

A toxikusság felismerésének feltétele: interdiszciplináris közeg és a fogalom bevezetése

Hatvani Péter

PPKE Bölcsészet- és Társadalomtudományi Kar

Nyelvészeti Doktori Iskola

2026. június 26.



Az előadás menete

I. A fogalom szemantikai eredete

II. Az interdiszciplináris tér

III. Nyelvészeti alapok

IV. Elméleti kiterjesztések

V. A magyar kontextus

A kiindulópontok

„A »toxikusság« kifejezést a kutatók mindenütt alkalmazzák, de szinte soha nem definiálják.”

— Recuero (2024)

„Fiatal fogalom, amelynek figuratív szemantikája még alakulóban van.”

— Redkozubova (2021)

A feladat kettős:

1. **Nyelvészeti:** milyen nyelvi mechanizmusok teszik „toxikussá” a közlést?
2. **Operacionális:** a fogalom legyen annyira egzakt, hogy annotátorok jó megbízhatósággal tudják alkalmazni

Miért nem elegendő egyetlen tudományterület?

- A **pragmatika** magyarázza az egyedi aktust, de az algoritmikus terjedést nem tudja mérni
- A **számítógépes nyelvészet** detektálja a felszíni formát, de pragmatikai elmélet nélkül nem tudja értelmezni, *miért* sért egy közlés

- A **jogfilozófia** normatív keretet ad, de nem operacionalizálhatót
- A **kommunikációpszichológia** a személyiségvonást írja le, de a szekvenciális diskurzuskontextust nem
- A **magyar digitális nyelvészet** adja azt az anyanyelvspecifikus morfológiai és lexikai tudást, amelyet az angol centrikus modellek nem tartalmaznak

A „toxikusság” fogalma pontosan a diszciplínák *metzéspontján* értelmezhető és operacionalizálható.

A „toxikus” szó szemantikai eredete

- Eredet: kémia / orvostudomány (*mérgező anyag*)
- Figuratív terjedés: 1970–80-as évek; tömeges diffúzió a 90-es években (Redkozubova 2021, COHA)
- **Oxford Dictionaries Év Szava, 2018** — 45%-os keresési növekedés

A kémiai metafora produktív:

toxikus személy · toxikus kapcsolat · toxikus férfiasság · toxikus pozitivitás · toxikus komment

A metafora elvégzi a fogalmi munkát:

Nem a *cselekvőre* (sértő, fenyegető), hanem a *közegre* irányítja a figyelmet, a diskurzuskörnyezetre, amely *megfertőződik*. Ez élesen eltér a *sértés* (puntuális, telikus) vagy a *gyűlöletbeszéd* (atelikus, célirányos) fogalmától; a „toxikusság” *aktionsart*-ja duratív, akkumulatív, környezeti.

Az interdiszciplináris tér: szöveg és aktus

- **Számítógépes nyelvészet / NLP**: hogyan detektálható automatikusan?
(*Jigsaw / Perspective API; Wulczyn et al. 2017*)
- **Pragmatika / udvariatlanságkutatás**: milyen illokúciós aktus?
(*Culpeper 2011; Austin 1962; Searle 1969*)
- **Kommunikációpszichológia**: milyen viselkedéstípus?
(*Infante & Wigley 1986; Guberman et al. 2016*)

Az interdiszciplináris tér: norma, nyilvánosság, magyar kontextus

- **Politikatudomány / incivilitáskutatás**: mit tesz a politikai diskurzussal?
(Rossini 2022; Anderson et al. 2018; Iyengar et al. 2019)
- **Jogfilozófia**: mikor avatkozzon be az állam?
(Waldron 2012; Butler 1997; Calvert 1997)
- **Szociológia / platformkutatás**: hogyan terjednek az algoritmusokon?
(Recuero 2024; Bourdieu 1991; Foucault 1970)
- **Magyar digitális nyelvészet és pragmatika**: mi jellemzi a *magyar* online közeget, és milyen eszközök állnak rendelkezésre?
(Terestyéni 2006; Balázs G. 2023; Domonkosi 2002; Nemeskey 2021; Yang et al. 2023; Hatvani & Yang 2025)

Fogalmi szomszédok — a hierarchia

Verbális agresszió (*Infante & Wigley 1986*)

az önkép elleni támadás; 10 típus: sértés, fenyegetés, gúny, átok...

↓ pragmatikai szintre emelve

Udvariatlanság / impoliteness (*Culpeper 2011*)

normasértés + offenzív perlokúciós hatás; szándék nem feltétel

↓ ha csoport-hovatartozás a célpont

Gyűlöletbeszéd (*Waldron, Butler*)

csoportrágalmazás; méltóság-sértés

↓ ha a nyilvános diskurzus szintjén

Incivilitás → intolerancia-diskurzus (*Rossini 2022*)

delegitimáció

↓ platformokon, algoritmikusan erősítve

Diszkurzív toxicitás (*Recuero 2024*)

terjedés + legitimáció + diskurzusszennyezés

Culpeper: a toxicitáskutatás nyelvészeti csapásiránya

Az udvariatlanság három feltétele (Culpeper 2011):

1. A közlés sérti a hallgató *szociális norma alapú elvárásait*
2. *Offenzív perlokúciós hatást* vált ki → negatív érzelmi következmény
3. A szándék nem feltétel → a *szándék észlelése* elégséges

Arcmegtámadás, nem csupán arcveszélyeztetés: Spencer-Oatey (2000) rapport-kerete:

Rapportdimenzió	Példa	Toxicitásbeli megfelelő
Minőségi arc (<i>quality face</i>)	„Te idióta"	sértés, kompetencia-támadás
Szociális identitásarc	csoportslur	csoport-alapú identitástámadás
Méltányossági jogok	„Megöllek"	fenyegetés, megfélemlítés
Társulási jogok	kiközösítés	elnémítás, kizárás

Kontextusfüggetlen vs. kontextusfüggő udvariatlanság: a slur és az explicit fenyegetés az első; a szarkazmus és a rituális ugratás a második. Ez magyarázza a gépi osztályozók szelektív kudarcát.

Az operacionális meghatározás: Jigsaw / Perspective API

„A toxic comment is a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.”

— Wulczyn et al. 2017 / Jigsaw LLC 2020

A definíció három vonása:

1. **Perlokúciós, nem lokúciós** — a hatást definiálja, nem a tartalmat
2. **Viselkedési proxy** — „elhagyja a vitát” = elhallgattatás / visszavonulás
3. **Alattribútumok** searle-i illokúciós kategóriáknak felelnek meg: *sértés* (expresszívum), *fenyegetés* (komisszívum), *csoport-identitástámadás* (aszertívum + expresszívum)

Korlátok:

- Kontextusfüggetlen: ~5%-ban változik a humán ítélet a szülő-komment ismeretében (Pavlopoulos et al. 2020)
- Szarkazmus-vak és identitáselőítéletes (kisebbségi nyelvváltozatok → téves pozitív osztályozás)
- **Temporal drift**: szemantikai elmozdulás idővel, német osztályozó: $F1 = 0.62 \rightarrow 0.21$ három év alatt (Hoffmann et al. 2025); a kontextusfüggő udvariatlanság lexikai felszíne gyorsabban változik, mint a modellek frissítése

Mi a toxikusság ontológiailag? — Három jelölt

1. Toxikusság mint agresszió (*Infante–Wigley*)

„az önkép megtámadása"

- PRO: érett pszichológiai skálák · Gyengeség: a halk szimbolikus erőszakot és a toxikus pozitivitást nem ragadja meg

2. Toxikusság mint antiszociális viselkedés (*Hardaker 2010; Jigsaw*)

„ami megzavarja a kooperatív diskurzust"

- PRO: perlokúciós, mérhető a diskurzus folytatásában · Gyengeség: egyetlen üzenetre nem alkalmazható szekvenciális kontextus nélkül

3. Toxikusság mint normasértés (*Watts; Culpeper*)

„ami sérti a közösség *politic behaviour*-normáit"

- PRO1: igazodik a laikus metanyelvi ítéletekhez · Gyengeség: a normák rejtettek, vitatottak és gyorsan változnak

A disszertáció megközelítése: háromtesztes keret

Egy M_n közlés *toxikus*, amennyiben a három teszt közül egy vagy több pozitív:

1. **Agresszióteszt:** M_n konvencionalizált udvariatlan formulát valósít meg (Culpeper 2011): sértés, fenyegetés, elnémítás *VAGY* Infante–Wigley verbális-agressziókatégoriát
2. **Antiszociabilitás-teszt:** M_n perlokúciós hatása: elhallgattatás, kilépés, megtorlás, eskaláció (a szálban retrospektív annotációval meghatározva)
3. **Normasértés-teszt:** M_n sérti az adott közösség elvárásait, ahogy kompetens anyanyelvi értékelők megítélik *szálkontextussal*

Konjunktív az erős esetekben: slur + fenyegetés + normaszegés: mind három tüzel

Diszjunktív a perifériákon: toxikus pozitivitás: 2+3 tüzel, 1 nem → *toxikus*; rituális ugratás barátok közt: 1 tüzel, 2+3 nem → *nem toxikus*

Butler: a sérülékeny alany és az ismételhetőség

„Amikor azt állítjuk, hogy a nyelv megsebzett minket, milyen igényt fogalmazunk meg?”

— Butler (1997), *Excitable Speech*

- **Nyelv útján jövünk létre** — ezért vagyunk kiszolgáltatottak a minket megnevező, osztályozó, lealacsonyító szavaknak
- **A forma és tartalom szétválasztása illuzórikus** — a rasszista sértés nem azonos propozícióval a semleges megfogalmazáshoz képest; a *hogyan* a *mit* része
- **Performativitás, nem propozicionalitás** — a sértés *valamit tesz* a hallgatóval, de hatása nem a szándékon múlik; az illokúciós erő a utalási lánc *előtörténetéből* ered
- **Ismételhetőség** (*iterability*, Derrida nyomán) — a sértő szó erejét korábbi sérelmek idézése adja

Következmény: a platformon terjedő sértő tartalom a utalási(citacionális) láncot *mechanikusan és nagy sebességgel* iterálja — a toxikusság ereje algoritmikusan felerősített ismételhetőség.

Recuero: platformizáció és diszkurzív toxicitás

Recuero (2024) — a toxikusság nem *amit* a diskurzus mond, hanem *amit tesz platformizált közegben*:

Dimenzió	Mechanizmus	Platform-korreláció
Terjedés (<i>spread</i>)	Algoritmikus amplifikáció; virális megosztás	Repost, idézett megosztás
Legitimáció	Láthatóság = tekintélyjelzés	Kedvelések, trendek, verifikált fiókok
Ártalom (<i>harm</i>)	Diskurzusszennyezés; elhallgattatás	Kivonulás, koordinált zaklatás

A fogalmi újítás: egy diskurzus lehet toxikus *anélkül*, hogy egyedileg sértő lenne — ha rendszeresen mérgezi a diskurzív környezetet (pl. „ártatlan kérdések", mémek, toxikus pozitivitás).

Foucault *dispositif*-fogalmát alkalmazva: a platformok maguk *termelik* a „toxikus" kategóriát, amelyet aztán detektálni kívánnak — a toxicitásosztályozó soha nem semleges megfigyelő.

A magyar kontextus: Terestyéni és a konfrontatív kommunikáció

Terestyéni Tamás (2006), *Kommunikációelmélet*:

- A kommunikáció alapvető hivatása: **kooperáció**: a konfrontáció a kommunikáció *elfajulása*
- A 2000-es évek magyar nyilvános kommunikációja: **konfrontatív kommunikációkultúra**: a kooperatív és konfrontatív módok egyensúlyának tartós felborulása; *torz kommunikációk*
- **Mozaiknyilvánosság**: fragmentált figyelem, eltűnő közös tartalmak; a habermasi nyilvánosság szerkezeti átalakulásának magyar megfelelője

A habermasi örökség: kommunikatív cselekvés (érvényességi igények: igazság, helyesség, őszinteség) vs. stratégiai cselekvés (sikerre orientált). A toxikusság e keretben, egy stratégiai cselekvés, amely a nyilvános szféra kommunikatív affordanciáit fegyverként használja.

A magyar kontextus: Balázs Géza és a netnyelv

Balázs Géza (2023), *Az online tér hatása az írott és beszélt magyar nyelvre*:

A netnemzedék kommunikatív tünetei:

- *Mozaikfogalmazás* — vágott-másolt vendégszövegek; koherenciatörések
- Érzelmi érvek uralma a logikai érvek felett; *csak úgy*-fecsegés
- Konfliktuskezelés gyengülése: „nem megbeszélés, hanem elkendőzés"

Netnyelv-specifikus toxicitásvektorok:

- *trollkodik, hájpol, diszlájkol*: magyarosított angol tőjű agresszió
- *bazzege, wazze* (< *bazd meg*): eufemizált tabunyelv, szólistás szűrőkön átmegy
- Kicsinyítő képzők mint implikacionális udvariatlanság: *-csi/-i* sértő/gúnyos alkalmazása, becézés, csúfolás
- *grammar náci* — csúfolódó álsértés (*mock-insult*) mint szubkultúra-marker

A szekvenciális dimenzió: az előrejelzés lehetősége

A társalgáselemzés (Sacks, Schegloff) alaptétele: **a diskurzus szekvenciálisan szervezett**: minden fordulót alakít és alakít az előző. A kérdés: *mérhető-e ez a szervezettség toxicitás-előrejelzésre?*

Zhang et al. (2018) — *Conversations Gone Awry*: az első két megszólalás már tartalmaz jelzéseket arról, hogy egy Wikipédia-vita személyes támadásba torkollik-e. Prediktív jegyek: *te*-névmás sűrűsége · mondatkezdő kötőszók · tagadások · udvariassági markerek hiánya.

Chang & Danescu-Niculescu-Mizil (2019, EMNLP) — CRAFT: szólszintű szekvenciális számítógépes modell, amely fordulóról fordulóra frissíti annak valószínűségét, hogy a *következő* üzenet személyes támadás lesz.

A disszertáció kérdése:

$M_1 \dots M_n$ alapján megjósolható-e M_{n+1} toxicitása magyar nyelvű online szálakban?

Magyar specificitás: T/V-váltások, diminutív szarkazmus, netnyelvi koherenciatörések — olyan morfológiai és pragmatikai jegyek, amelyeket angol modell nem tud kinyerni.

Annotációs következmények

A háromtesztes keretből következő annotációs séma:

Tengely	Annotátor kérdése	Rögzített adat
Agresszió	Valósít-e meg M_n konvencionalizált udvariatlan formulát vagy Infante-kategóriát?	bináris + bizalmi érték
Antiszociabilitás	Vezet-e M_n elhallgattatáshoz, kilépéshez, eskalációhoz?	bináris + bizalmi érték
Normasértés	Sérti-e M_n a közösség elvárásait (száلكontextussal)?	bináris + bizalmi érték

Módszertani előny az előző munkához képest:

Hatvani & Yang (2025): Jigsaw-import séma, elméleti döntési kritériumok nélkül $\rightarrow \kappa_{\text{free}} = 0.525$

Jelenlegi séma: három elméleti döntési kritérium + száلكontextus kötelező biztosítása

Cél: Krippendorff $\alpha \geq 0.65$ tengelyenként

A disszertáció pozíciója

Közvetlen előzmény: Hatvani & Yang (2025) — magyar toxikus kommentek üzenetszintű osztályozása (huBERT-SetFit, F1 = 0.937)

A PhD három bővítési iránya:

1. **Üzenetszintről szálszintre** → *következő* üzenet toxikusságának előrejelzése; CRAFT-architektúra magyar adaptációja
2. **Jigsaw-importból pragmatikailag megalapozottra** → háromtesztes annotációs séma; elméleti döntési kritériumok
3. **Kis egyensúlyozott korpuszból nagy naturalisztikusra** → magyar *beszélt+írott* korpusz; *Conversations Gone Awry*-mintavételi logika; 1-2M szál

Összefoglalás

Kérdés	Válasz
Miért szükséges interdiszciplináris keret?	Mert a fogalom egyetlen területen sem definiálható és operacionalizálható egyszerre
Mi a „toxikusság” szemantikailag?	Kémiai metafora: duratív, akkumulatív, környezeti — a <i>közegre</i> fókuszál, nem a cselekvőre
Hogyan helyezkedik el a szomszédfogalmak között?	Verbális agresszió → udvariatlanság → gyűlöletbeszéd → incivilitás → <i>diszkurzív toxicitás</i>
Mi a disszertáció operacionális definíciója?	Háromtesztes keret: agresszió + antiszociabilitás + normasértés
Mi a szekvenciális tézis?	M_{n+1} toxicitása előrejelezhető $M_1 \dots M_n$ alapján, magyar-specifikus morfológiai és pragmatikai jegyekkel

Hivatkozások (válogatott) — I.

- Anderson, A.A. et al. (2018). Toxic talk. *Int. J. Public Opinion Research* 30(1).
- Austin, J.L. (1962). *How to Do Things with Words*. Harvard UP.
- Balázs G. (2023). *Az online tér hatása az írott és beszélt magyar nyelvre*. ELTE.
- Bourdieu, P. (1991). *Language and Symbolic Power*. Polity.
- Butler, J. (1997/2021). *Excitable Speech*. Routledge Classics.
- Calvert, C. (1997). Hate speech and its harms. *Journal of Communication* 47.
- Chang, J.P. & Danescu-Niculescu-Mizil, C. (2019). Trouble on the horizon. *EMNLP-IJCNLP*.
- Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence*. CUP.
- Domonkosi Á. (2002). *Megszólítások és beszédpartnerre utaló elemek*. DE.
- Foucault, M. (1970/1981). The order of discourse. In Young (ed.), *Untying the Text*. RKP.
- Guberman, J. et al. (2016). Quantifying toxicity and verbal violence on Twitter. *CSCW Companion*.

Hivatkozások (válogatott) — II.

- Hardaker, C. (2010). Trolling in asynchronous CMC. *Journal of Politeness Research* 6(2).
- Hatvani P. & Yang Z.Gy. (2025). Automated detection of toxic comments in Hungarian. *Ann. Math. Inf.* 61: 108–117.
- Hoffmann, M., Fillies, J., John, J., Maronikolakis, A., Navilarekal, A., Udupa, S. & Hammer, N. (2025). Temporal Robustness in Hate Speech Detection: Updating German Classifiers with Advanced AI Infrastructures. *ECAI 2025*, FAIA 413, pp. 1171–1178. DOI: 10.3233/FAIA250929.
- Infante, D.A. & Wigley, C.J. (1986). Verbal aggressiveness. *Communication Monographs* 53(1).
- Iyengar, S. et al. (2019). The origins and consequences of affective polarization. *Annual Review of Political Science* 22.

Hivatkozások (válogatott) — III.

- Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N. & Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? In *Proceedings of ACL 2020*, pp. 4149–4161.
- Recuero, R. (2024). The Platformization of Violence: Toward a Concept of Discursive Toxicity on Social Media. *Social Media + Society* 10(1). DOI: 10.1177/20563051231224264.
- Redkozubova, E.A. (2021). Toxic discourse, language ecology and slang. *HSS* 86(3).
- Rossini, P. (2022). Beyond incivility. *Communication Research* 49(3).
- Sacks, H., Schegloff, E.A. & Jefferson, G. (1974). A simplest systematics for turn-taking. *Language* 50(4).
- Searle, J.R. (1969). *Speech Acts*. Cambridge UP.
- Spencer-Oatey, H. (2000). *Culturally Speaking*. Continuum.
- Terestyéni T. (2006). *Kommunikációelmélet*. AKTI–Typotex.

Hivatkozások (válogatott) — IV.

- Waldron, J. (2012). *The Harm in Hate Speech*. Harvard UP.
- Watts, R.J. (2003). *Politeness*. Cambridge UP.
- Wulczyn, E. et al. (2017). Ex machina: Personal attacks seen at scale. *ICWWW*.
- Yang Z.Gy., Laki L.J., Váradi T. & Prószéky G. (2023). Mono- and multilingual GPT-3 models for Hungarian. In *Text, Speech, and Dialogue (TSD 2023)*, pp. 94–104. Springer.
- Zhang, J. et al. (2018). Conversations gone awry. *ACL 2018*.