Üveges István

Közérthetőség és automatizáció
- kísérletek a jog, természetesnyelv-feldolgozás és informatika határán

Doktori (Ph.D.) értekezés tézisei

Témavezető: Dr. Vincze Veronika

Szegedi Tudományegyetem

Nyelvtudományi Doktori Iskola

Szeged, 2024

1. Az értekezés célja

A jogi szövegek természetükből adódóan olyan magatartási normákat határoznak meg, amelyek betartása mindenkire nézve kötelező, és amelyek ismertét és megértését a jogalkotó minden állampolgár részéről feltételezi és elvárja. Ennek ellenére az ilyen szövegek megalkotása során legtöbbször csak a szaknyelvet ismerők szempontjai érvényesülnek, a laikusok nézőpontja nem. A problémára megoldást jelenthet, ha a közérthető fogalmazás automatizációs lehetőségeit vizsgáljuk.

Egy olyan szoftver, amely egyfajta helyesírás ellenőrzőként segít a fogalmazónak könnyebben érthetővé tenni a szöveget, jelentős emberi munkaerő megtakarítását teszi lehetővé. Az értekezés a jogi és nyelvészeti szakirodalomra támaszkodva keresi a választ arra a kérdésre, hogy mely nyelvi, nyelvhasználati jegyek megváltoztatása segítheti a jogi szövegek közérthetővé tételét. Ennek során a kutatás nagyban támaszkodik a Plain Language Movement által meghatározott alapelvekre és javaslatokra. Egyszersmind felügyelt-, és felügyelet nélküli gépi tanulási megoldásokkal vizsgálja a közérthetőséget, mint klasszifikációs problémát.

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 és 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 is.

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 különböző programnyelveken implementált elemzési láncokról. Saját gépi tanuló modellek építéséhez is rendelkezésre állnak korpuszok a legkülönfélébb kutatási kérdések esetén, legyen szó akár politikai szövegek szentiment- és érzelemelemzéséről, kérdés-válasz benchmark vagy például szövegek absztraktív összefoglalóit elkészítő modellek kiértékeléséhez használható korpuszokról.

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 olyan célok kitűzése, amelyek nem csak nyelvtudományi szempontból megalapozottak és jogtudományi szempontból elfogadhatóak, de a természetesnyelv-feldolgozás jelenleg rendelkezésre álló eszközkészletének képességeit figyelembe véve reális fejlesztési célként is megragadhatók.

A fentiek megoldásához elengedhetetlen volt egyrészről a közérthetőség kérdését nyelvészeti és jogi szempontból tárgyaló (magyar és nemzetközi) szakirodalom eredményeinek szintetizálása, másrészről 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 vált dominánssá, így ezek alkalmazásához megfelelő mennyiségű tanítóadatot is biztosítani kellett.

2. Felvetett 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.

Ahogyan Tóth Judit megfogalmazza: „a közérthetőség olyan tulajdonsága valamely mondatnak, tantételnek, értelmi vagy erkölcsi igazságnak, szabálynak, amelynél fogva azok jelentése, tartalma mindenki által könnyen felfogható". A témával a leginkább behatóan a Plain Language Movement nemzetközi hatású mozgalma foglalkozik. Az irányzat képviselői kritikai észrevételek mellett olyan konkrét javaslatokkal is szolgálnak, amelyek alkalmazása elősegítheti a laikusok szövegértését a jogi nyelvvel való találkozás során. Ha általánosan akarjuk megragadni, akkor az érthetőség alappilléreit a következők jelenthetik:

- 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 ismert a közérthető fogalmazásnak nemzetközileg elfogadott, bár pragmatikai központú definíciója is, a következők 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.”

Fontos azonban tekintetbe venni a jogi nyelv rétegzettségét, 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 előírásként is megfogalmazza, hogy a szabályozási tartalom a jogszabály címzettjei számára egyértelmű legyen. Ugyanitt említhetjük 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.” .

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.

A hazai Jog és nyelv kutatások, mint a magyarországi kontextusban leginkább témába vágó kutatási irányzat szakirodalmának rövid előzetes áttekintése 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 priorizálása 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 Közéthetőségi 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?

3. A kutatás eredményei

A kutatás kezdetén tehát négy kiinduló hipotézist fogalmaztam meg, amelyek nagyrészt egymásra épülve voltak hivatottak feltérképezni azokat a legfontosabb kérdéseket, amelyek a közérthető fogalmazás automatizálási lehetőségei kapcsán felmerülnek.

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 az i. hipotézis világosan igazolódott.

Az olyan általános érvényű meghatározásokon felül, mint például a közérthetőség általánosan elfogadott, nemzetközi definíciója, 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.

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. A 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. A fentiek alapján a ii. hipotézis is egyértelműen igazoltnak tekinthető.

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.

Az elvégzett gépi tanulási kísérletek 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 kutatás során finomhangolt huBERT modell, 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 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.

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.

Látható volt, hogy a közérthetőséget támogató szabályok jelentős része végül nem kerülhetett kód szinten megvalósításra. 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:

A I. kutatási kérdés szerint a két nézőpont (laikus és szakértő) 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.

A II. kutatási kérdésre erős befolyással bírt, 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 törvényszövegek is közérhető megfogalmazásban íródnak. Magyarországon ezzel szemben azt láttuk, hogy a hasonló törekvések szórványosak, céljuk általában valamilyen emberi jogi vagy uniós normának való megfelelés.

A III. kutatási kérdésre vonatkozóan 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.

A vizsgált szakirodalmi irányzatok megközelítési módjukban (pl. kognitív irányzatok vs. gyakorlati alapokon kiinduló kezdeményezések, mint a PLM) széttagoltnak bizonyultak (IV-kutatási kérdés). 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.

Az V. kutatási kérdést vizsgálva részletesen értékeltem 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.

A VI. kutatási kérdésre tekintettel 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.

A VII. kutatási kérdés kapcsán az egyes kutatási irányok által az érthetőséghez rendelt konkrét nyelvi jellegzetességeket feltárása után részletesen ismertettem azokat a szabályszerűségeket, amelyek implementálhatónak bizonyultak.

4. Konklúzió

A dolgozat célja annak vizsgálata volt, hogy egy olyan, elsőre általánosnak és talán nehezen megfoghatónak tűnő 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, csakúgy, mint a normavilágosság és a közérthetőség fogalmi párosának relációját körbejárva a téma 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 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. 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 varianciát mutat a kérdést szabályozó konkrét törtvényekig (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 egyrészről a jogtudatosság fontosságára az állampolgárok részéről, másrészről 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 a problémát gépi tanulási, és szabályalapú megoldások szempontjából is megvizsgáltam.

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.

Ezt követően a korábban leghatékonyabbnak ítélt gépi tanul modell kvalitatív értékelésén keresztül mutattam be a felmerülő problémákat a konkrét személyekhez kötődő közérthetőségi intuíció, valamint az általános közérthetőségi szempontú értékítélet rekonstruálása 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 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 mozgatását, mivel lokális szerverként, esetleg docker containerbe csomagolva is futtatható.

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.

5. Az értekezés témájában megjelent publikációk listája

Üveges, István. Szabályalapú és gépi tanulásra alapozott megoldások a közérthető fogalmazás elősegítése érdekében, MAGYAR JOGI NYELV 7 : 1 pp. 12-20. , 9 p. (2023)

Üveges, István, Comprehensibility and Automation: Plain Language in the Era of Digitalization, TALTECH JOURNAL OF EUROPEAN STUDIES 12 : 2 pp. 64-86. , 23 p. (2022)

Üveges, István, Közérthetőség mint osztályozási probléma (?) - gépi tanulási kísérlet kézzel címkézett korpuszon In: Berend, Gábor; Gosztolya, Gábor; Vincze, Veronika (szerk.) XVIII. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2022 Szeged, Magyarország : Szegedi Tudományegyetem, Informatikai Intézet (2022) 644 p. pp. 619-631. , 12 p.

Üveges, István, A közérthetőség fogalmának megjelenése az Európai Unió és tagállamai jogforrásaiban, MAGYAR JOGI NYELV 5 : 1 pp. 25-31. , 7 p. (2021)

Üveges, István, A Plain Language Movement kulturális kontextusa: Társadalmi háttér, történeti irányok és eredmények az Egyesült Államokban, MAGYAR JOGI NYELV 2020 : 2 pp. 16-25. , 10 p. (2021)

Üveges, István, Automatizálható a közérthető megfogalmazás? - Jog, számítógépes nyelvészet és pszicholingvisztika találkozása, MAGYAR JOGI NYELV 2020 : 1 pp. 1-8. , 8 p. (2020)

Üveges, István, Közérthetőség a jogi nyelvben: követelmény és/vagy kultúra? MAGYAR JOGI NYELV 2019/2. pp. 20-26. , 6 p. (2019)

István Üveges

Comprehensibility and automation
- experiments at the interface of law, natural language processing and informatics

Synopsis for the doctoral (PhD) dissertation

Supervisor: Dr. Veronika Vincze

University of Szeged

Doctoral School of Linguistics

Szeged, 2024

1. Aim of the thesis

Legal texts, by their very nature, lay down rules of conduct which must be observed by all and which the legislator presupposes and expects all citizens to know and understand. However, in most cases, such texts are drafted from the point of view of those who know the language and not from the layman's point of view. A solution to this problem could be found by looking into the possibilities of automating the drafting of plain language.

Software that acts as a kind of spell-checker, helping the author to make the text easier to understand, could save considerable human labor. The thesis draws on the legal and linguistic literature to answer the question of which linguistic features of language and language use can be changed to make legal texts more understandable. In doing so, the research draws heavily on the principles and proposals set out by the Plain Language Movement. At the same time, it investigates comprehensibility as a classification problem using both supervised and unsupervised machine learning solutions.

The scientific significance of the research lies in the fact that the paper attempts to find a solution to an abstract semantic-pragmatic issue such as the comprehensibility of official texts from the point of view of the average recipient, primarily by using the tools of natural language processing (computational linguistics). Such an undertaking will inevitably have to consider the current limitations of the available natural language parsers, the fact that many linguistic phenomena cannot be captured effectively by automatic tools even with the most modern technologies, and the general lack of data that characterizes the issue of intelligibility.

The latter refers to the fact that, while a wide range of analysis tools is now available for basic tasks such as word-genre tagging, word recognition or morphological analysis, whether language models or analysis chains implemented in different programming languages. There are also corpora available for building your own machine learning models for a wide variety of research questions, be it sentiment and sentiment analysis of political texts, question-answer

benchmarks, or corpora for evaluating models for abstract summaries of texts.

However, if we are looking for a corpus that contains a selection of texts from a public accessibility point of view, we cannot find one available in Hungarian or in any other language (known to the author).

In addition, algorithmic feasibility requires setting goals that are not only linguistically sound and jurisprudentially acceptable, but also realistic as development goals given the capabilities of the currently available natural language processing toolbox.

To solve the above problems, it was essential to synthesize the results of the (Hungarian and international) literature on the issue of intelligibility from a linguistic and legal point of view, on the one hand, and to implement an IT toolkit capable of adequately solving the task formulated, even under the present technological constraints, on the other. Given the dominance of machine learning and deep learning algorithms in the last decade, it was necessary to provide the necessary amount of teaching data for the latter.

2. Research questions and hypotheses

The main motivation for this thesis was to identify the basic principles and the possibilities of applying the requirements of accessibility to certain texts of the legal system and to develop practical applications that could support the creation of texts that take these aspects into account (i.e. texts that have undergone intralingual translation).

As Judit Tóth puts it, 'intelligibility is the property of a sentence, a doctrine, an intellectual or moral truth, a rule, whereby its meaning and content can be easily understood by all'. The subject is most thoroughly addressed by the Plain Language Movement, which has had an international impact. In addition to critical comments, the movement's representatives also make concrete suggestions which, if

applied, could help lay people to understand the text when they encounter legal language. To put it in general terms, the basic pillars of comprehensibility could be the following:

- the ease of comprehension of simpler text,
- easier clarity of a structured and ordered text,
- the reduction in processing effort for shorter, more concise text, and
- the appropriateness of using linguistic-stylistic devices that indicate emotional motivation.

However, there is also an internationally accepted, albeit pragmatically centered, definition of plain language, as follows:

"A text is called plain language if the target audience:

- can find what they need;
- understand what they have found; and
- can use the information to meet their own needs."

It is important, however, to consider the layering of legal language, which predicts that the comprehensibility efforts that the essay seeks to promote may not be uniformly applied across the entire spectrum of the administrative/legal domain.

Act CXXX of 2010 also stipulates that the regulatory content must be clear to the addressees of the legislation. The same can be said of the previous legislative act, which even specified the criterion of intelligible drafting as "...legislation shall be drafted in accordance with the rules of the Hungarian language, in a clear and comprehensible manner.".

The above has led to the following hypotheses:

i. Based on the national and/or international literature, there is not only a definition of intelligibility, but also a specific set of linguistic features that impair the ease of understanding a text.

ii. There is a layer of legal language in which the objective of intelligibility can be put into practice.

A brief preliminary review of the literature on domestic Law and Language research as the most relevant research stream in the Hungarian context led to the following research questions:

I. Professional languages operate within a tight system of expectations, and those working in the profession rightly demand accuracy and professional correctness from them. Clarity seems to work against this; it adapts the text to the (lay) point of view of the recipient. How can these two aspects be reconciled, while prioritizing the latter in the legal domain?
II. Consequently, what is the jurisprudential position on such an intralingual translation in the domestic and international context?
III. What are the layers of the legal domain in which the primacy of intelligible drafting can be legitimized and implemented?
IV. Which preferred and dispreferred linguistic features appear in the domestic and international literature on plain language?

Following a review of the available literature (mainly related to the results of the Plain Language Movement abroad), it has become clear that the information materials of individual agencies may be the medium which represents the broadest interface of the legislator (or the state apparatus) with the lay recipient.

Following contacts with linguistic experts working in the Public Access Task Force of the National Tax and Customs Administration, a database of texts that had been reformulated in the field to make them more accessible to the public was available. As the size of the database seemed sufficient for the application of machine learning techniques, the following hypotheses had to be tested:

iii. Given adequate training data, it is possible to automatically sort (classify) texts that are comprehensible and in need of restructuring using a machine learning model.

iv. It is possible to build software around such a model that can support the author of the text (as a kind of special spell-checker) with comprehensibility suggestions.

The problem was motivated by the need to answer the following questions:

V. Which of the (supervised) machine learning algorithms works reliably enough to support the work of experts in a meaningful way?

VI. Which design principles can be used to implement a public accessibility checking / suggestion software?

VII. If there are specific suggestions in the literature related to the level of comprehensibility for each language, which of them and how can they be algorithmized?

3. Results

At the beginning of the research, I formulated four initial hypotheses, which were largely built on each other and were intended to explore the most important questions that arise in connection with the automation of plain language.

Hypothesis (i) was in fact that not only definitions/attempts at definitions of comprehensible drafting can be found in the study of the relevant literature, but also specific linguistic/language use features that negatively affect the comprehensibility of the text. A review of relevant research trends since the 1920s clearly confirms hypothesis i.

In addition to general definitions, such as the generally accepted international definition of intelligibility, the relevant research also formulates specific systems of criteria. Several linguistic phenomena,

drafting choices and preferred and dispreferred language usage features have been identified which can be used to identify specific points in a text which make it difficult to understand.

Hypothesis ii was the cornerstone of practical applicability. Unless a text type within the legal domain could be found which (especially in the domestic context) could accommodate intelligibility suggestions, further investigation could not go beyond basic research. As identified, this type of text is a group of functional texts that are dedicated to the lay reader. There is also a specific institutional programme under the aegis of the National Tax and Customs Administration, which makes the texts indicated available on its website only after a 'screening' process for public accessibility. The construction of a specific corpus of such texts has also been made possible thanks to the cooperation with the linguistic experts working in the Office's Accessibility Programme. Hypothesis (ii) can therefore also be considered as clearly confirmed.

Hypothesis iii had several pitfalls. Indeed, the fact that it is possible to collect teaching data for a task does not necessarily mean that the task itself is one that can be implemented by machine learning. For example, there may be several problems with the available data if they are inconsistent, or the problem goes beyond the capabilities of the chosen model.

Based on the machine learning experiments carried out, it can be stated that it was possible to create a model capable of public understanding classification, as confirmed by the results of the family of models tested. However, the efficiency with which the constructed model can operate is critical for its practical applicability. In particular, for the huBERT model, huBERTPlain, which was fine-tuned during the research, it was observed that the model performance (0.73 macro mean F1) was significantly below the usual results in binary classification tasks (typically > 0.9 macro mean F1). However, the accuracy and coverage achieved is sufficiently good to support expert work, as confirmed by manual validation.

The performance, which is significantly below the usual performance in binary classification tasks, was therefore accompanied by the fact that the best model currently available may be able to support expert work efficiently. Because of this dichotomy, hypothesis iii can be considered partially confirmed. The cleaning of the training data, and possibly the use of more resource-intensive models that may emerge later, may change this.

Hypothesis iv mainly reflects on the interconnectivity of the rule system hypothesized in i and the machine learning model hypothesized in iii. At the same time, it concretizes the expectation that the revealed, generalizable rules can be implemented in a programmed form.

It could be seen that a significant number of the rules supporting understandability could not be implemented at code level in the end. This was partly due to the insufficient amount of publicly available databases (e.g. terminology databases) and partly due to the limitations of the capacity of today's NLP tools. However, even with these limitations, it proved feasible to build a system that uses the machine-learned model to pre-filter the resulting texts and then provides rule-based suggestions to make them more understandable in cases that the model finds problematic. Such rules were:

- filtering impersonal structures / nominalization (using word-genre analysis)
- monitoring the proportion of abstract-related words (lexicon-based)
- Marking (legal) abbreviations - suggestion for resolution (lexicon based)
- Search for legal references - outsourcing to footnotes (regex based)
- filtering archaisms, multiple negations and function verbs (lexicon based)
- flagging of excessively long sentences and excessively many tag phrases (using a language model with empirical thresholds)

- content summarization at the beginning of the text: extractive extraction (using an unsupervised machine learning model).

It is also important to note that the machine learning model chosen as the best (as well as all ML models) has limited uses. The main reason for this is the domain dependence that follows from the nature of the learning data. Given that they could only include functional texts from the NAV's public accessibility programme (due to the unavailability of other data), the model can only be able to work effectively on identical or very similar texts. It cannot be expected, for example, to be able to classify the reasoning of court decisions with the same accuracy in terms of clarity.

Partly referring to the reason for the partial justification of hypothesis iii, and partly adding the limitations described here in relation to rule implementation and the machine-learned model, hypothesis iv can again be considered only partially justified.

Also, in the opening reflections of the thesis, a total of seven research questions emerged, which guided both the research process and the structure of the thesis. The first of these was as follows:

Research Question I was that it was important to determine which of the two perspectives (lay and expert) was considered to be primary for which text types, given the layering of the legal domain. The requirement of precision or even clarity of the norm (in the case of criminal law) evaluates texts from the point of view of the legal practitioner, while comprehensibility evaluates texts from the point of view of the average recipient. By defining the type of text (functional texts) that is specifically addressed to a lay audience (in legal terms), the apparent contradiction can be resolved.

Research question II was strongly influenced by the cultural context in which it was posed. The example of the United States illustrates that all communication channels used by US agencies to address citizens are subject to a mandatory public accessibility review. In the Scandinavian countries, it is not uncommon for specific legal texts to

be written in plain language. In Hungary, on the other hand, we have seen that such efforts are sporadic, and their aim is usually to comply with some human rights or EU norm.

For Research Question III, the appropriate text type is the group of functional texts mentioned above, which have a dedicated purpose to inform or guide a well-defined target audience (e.g. taxpayers) who are not familiar with the subject matter, in the implementation of an activity.

The literature trends examined were found to be fragmented in their approach (e.g. cognitive trends vs. practice-based initiatives such as PLM) (Research Question IV). Nevertheless, a number of common linguistic features were listed (e.g. multiple compound, long sentences, industry-specific abbreviations, etc.) which have a negative impact on comprehensibility in general and on comprehensibility in particular.

In addressing Research Question V, I evaluated in detail the performance of models belonging to the three characteristic types of machine learning algorithms. As formulated here, the BERT model, which uses context-dependent embeddings and is fine-tuned for the purpose, was found to be the most reliable.

The API nature of the application developed with respect to Research Question VI was justified by the need to ensure its wide usability, as it can be embedded behind any user interface (e.g. a web application). To take advantage of the transparency provided by the "handwritten" rules and the generalization capability inherent to machine-learned models, both approaches have been accommodated in the application. The fact that the machine-learned model is dependent on the output of the rules to be run ensures consistency between the two approaches.

In the context of Research Question VII, after exploring the specific linguistic features assigned to comprehensibility by each research direction, I have described in detail the rule features that have been shown to be implementable.

4. Conclusion

The aim of this thesis was to investigate whether a concept as general and perhaps elusive as "intelligible expression" could be captured in some way by specific linguistic features. In addition, it was inevitable to establish how these clues (together with other NLP tools) could help to create a programme that could effectively support expert drafting or post-editing tasks aimed at improving comprehensibility.

The starting point for this topic was the need for strong drafting, or the exploration of the role of clarity, coherence and consistency of drafting for the legislator. By exploring this question, as well as the relationship between the conceptual pair of norm clarity and intelligibility, the topic led to the interpretation that what is a norm-clarity text for the legislator is an intelligible text for the average person.

Given the fact that the various disciplines have tended to discuss the factors that influence the intelligibility of texts in isolation, and that the literature has not previously integrated the positions of the various schools of thought on the subject, it was appropriate to bring together all the major positions that have made some form of claim about the nature of intelligible communication. Experience has shown that these tendencies identify broadly similar lexical, syntactic and text organization elements as factors that complicate interpretation.

To assess how the efforts to put plain language at the heart of official communication can be implemented in the domestic context, it was useful to look at similar initiatives in other countries. In the Anglo-Saxon world, the role of plain language in the formal era is mainly ensured by the emphasis on the efficiency of communication and the 'service state' approach, while in the case of the Scandinavian countries and typically the continental legal system, the main emphasis is on its supportive character for the democratic capacity of citizens to assert their interests. The regulation of easily understandable official/legal documents varies considerably, from specific laws governing the issue (such as those in force in Sweden

and the United States) to sporadic initiatives such as those that have characterized the Hungarian situation (especially since the end of the 2010s). This highlights the importance of legal awareness on the part of citizens, on the one hand, and the inevitability of the state's role in the field, on the other, if the intelligibility of the law and of official communications in general is seen as a democratic right for the benefit of citizens.

Supporting the work of experts in this field can only be done effectively by making the widest possible use of the available NLP tools. For this reason, after reviewing the criticisms of the public accessibility proposals, I have examined the problem from the perspective of machine learning and rule-based solutions.

In the field of machine learning models, I tested both TF-IDF and models using context-independent and context-dependent embeddings, starting from classical machine learning algorithms. Among the latter, the huBERT model, fine-tuned by huBERTPlain, is the neural network-based model that was able to achieve the most promising results in the separation of the training data received from the National Tax and Customs Administration's Public Understanding Task Force, where the aim was to automatically separate sentences that needed reformulation from those that were already intelligible. After this separation, the sentences considered problematic were the input for the manual rule system, which aims to support the drafting work with automatic suggestions.

Then, through a qualitative evaluation of the machine-learning model, which was previously considered to be the most efficient, I presented the problems encountered in terms of reconstructing the intelligibility intuition associated with specific individuals, as well as the overall intelligibility-related value judgments.

When implementing the manual rules, it was important to ensure that they were supported by the different research directions (psycholinguistics, PLM, law and language, corpus linguistics). In addition, in several cases, practical limitations were imposed by the current limitations of the NLP tools currently available or chosen for

the development of the application. A good example of the latter is the syntactic analysis module of the transformer-based language model available for spaCy. This model was chosen at an early stage of the implementation phase because it represents one of the SOTA solutions available for the Hungarian language, but the consequences of its incomplete development were only faced at an advanced stage of the work. Nonetheless, the manual rule system produced may be suitable for use in conjunction with huBERTPlain to support the review of information material dealing specifically with tax issues from a comprehensibility perspective, and, if the model is replaced, it may become a general tool that can be used independently of the domain for similar purposes.

I implemented the application as an API independent of the specific user interface, which helps to avoid the forced movement of (possibly sensitive) data, as it can be run as a local server, possibly wrapped in a docker container.

This thesis discusses the nature and automatability of the concept of comprehensible formulation from a new perspective that has not been previously reported in the literature. In addition, the results of the dissertation raise a number of further questions about the relationship between automation, which is becoming increasingly widespread as a result of digitalization, and complex issues such as the possibilities of machine support for expert work in the legal domain, the uncertainty as to what language competence neural network-based language models are capable of representing, and the borderline between tasks that do not require human intuition and those that can be effectively approached using NLP tools.

Overall, the thesis attempted to provide a solution method and approach that has been missing from the Hungarian literature, moving within the interdisciplinary framework of jurisprudence, linguistics and informatics. The main tool for this was the integration of linguistics and jurisprudence and the introduction of an automation and IT perspective. The results may be useful for public actors interested in improving efficiency and promoting the requirements of

the rule of law, as well as for theorists working on the issue of communicating in a way that is accessible to the public. In particular, the presentation of a yet undeveloped approach to the topic and an assessment of the known limitations of the method could be helpful.

5. List of publication on the subject of the thesis

Üveges, István. Szabályalapú és gépi tanulásra alapozott megoldások a közérthető fogalmazás elősegítése érdekében, MAGYAR JOGI NYELV 7 : 1 pp. 12-20. , 9 p. (2023)

Üveges, István, Comprehensibility and Automation: Plain Language in the Era of Digitalization, TALTECH JOURNAL OF EUROPEAN STUDIES 12 : 2 pp. 64-86. , 23 p. (2022)

Üveges, István, Közérthetőség mint osztályozási probléma (?) - gépi tanulási kísérlet kézzel címkézett korpuszon In: Berend, Gábor; Gosztolya, Gábor; Vincze, Veronika (szerk.) XVIII. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2022 Szeged, Magyarország : Szegedi Tudományegyetem, Informatikai Intézet (2022) 644 p. pp. 619-631. , 12 p.

Üveges, István, A közérthetőség fogalmának megjelenése az Európai Unió és tagállamai jogforrásaiban, MAGYAR JOGI NYELV 5 : 1 pp. 25-31. , 7 p. (2021)

Üveges, István, A Plain Language Movement kulturális kontextusa: Társadalmi háttér, történeti irányok és eredmények az Egyesült Államokban, MAGYAR JOGI NYELV 2020 : 2 pp. 16-25. , 10 p. (2021)

Üveges, István, Automatizálható a közérthető megfogalmazás? - Jog, számítógépes nyelvészet és pszicholingvisztika találkozása, MAGYAR JOGI NYELV 2020 : 1 pp. 1-8. , 8 p. (2020)

Üveges, István, Közérthetőség a jogi nyelvben: követelmény és/vagy kultúra? MAGYAR JOGI NYELV 2019/2. pp. 20-26. , 6 p. (2019)