 38, Number 3. (2011): 376-404.
- Balogh Dorka. „Alanyi szerkezetek a magyar jogi nyelvben.” In Szabó Miklós – Vinnai Edina (szerk.) *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.*, 71-94. Bíbor Kiadó: Miskolc, 2018.
- Bandle, Oskar – Braunmüller, Kurt – Jahr, Ernst Hakon – Karker, Allan – Naumann, Hans-Peter (eds.) *The Nordic Languages: An International Handbook of the History of the North Germanic Languages*. Vol. 2. De Gruyter, 2005.
- Barta Botond – Lakatos Dorina – Nagy Attila – Konor Nyist Milán – Ács Judit. “HunSum-1: an Abstractive Summarization Dataset for Hungarian.” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.) *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023*, 231-43. Szegedi Tudományegyetem: Szeged, 2023.
- Blutman László. „Egy empirikus jogértelmezéstan szükségessége.” *Jogtudományi Közlöny*, 2007/1 (2007): 3–13.
- Bowen, Betsy A. – Jereb, Barry – Felsenfeld, Carl. *Plain Language: Principles and Practice*. Wayne State University Press, 1991.
- Bowers, Jeffrey S. – Davis, Colin J. „Is speech perception modular or interactive?” *Trends in Cognitive Sciences*, Vol. 8, Issue 1 (2004): 3 -5.
- Britton, Bruce K. – Gülgöz, Sami. „Using Kintsch’s computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures.” *Journal of Educational Psychology*, 83 Vol. 3 (1991): 329–45.
- BMAS – Bundesministerium für Arbeit und Soziales (ed.) „Leichte Sprache. Ein Ratgeber.” Elérés: 2022.06.17. <https://www.bmas.de/DE/Service/Publikationen/Broschueren/a752-leichte-sprache-ratgeber.html;jsessionid=EC26A3B117EFAABC1DA541BED2874867.delivery2->



[replication#:~:text=Leichte%20Sprache%20hilft%20vielen%20Menschen,von%20Texten%20in%20Leichter%20Sprache.](#)

Budai Balázs Benjámín. „Lakossági elégedettség és bizalom függvényei a kommunikáció során” *Új magyar közigazgatás* 10 (4) (2017): 1-10.

Bussmann, Hadumod (Ed.) *Dictionary of Language and Linguistics*. Routledge: London – New York, 1998.

Byrd, Dani – Toben, H. Mintz. *Discovering speech, words, and mind*. Wiley-Blackwell: Chichester, West Sussex, 2010.

Carnap, Rudolf. „Jelentés és szinonímia.” Horányi Özséb – Szépe György (szerk.) *A jel tudománya*, 117-26. Gondolat Kiadó: Budapest, 1975.

Chambers, Jack K. „Sociolinguistics.” In: Ritzer, George (Ed.) *The Blackwell Encyclopedia of Sociology*. John Wiley & Sons Ltd., 2015.

Charrow, Robert P. – Charrow, Veda R. “Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions.” *Columbia Law Review*, 79 (1979): 1306-74.

Charrow, Veda R. – Erhardt, Myra K. – Charrow, Robert P. *Clear & Effective Legal Writing*, 4th ed. Aspen Publishers: New York, 2007.

Chauhan, Priyavrat. „Sentiment analysis: A comparative study of supervised machine learning algorithms using rapid miner.” *International Journal for Research in Applied Science and Engineering Technology* V. (2017): 80–89.

Chase, Stuart – Schlink, Frederick John. *Your Money's Worth: A Study in the Waste of the Consumer's Dollar*. Macmillan, 1927.

Chiche, Alebachew – Yitagesu, Betselot. „Part of speech tagging: a systematic review of deep learning and machine learning approaches.” *Journal of Big Data* 9 (2022)

Chomsky, Noam. *Syntactic Structures*. The Hague: Mouton, 1957.

Chomsky, Noam. *Rules and Representations*. Columbia University Press: New York, 1980.

Chow, Wing-Yee – Smith, Cybelle – Lau, Ellen – Phillips, Colin. „A ‘bag-of-arguments’ mechanism for initial verb predictions.” *Language, Cognition and Neuroscience*, 31(5) (2016): 577–96.

- Cohen, Jacob. "A coefficient of agreement for nominal scales." *Educational and psychological measurement* 20.1 (1960): 37-46.
- Connaster, Bradford F. „Last rites for readability formulas in technical communication.” *Journal of technical writing and communication*. 29, no. 3 (1999): 271-87.
- Cutts, Martin. *Oxford Guide to Plain English*. Oxford University Press, 1995.
- Cutts, Martin. *The Plain English Guide*. Oxford University Press, 1999.
- Dale, Edgar – Chall, Jeanne S. „A formula for predicting readability.” *Education Research Bulletin*, 27 (1948): 37-54.
- Danyi Magdolna. „A tagolatlan vagy egy főtagú hiányos nominális mondat Pilinszky János költői nyelvében.” *Hungarológiai Közlemények* 2010/2. (2010): 93-103.
- Devlin, Jacob – Chang, Ming-Wei – Lee, Kenton – Toutanova, Kristina. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. Association for Computational Linguistics (ACL): Minneapolis, Minnesota, 2018.
- Dobos Csilla. „Állítmányi szerkezetek a magyar jogi nyelvben” In Szabó Miklós – Vinnai Edina (szerk.) *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.*, 37-69. Bíbor Kiadó: Miskolc, 2018.
- Dorney, Jacqueline M. “ERIC/RCS Report: The Plain English Movement.” *The English Journal* Vol. 77, No. 3 (1988): 49-51.
- Dömötör Andrea – Nemeskey Dávid. “Tagmondatokra bontás és NP-chunking függőségi alapon.” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.). *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, 457-69. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2023.
- Dubay, William H. *The Principles of Readability*. Costa Mesta: Impact Information, 2004.
- Edward Gibson – Neil J. Pearlmutter. „Constraints in sentence comprehension.” *Trends in Cognitive Sciences* 1998/7 (1998): 262–68.

Ettinger, Allyson. „What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models.” *Transactions of the Association for Computational Linguistics* 2020; 8 (2020) 34–48.

European Commission: „English Style Guide.” Elérés: 2021.07.20.

[https://ec.europa.eu/info/sites/info/files/styleguide\\_english\\_dgt\\_en.pdf](https://ec.europa.eu/info/sites/info/files/styleguide_english_dgt_en.pdf)

Farrall, Melissa Lee: *Reading Assessment: Linking Language, Literacy, and Cognition*. John Wiley & Sons, 2012.

Federmeier, Kara D. – Kutas, Marta. „A rose by any other name: Long-term memory structure and sentence processing.” *Journal of memory and Language*, 41(4) (1999): 469–95.

Fekete Balázs. „A kortárs 'jogi kultúra' fogalom eredete: Lawrence M. Friedman a jogi kultúráról.” *Iustum Aequum Salutare* 13. évf. 1. szám (2017): 37-53.

Felsenfeld, Carl –Cohen, David S. – Fingerhut, Martin. „The Plain English Movement in the United States: Comments.” *Canadian Business Law Journal* 6 (1981): 408-52.

Felsenfeld, Carl – Siegel, Alan. *Writing contracts in plain English*. West: St. Paul: Minnesota, 1981.

Ficsor Krisztina. „A normák határozott megfogalmazásának problémája a büntetőjogban – A normavilágosság fogalma az Alkotmánybíróság döntéseiben és a bírói gyakorlatban.” *Pro futuro* 2018/3 (2018): 37-59.

Finch, James E. „A history of the consumer movement in the United States: its literature and legislation.” *Journal of Consumer Studies and Home Economics* 9. (1985): 23-33.

Fischler, Ira – Bloom, Paul A. – Childers, Donald G. – Roucos, Salim E. – Perry Jr., Nathan W. „Brain potentials related to stages of sentence verification.” *Psychophysiology*, 20(4) (1983): 400–9.

Frege, Gottlob. „Über Sinn und Bedeutung” *Zeitschrift für Philosophie und philosophische Kritik* Vol. VIII, Issue 2. (1892): 25–50.

Friedman, Lawrence M. *The Legal System. A Social Science Perspective*. Russel Sage Foundation: New York, 1975.

- Forche, Christian F. “On the emergence of Euro-English as a potential European variety of English: attitudes and interpretations.” *Jezikoslovlje* 13. (2012): 447–78.
- Gabig, Cheryl Smith. “Telegraphic Speech.” In Volkmar, F.R. (eds) *Encyclopedia of Autism Spectrum Disorders*. 3076-77. Springer: New York, NY., 2013.
- Garner, Bryan A. *Legal Writing in Plain English*. University of Chicago Press: Chicago, 2001.
- Gellér Balázs József. *Legality on Trial. A Theoretical Analysis of the Legality of Substantive Criminal Norms*. Eötvös University Press: Budapest, 2012.
- Goodfellow, Ian - Jean Pouget-Abadie - Mehdi Mirza - Bing Xu - David Warde-Farley - Sherjil Ozair - Aaron Courville - Yoshua Bengio. „Generative adversarial networks.” *Communications of the ACM*, Volume 63, Issue 11 (October 2020): 139-44.
- Google, Gemini Team: „Gemini: A Family of Highly Capable Multimodal Models.” arXiv. 2023. <https://doi.org/10.48550/arXiv.2312.11805>
- Gosztolya Gábor – Tóth László. „Spoken term detection based on the most probable phoneme sequence.” In *Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on Applied Machine Intelligence and Informatics*, 101-6. IEEE, 2011.
- Grasso, Arianna. „Plain English and the EU: Still Trying to Fight the Fog?” In Marino, Silvia – Biel, Łucja – Bajčić, Martina – Sosoni, Vilelmini (eds.) *Language and Law*. 359-76. Springer: Cham, 2018.
- Grice, H. Paul. „Logic and conversation,” In D. Davidson and G. Harman (eds.) *The Logic of Grammar*, Encino: Dickenson, 1967.
- Grice, H. Paul. *Studies in the Way of Words*, Cambridge, Mass: Harvard University Press, 1989.
- Guo, Yue – Qiu, Weijian – Wang, Yizhong – Cohen, Trevor A. „Automated Lay Language Summarization of Biomedical Scientific Reviews.” Elérés: 2022.11.10. ArXiv abs/2012.12573
- Gyuris Beáta – Maleczki Márta – Varasdi Károly. *Formális szemantika*. JATEPress: Szeged, 2008.

Hahn, Udo, – Inderjeet Man. „The challenges of automatic summarization.” *IEEE Computer* 33 (11), (2000): 29-36.

Heltai Pál - Nagy Róbert. „Nyelvművelés Angliában és Amerikában: A Plain English Campaign.” *Magyar Nyelvőr* 121. évf. 4. sz. (1997): 396-409.

Hettinger Sándor: *A jogi nyelvi közvetítés buktatói a többnyelvű európai uniós környezetben.* In: *Comparative Law Working Papers*, (4) 1. (2020) 1-7.

Hohmann Balázs. *Az átláthatóság értelmezése és követelményrendszere a közigazgatási hatósági eljárások tükrében.* Novissima Kiadó: Budapest, 2022.

Holder, Jeremy. *Ashworth's Principles of Criminal Law.* Oxford University Press: Oxford, 2016.

Howard, Jeremy – Ruder, Sebastian. „Universal Language Model Fine-tuning for Text Classification” In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328-39. Association for Computational Linguistics, 2018.

Hunyadi László - Bertok Kornél - Németh T. Enikő - Szekrenyes István - Abuczki Ágnes - Nagy Gábor - Nagy Norbert - Németi Péter - Bodog Alexa. „The outlines of a theory and technology of human-computer interaction as represented in the model of the HuComTech project.” In *2011 2nd International Conference on Cognitive Infocommunications, CogInfoCom*, 2011.

Imre Attila. „Jogi szakszövegek és terminológiai adatbázisok.” *Hungarológiai Közlemények* 2014/4. Sapientia Erdélyi Magyar Tudományegyetem, Bölcsészettudományi Kar: Újvidék (2014): 13-23.

International Organization for Standardization. *ISO 24495-1:2023 - Plain Language*, 2023 (<https://www.iso.org/standard/78907.html> Elérés: 2024.02.05.)

Jakab András. „A norma szerkezetének vizsgálata.” *Jogelméleti Szemle* [<http://jesz.ajk.elte.hu>] 2001/4., 2001.

Jiao Sun - Q. Vera Liao - Michael Muller - Mayank Agarwal - Stephanie Houde - Kartik Talamadupula -Justin D. Weisz. „Investigating Explainability of Generative AI for Code through Scenario-based Design” In *27th International Conference on Intelligent User*

*Interfaces (IUI '22)*. 212–28. Association for Computing Machinery, New York, NY, USA, 2022

Johannessen, Marius – Berntzen, Lasse – Ødegård, Ansgar. „A Review of the Norwegian Plain Language Policy.” In *16th International Conference on Electronic Government (EGOV), Sep 2017, St. Petersburg, Russia*. 187-198. Springer, 2017.

Johansson, Stig: *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo, 1978.

Jordan, Michael I. – Mitchell, Thomas. „Machine learning: Trends, perspectives, and prospects” *Science*, Vol 349, Issue 6245 (2015): 255-260.

Joulin, Armand – Grave, Edouard – Bojanowski, Piotr – Mikolov, Tomas. „Bag of Tricks for Efficient Text Classification.” In *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (2016)*: 427-31.

Jurafsky, Daniel – Martin, James H. „Speech and Language Processing (3rd ed. draft)” Elérés: 2022.07.10. <https://web.stanford.edu/~jurafsky/slp3/>

Kaczmarek, Karolina: „Az utasítás kifejezőeszközei a jogi szakszövegben a deontikus modalitás tükrében.” *Filológia* 2010/4 (2010): 224-38.

Kallett, Arthur – Schlink, Frederick J. *100,000,000 Guinea Pigs*. Vanguard Press: New York 1933.

Kanungo, Tapas – Orr, David. „Predicting the readability of short web summaries.” In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*. 202–11. Association for Computing Machinery: New York, NY, USA, 2009.

Karsai Krisztina. *Az alapelvek rendszere az európai büntetőjogban* (MTA doktori értekezés) Szeged, 2015.

Kas Bence – Lukács Ágnes. „Processing relative clauses by Hungarian typically developing children.” *Language and Cognitive Processes* 27. (2012): 500–38.

Kálmán László – Trón Viktor. *Bevezetés a nyelvtudományba. 2.*, bővített kiadás, Tinta Kiadó: Budapest, 2007.

Kemény Gábor. „Az irodalmi impresszionizmus nyelvi sajátosságai és Krúdy Gyula stílusának impresszionista jellege.” In *Bonarum Cultores Artium*. 305-14. Miskolci Egyetem BTK ATTI: Miskolc, 2020.

Keszler Borbála (szerk.) *Magyar grammatika*. Nemzeti Tankönyvkiadó: Budapest, 2000.

Kimble, Joseph. *Writing for Dollars, Writing to Please: The Case for Plain Language in Business, Government, and Law*. US: Carolina Academic Press, 2012.

Kjærgaard, Anne. „The Organisation of the Plain Language Movement in Denmark.” In Pirkko, Nuolijärvi – Stickel, Gerhard (eds.) *Language Use in Public Administration — Theory and Practice in the European States*. 123-34. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences, 2015.

Klare, George R. „A second look at the validity of readability formulas.” *Journal of Reading Behavior* 8 (2) (1976) 129-52.

Klenk Márk – Tamás Dóra Mária – Polácska Edina. „A hiteles fordítás terminológiai támogatásának lehetőségei” In Szoták, Szilvia (szerk.) *A hiteles fordítás, mint közfeladat* 113-130. Budapest, Magyarország: Országos Fordító és Fordításhitelesítő Iroda Zrt. (OFFI), 2020.

Kuperberg, Gina R. –Jaeger, T. Florian. “What Do We Mean by Prediction in Language Comprehension?” *Language, Cognition and Neuroscience* 31 (1) (2016): 32–59.

Kurtán Zsuzsa. „Központosítás a magyar jogi nyelvben” In Szabó Miklós – Vinnai Edina (szerk.): *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.* 153-79. Bíbor Kiadó: Miskolc, 2018.

Kurtán Zsuzsa – Tamás Dóra – Vinnai Edina. „Konferenciabeszámoló: Az igazságügyi nyelvészet és társterületei (2019. november 6.)” *Magyar Jogi Nyelv 2020: 1* (2020): 34-37.

Landqvist, Mats – Spetz, Jennie. „Ten Years with the Swedish Language Act.” *Current Issues in Language Planning*, Volume 21, Issue 5 (2020): 532-47.

Lanstyák István. „A funkcióigés szerkezetek néhány általános kérdéséről.” *Társadalomtudományi Szemle*, XXI. évfolyam 2019/4. (2020): 61–91.

*Law Reform Commission of Victoria on Plain English and the Law*. Melbourne: Law Reform Commission of Victoria (1986)

Liu, Yinhan – Ott, Myle – Goyal, Naman – Du, Jingfei – Joshi, Mandar – Chen, Danqi – Levy, Omer – Lewis, Mark – Zettlemoyer, Luke – Stoyanov, Veselin. „RoBERTa: A Robustly Optimized BERT Pretraining Approach.” Elérés: 2023.01.12. ArXiv abs/1907.11692

Lindquist, Hans – Levin, Magnus. *Corpus Linguistics and the Description of English* (2nd edition) Edinburgh University Press, 2018.

Leuckert, Sven – Neumaier, Theresa. “Copula deletion in English as a Lingua Franca in Asia.” *10plus1: Living Linguistics 2* (2016): 86-103.

Lutz, Benedikt. *Plain Language: „An Important Basis of E-Democracy and Open Government”* In Parycek, Peter – Edelmann, Noella (eds.): *CEDEM 16 - Conference for E-democracy and Open Government*. Danube University: Krems, Austria (2016): 165-75.

Markovich Réka. *Deontic Logic and Formalizing Rights*. (PhD thesis), 2017.

Markovich Réka: „A jogszabályok logikai mélystruktúrája” In Szabó Miklós – Vinnai Edina (szerk.) *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga*. Miskolc, ME – MAB, 2018. május 25., 181-98. Bíbor Kiadó: Miskolc, 2018.

Maxwell, Martha. „Readability: Have we gone too far?” *Journal of Reading Behavior* 21 (1978): 525-530.

Márki Dávid – Szaniszló Réka (szerk.). *Közjogi fogalmak közérthetően*. Szeged: Iurisperitus Kiadó, 2020.

Mellinkoff, David. *The Language of the Law*. Little, Brown and Co.: Boston, 1963.

Mertz, Elizabeth. *The Language of Law School: Learning to "Think Like a Lawyer"*. Oxford University Press: New York, 2007.

Mesmer, Heidi Anne E. *Tools for Matching Readers to Texts: Research-Based Practices*. The Guilford Press, 2008.

Miháltz Márton – Csaba Hatvani – Judit Kuti – Szarvas György – Csirik János – Prószéky Gábor – Váradi Tamás. „Methods and Results of the Hungarian WordNet Project.” In *Proceedings of The Fourth Global WordNet Conference*, 311–21. Szeged, Hungary, 2008.

Mike, Lewis – Yinhan, Liu – Naman, Goyal – Marjan, Ghazvininejad – Abdelrahman, Mohamed – Omer, Levy – Veselin, Stoyanov – Luke, Zettlemoyer: „BART: Denoising



Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880. Association for Computational Linguistics (ACL), 2020.

Minya Károly - Vinnai Edina. „Hogyan írjunk érthetően? Kilendülés a jogi szaknyelv komfortzónájából.” *Magyar Jogi Nyelv* 2018/1 (2018): 13-18.

Moon, Rosamund. *Words, Grammar, Text: Revisiting the Work of John Sinclair*. John Benjamins Publishing Company, 2009.

Murawski, Thomas A. *Writing Readable Regulations*, Carolina Academic Press: Durham, NC, 1999.

Nader, Ralph. *Unsafe at Any Speed: The Designed-in Dangers of the American Automobile*. Grossman Publishers, 1965.

Nagy Ferenc. „A nullum crimen/nulla poena sine lege alapelvről.” *Magyar Jog* 1995/5 (1995): 257–270.

Nagy Ferenc: *Anyagi büntetőjog – Általános rész I.*, Iurispertius Bt.: Szeged, 2013.

Nemeskey Dávid Márk. “Natural Language Processing methods for Language Modeling.” PhD thesis. Eötvös Loránd University, 2020.

Neumann, Stella. „Improving the Comprehensibility of German Court Decisions.” In Grewendorf, Günther – Rathert, Monika (eds.): *Formal Linguistics and Law*. 55–80. Mouton de Gruyter: Berlin, 2009.

Nord, Andreas. „Plain Language and Professional Writing: A research Overview.” Institutet för språk och folkminnen, Report, Elérés: 2020.07.20.

<https://www.gu.se/english/research/publication?publicationId=273574>

Novák Attila – Novák Borbála: „MILQA kérdés-válasz benchmark adatbázis.” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.) *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023*, 203-16. Szegedi Tudományegyetem: Szeged, 2023.

Ogden, Charles Kay – Richards, Ivor Armstrong. *The Meaning of Meaning*. Harcourt Brace Jovanovich: Florida, 1923.

Olivares-Beltrán, Andrea – Morell, Teresa. „The Role of English as a Lingua Franca in European Projects: The Case of the RSGAE Project.” *EPiC Series in Language and Linguistics*. Vol. 2. (2017): 135–42.

OpenAI: “GPT-4 Technical Report.” arXiv. 2023.

<https://doi.org/10.48550/arXiv.2303.08774>

Oravecz Csaba – Váradi Tamás – Sass Bálint: „The Hungarian Gigaword Corpus.” *Proceedings of LREC 2014*, 1719-23. (2014)

Oravecz Csaba – Sass Bálint – Váradi Tamás. „Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében.” In Tanács Attila – Varga Viktor - Vincze Veronika (szerk.): *MSZNY 2015, XI. Magyar Számítógépes Nyelvészeti Konferencia*, 109-21. JATEPress: Szeged, 2015.

Orosz György – Szántó Zsolt – Berkecz Péter – Szabó Gábor – Farkas Richárd. „HuSpaCy: an industrial-strength Hungarian natural language processing toolkit” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.) *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. 59-73. Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet, 2020.

Papp Imre - Dezső Márta - Fűrész Klára - Sári János - Somody Bernadette - Szegvári Péter - Takács Imre - Kukorelli István (szerk.) *Alkotmánytan I.*, Osiris Kiadó: Budapest, 2007.

Piehl, Aino – Sommardahl, Eivor. „Working towards Clear Administrative Language in Finland – Bilingually.” In Pirkko, Nuolijärvi – Stickel, Gerhard (eds.) *Language Use in Public Administration — Theory and Practice in the European States*. 73-82. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences, 2015.

Pak, Myeongsuk – Kim, Sanghoon. „A review of deep learning in image recognition” *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, Kuta Bali, Indonesia (2017) 1-3.

Peters, Matthew E. – Neumann, Mark – Iyyer, Mohit – Gardner, Matt – Clark, Christopher – Lee, Kenton – Zettlemoyer, Luke. „Deep Contextualized Word Representations” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227-37. New Orleans, Louisiana: ACL, 2018.

Petrétei József – Tilk Péter. *Magyarország alkotmányjogának alapjai*. Pécs: Kodifikátor Alapítvány, 2014.

Pinker, Steven. *A nyelvi ösztön - Hogyan hozza létre az elme a nyelvet?* Budapest: Typotex Kiadó, 2013.

Pléh Csaba. *A mondatmegértés a magyar nyelvben*. Budapest: Osiris, 1998.

Pléh Csaba. *A lélek és a nyelv*. Budapest: Akadémiai Kiadó, 2013.

Pléh Csaba – Lukács Ágnes. *Pszicholingvisztika – magyar pszicholingvisztikai kézikönyv 1*. Budapest: Akadémiai Kiadó, 2014.

Pottmann, Daniel. „Leichte Sprache and Einfache Sprache – German plain language and teaching DaF German as a foreign language.” *Studia Linguistica*. 38. (2020): 81-94.

Radford, Alec – Narasimhan, Karthik. „Improving Language Understanding by Generative Pre-Training.” 2018.

Radünzel, Claudia. „Leichte Sprache.” *Zeitschrift für Slawistik*. 62. (2017): 48-95.

Qiang, Jipeng - Qian, Zhenyu - Li, Yun - Yuan, Yunhao - Wu, Xindong. "Short Text Topic Modeling Techniques, Applications, - Performance: A Survey," *IEEE Transactions on Knowledge - Data Engineering*, vol. 34, no. 3 (1 March 2022): 1427-45.

Ring Orsolya – Szabó Martina Katalin – Guba Csenge – Váradi Bendegúz – Üveges István. „Approaches to Sentiment Analysis of Hungarian Political News at Sentence Level with Dictionary-based Method and with Machine Learning” (bírálat alatt) (2022)

Ring Orsolya – Vincze Veronika – Guba Csenge – Üveges István: „HunEmPoli magyar nyelvű, részletesen annotált emóciókorpusz.” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.): *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023*, 187-201. Szeged: Szegedi Tudományegyetem, 2023.

Rizzo, Camino Rea – Pérez, M. José Marín. „Structure and design of the British Law Report Corpus (BLRC): a legal corpus of judicial decisions from the UK.” *Journal of English Studies*, 2012/10. (2012): 131-145.

Sajgál Mónika. „Alárendelés a magyar jogi nyelvben” In Szabó Miklós – Vinnai Edina (szerk.): *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.* 123-151. Miskolc: Bíbor Kiadó, 2018.

- Sammut, Claude – Webb, Geoffrey I. „TF-IDF” In Sammut, Claude – Webb, Geoffrey I. (eds.) *Encyclopedia of Machine Learning*. 986–87. Boston, MA: Springer, 2010.
- Sarangi, Srikant – Stefan Slembrouck: „Language, Bureaucracy, and Social Control.” London: Routledge, 2014.
- Sebők Miklós – Ring Orsolya – Máté Ákos. *Szövegbányászat és mesterséges intelligencia R-ben*. Budapest, Magyarország: Typotex Kiadó, 2021.
- Simon Eszter – Farkas Richárd – Halácsy Péter – Sass Bálint – Szarvas György – Varga Dániel: „A HunNER korpusz.” In Alexin Zoltán – Csendes Dóra (Szerk.): *IV. Magyar Számítógépes Nyelvészeti Konferencia, 373-376*. Szeged: Szegedi Tudományegyetem, 2006.
- Simon Eszter – Vadász Noémi. „Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus.” In Ekštejn, Kamil – Pártl František – Konopík Miloslav (eds) *Text, Speech, and Dialogue. TSD 2021. Lecture Notes in Computer Science*, vol 12848. Cham: Springer, 2021.
- Sipos Lajos (főszerk.). „Pannon enciklopédia: Magyar nyelv és irodalom.” Budapest: Dunakanyar, 2000.
- Sun, Nan – Lin, Guanjun – Qiu, Junyang – Rimba, Paul. „Near real-time twitter spam detection with machine learning techniques.” *International Journal of Computers and Applications* 0(0), (2020): 1–11.
- Szabó Dávid. „A francia argó.” In Kis Tamás (szerk.). *A szlengkutatás útjai és lehetőségei*. Debrecen: Kossuth Egyetemi Kiadó, 1997.
- Szabó Miklós. „A jog argumentatív természete.” *Jogesetek magyarázata* 1. (2010): 81-7.
- Szabó Miklós - Vinnai Edina. *A törvény szavai. Prudentia Iuris* 33. Miskolc: Bíbor Kiadó, 2018.
- Szilágyi H. István. „A jogtudat-kutatások elméleti kérdései.” *MTA Law Working Papers*, 2016/12., 2016.
- Szilágyi Péter. *Jogi alaptan*. Budapest: Osiris Kiadó, 2006.
- Tamás Dóra. „A IUSterm jogi és közigazgatási adatbázis háttér munkálatainak tapasztalatairól.” In Szoták Szilvia (szerk.) *Diszciplínák találkozása: nyelvi közvetítés a XXI.*

században, 260-81. Budapest: Országos Fordító és Fordításhitelesítő Iroda Zrt. (OFFI), 2019.

Tekfi, Chaffai: „Readability formulas: an overview” *Journal of Documentation*, Vol. 43 No. 3, (1987): 261-73.

Tibshirani, Robert – Walther, Guenther – Hastie, Trevor: „Estimating the Number of Clusters in a Data Set Via the Gap Statistic.” *Journal of the Royal Statistical Society Series B*. 63. (2001): 411-23.

Tikk, Domonkos (szerk.). *Szövegbányászat*. Budapest: Typotex Kiadó, 2007.

Tóth Judit. *Jogalkotástan*. Szeged: Szegedi Tudományegyetem ÁJTK, 2016.

Tóth Judit. „A közérthetőség úgy, ahogy a jogalkotó elképzei.” *Közjogi Szemle* 15:4 (2022): 20-7.

Tóth Judit. „A közérthető közigazgatási eljárás és az átláthatóság – recenzió.” *Magyar Jogi Nyelv* 2023:1 (2023): 26-28.

Trócsányi László – Schanda Balázs (szerk.). „Bevezetés az alkotmányjogba.” Budapest: HVG-Orac, 2014.

Üveges István – Vincze Veronika – Ring Orsolya – Guba Csenge. „Aspect-based emotion analysis of Hungarian parliamentary speeches” In Rehbein, Ines –Lapesa, Gabriella – Klamm, Christopher –Ponzetto, Simone (eds.). *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022) Potsdam, Germany*, 75-84. University of Mannheim – University of Stuttgart, 2022.

Üveges István. „A Plain Language Movement kulturális kontextusa: Társadalmi háttér, történeti irányok és eredmények az Egyesült Államokban.” *Magyar Jogi Nyelv* 2020:2 (2021): 16-25.

Üveges István – Csányi Gergely Márk – Ring Orsolya – Orosz Tamás. „Szövegaugmentálási módszerek összehasonlítása politikai szövegek szentimentanalízise során” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.). XVIII. Magyar Számítógépes Nyelvészeti Konferencia. 521-34. Szeged: Szegedi Tudományegyetem, 2022.

Vapnik, Vladimir. „Estimation of Dependences Based on Empirical Data.” New York: Springer, 2006.

Vaswani, Ashish – Shazeer, Noam – Parmar, Niki – Uszkoreit, Jakob – Jones, Llion – Gomez, Aidan N. – Kaiser, Lukasz – Polosukhin, Illia. „Attention is all you need” In I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett (eds.) *Advances in Neural Information Processing Systems (30)*, Curran Associates, Inc., 2017.

Varga Csaba: *Ismereteink korlátozottsága és nyelvünk határozatlansága*. In Varga Csaba (szerk.). *Előadások a jogi gondolkodás paradigmáiról*. Budapest: Osiris Kiadó, 1999.

Varga Zsófia. „Terminológiai problémák a jogi szakfordításban: Korlátolt felelősségű társaság-e a limited liability company?” *Magyar Terminológia* 6: 1. (2013): 1-40.

Vincze Veronika – Lucza Mónika – Csendes Dóra – Kiss Gabriella. „Szótározási dilemmák a MetaMorpho magyar-angol fordítóprogram névszói adatbázisának építésében.” In Alexin Zoltán – Csendes Dóra (szerk.). *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006)*. 180-89. Szeged: Szegedi Tudományegyetem, 2006.

Vincze Veronika - Szauter Dóra - Almási Attila - Móra György - Alexin Zoltán - Csirik János. „Hungarian Dependency Treebank.” In Barbara Plank (eds.) *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10) Valletta, Malta*, European Language Resources Association (ELRA), 2010.

Vincze Veronika. „A Miskolc Jogi Korpusz nyelvi jellemzői.” In Szabó Miklós – Vinnai Edina (szerk.): *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.*, 9-36. Miskolc: Bíbor Kiadó, 2018.

Vincze Veronika – Kicsi András – Főző Eszer – Vidács László. „A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei.” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.). *XVII. Magyar Számítógépes Nyelvészeti Konferencia.*, 275-288. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2021.

Vinnai Edina. „A diskurzus kötött rendje: Kihallgatás és jegyzőkönyvezés a magyar büntetőeljárásban.” In Szabó Miklós (szerk.) *Nyelvében a jog: Nyelvhasználat a jogi eljárásban.*, 153-191. Miskolc: Bíbor Kiadó, 2010.

Vinnai Edina. „Nyelvhasználat a jogi eljárásban.” (phd értekezés) Elérés: 2021.08.17. <http://midra.uni-miskolc.hu/document/12225/4264.pdf>

Vinnai Edina. „A magyar jogi nyelv kutatása.” *Glossa Iuridica*. 2014: 1 (2014a): 29–48.

Vinnai Edina. „Az első „jog és nyelv” kutatás hazánkban.” *Alkalmazott Nyelvészeti Közlemények, Miskolc*, IX. évfolyam, 1. szám (2014b): 60-67.

Vinnai Edina. „Harc a szavakért – közérthetőség a jogban.” *Alkalmazott Nyelvészeti Közlemények, Miskolc*, XII. évfolyam, 1. szám (2017a): 42–53.

Vinnai Edina. „Jog és nyelv határán – A jogi nyelvhasználat nemzetközi és hazai kutatása.” Budapest: Gondolat Kiadó, 2017b.

Vinnai Edina. „»Megértette a figyelmeztetést?« A figyelmeztetések és tájékoztatások közlése a büntetőeljárásokban” In Szabó Miklós – Vinnai Edina (szerk.): *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.*, 281-295. Miskolc: Bíbor Kiadó, 2018.

Vinnai Edina. „Megjelent a közérthető fogalmazás nemzetközi szabványa” *Magyar Jogi Nyelv 2023: 1* (2023): 31-32.

Vogel, Friedemann – Hamann, Hanjo – Gauer, Isabelle: “Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies.” *Law & Social Inquiry* 4 (2018): 1340–1363.

Wang, Qi – Ma, Yue – Zhao, Kun – Tian, Yingjie: “A Comprehensive Survey of Loss Functions in Machine Learning” *Annuals of Data Science* 9 (2), (2020): 187-212.

Willerton, Russel: „Plain Language and Ethical Action - A Dialogic Approach to Technical Content in the 21st Century.” New York, NY: Routledge, 2015.

Williams, Joseph M. „Toward Clarity and Grace.” University of Chicago Press, 1995.

Wydick, Richard C. „Plain English For Lawyers.” Durham, NC: Carolina Academic Press, 1979.

Yang Zijian Győző – Perlaki Attila – Laki László János. „Automatikus összefoglaló generálás magyar nyelvre BERT modellel” In Berend Gábor – Gosztolya Gábor – Vincze Veronika (szerk.): *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, 343-53. Szeged: JATEPress, 2020.

Yasuhiko Igarashi – Kenji Nagata – Tatsu Kuwatani – Toshiaki Omori – Yoshinori Nakanishi-Ohno – Masato Okada. „Three levels of data-driven science.” *Journal of Physics: Conference Series* 699 (2016): 012001.

Zeng, Guoping. „On the Confusion Matrix in Credit Scoring and Its Analytical properties.” *Communications in Statistics - Theory and Methods* 49 (9) (2019): 2080–2093.

Zhu, Jinhua – Xia, Yingce – Wu, Lijun – He, Di – Qin, Tao – Zhou, Wengang – Li, Houqiang – Liu, Tie-Yan: „Incorporating BERT into Neural Machine Translation” Elérés: 2023.01.17. <https://arxiv.org/abs/2002.06823>

Zódi Zsolt. „A jog érthetőségének határai. Meg tudják-e oldani a nyelvészek a jogi szövegek érthetlenségének problémáját?” In Szabó Miklós – Vinnai Edina (szerk.): *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25.*, 241-260. Miskolc: Bíbor Kiadó, 2018.

Zsibrita János – Vincze Veronika – Farkas Richárd: „magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian” In Mitkov, R., Angelova, G. and Bontcheva, K. (eds.) *Proceedings of RANLP 2013.*, 763-771. Hissar, Bulgaria: INCOMA Ltd., 2013.