

# Az automatikus és az emberi indexelés különbségei az NLM Medical Text Indexer (MTI) vizsgálatának segítségével: előzetes tanulmány

VIZVÁRI DÓRA

[vizvari.dora@outlook.com](mailto:vizvari.dora@outlook.com)

**Referált közlemény: Chen E, Bullard J, Giustini D. Automated indexing using NLM's Medical Text Indexer (MTI) compared to human indexing in Medline: a pilot study. J Med Libr Assoc. 2023 Jul 10;111(3):684–694. doi: [10.5195/jmla.2023.1588](https://doi.org/10.5195/jmla.2023.1588).**

2002-ben mutatták be a National Library of Medicine (NLM) automatizált indexelést támogató szoftverét, a Medical Text Indexer-t (MTI). Ezt a szoftvert kezdték el használni a Medline közlemények indexelésének támogatására, de eleinte csak részlegesen vezették be a tárgyszavazási folyamatban. 2022-re érett meg a technológia arra, hogy teljesen automatizált indexelés-támogatásra álljanak át. A szoftver legfrissebb verziója már az MTI-Auto (MTIA) nevet viseli. Erre az eszközre támaszkodnak ma a PubMed MeSH leképezési (mapping) és a „Hasonló közlemények” (Related Citation) funkciói is.

Bevezetése óta a kutatók folyamatosan tesztelik az MTI-t, de mindeddig kevés tanulmány született arról, hogy a tapasztalatokat összevetették volna a tisztán emberi kézzel történő indexelés eredményességével.

A közlemény szerzői előzetes (pilot) tanulmányukkal szeretnék lefektetni a későbbi vizsgálódások alapjait. Céljuk, hogy segítsenek az egészségügyi információs szakembereknek abban, hogy megérthessék az orvosi szakirodalom automatizált indexelésének hatásait. Igyekeznek felmérni, hogy hogyan befolyásolhatja az információ-visszakeresést a gépi indexelés az olyan eszközökben, mint a PubMed (Medline).

## Módszer

A folyóiratok, amelyekben ezek a közlemények megjelentek, mind megtalálhatók az Abridged Index Medicus (AIM) listában. Ezt a gyűjteményt ugyan az NLM 2020-ban lezárta, mégis ezt választották kiindulási pontnak, mert az AIM széles témaköri lefedettséget biztosít, tartalmazza a legalapvetőbb klinikai folyóiratok széles körét, és az ezekben a folyóiratokban



(kép: Pixabay+bmj)

publikált cikkekhez rendelkezésre állnak a teljes MeSH rekordok is.

A 2000-es évet választották, mert ez volt az utolsó olyan teljes lezárt év, amikor még nem került bevezetésre az MTI.

Igyekeztek biztosítani, hogy a vizsgálandó közlemények alacsony és magas minőséget is képviseljenek. Ezt a széleskörűséget a folyóiratok impakt faktora (IF) alapozták, az AIM lapjait rangsorolták a 2020-as IF-értékeik alapján, majd az elemzésbe bevont folyóiratok közé választották a tíz legmagasabb és a tíz legalacsonyabb IF értékű folyóiratot.

Azt, hogy az egyes negyedévekből melyik öt közlemény legyen elemezve, úgy döntötték el, hogy eleve kizárták a tárgyszavakkal és absztraktokkal nem rendelkező közleményeket, a rostán fennmaradt cikkeket pedig rangsorolták a PubMed relevanciarangsora (ranking) alapján, majd az első közlemény került kiválasztásra.

A gépi és emberi indexelés összehasonlításához a 2011-es megjelenésű MTI First Line Indexer-t (MTIFL)

használták, mert az elemzés készítői számára sajnos csak ez az MTI verzió volt szabadon elérhető. Ez az eszköz lehetővé teszi, hogy a felhasználók tízezer karakteres szöveg alapján MeSH tárgyszavakat kapjanak eredményül. A vizsgálandó húsz közlemény címét és absztraktjait másolták be, így készült el az az output, amely a gépi indexelő eszköz vizsgálatát lehetővé tette.

Többféle eredmény is kérhető a lekérdezés során, lehet egy szűkebb adathalmazt választani (Just the Facts (JTF)), és van lehetőség a teljes elemzést (Full Listing) is kinyerni. Mindkét listában a tárgyszavak rangsorolva vannak az alapján, hogy melyek a legfontosabbak a vizsgált közlemény témájának tekintetében.

Minden közleményről 14-féle adatot állítottak össze. A közlemények azonosítására szolgáló PMID mellett szerepelt a folyóirat témaköre, a folyóirat címe és a közlemény megjelenésének hónapja. A tárgyszavazás összehasonlításához megszámozták a gép és emberi indexelés során adott tárgyszavak számát, hány olyan szó volt, amelyet csak a gép, vagy csak az ember választott ki, hány olyan volt, amelyben egyezett a két indexelési módszer. A tárgyszavak száma mellett felsorolták magukat a tárgyszavakat is, megvizsgálták, hogy mely tárgyszavak voltak, amelyeket vagy csak a gép, vagy csak az ember választott ki és melyek voltak azok a tárgyszavak, amelyben megegyezett a két módszer.

Ez az adatmennyiség már elegendő információt adott egy kiinduló elemzés számára. A vizsgálat során elemezték az emberi és gépi indexelés során kiválasztott tárgyszavak különbségeit, vizsgálták, hogy ez az eredmény hogyan változik a magasabb, illetve az alacsonyabb impakt faktorú folyóiratok esetén. Igyekeztek megnézni, hogy hogyan tér el az emberi és a gépi indexelés eredménye, illetve milyen anomáliák és minták figyelhetők meg a két módszer között. Vizsgálták a nemi és életkori címkék használatát, a tárgy-szó-szinonimák kiválasztásának eltérését is.

## Eredmények

A közlemény szerzői részleteiben is elemzik az egyes közlemények tárgyszavazását, az emberi és gépi indexelés eltéréseit.

### Közlemények minősége alapján

Az eredmények azt mutatják, hogy az MTI eredményessége eltérést mutat a magasabb és alacsonyabb IF értékű folyóiratokban megjelent közlemények indexelése során.

Az emberi indexelési folyamathoz képest több tárgyszót választott ki a magasabb IF-ú folyóiratok

esetében. A IF érték alapján magasabban értékelt folyóiratok közleményeinek esetében több volt az egyezés az emberi és a gépi indexelés között.

### Címkézés

A vizsgálatba bevont közlemények közül 18 emberi csoporttal foglalkozott: korosztályra vonatkozó és nemi címkézésre is lehetőséget adtak a közlemények. A gépi indexelés ezen a téren mutatkozott a legkevésbé megbízhatónak. A nemre vonatkozó címkézések során a „férfi” címkéket általában előrébb rangsorolta, mint a „női” címkéket.

Többször is előfordult, hogy a gépi indexelés nem vette figyelembe az „idősek” (Aged) csoportokra vonatkozó címkézési lehetőséget.

### Szinonimák használata

A szinonimák tekintetében az emberi indexelő gyakrabban választotta a részletesebb tárgyszót, amíg a gép az általánosabb megközelítésű szinonimát rendelte a közlemény mellé. De ezt sem tette következetesen.

### Következtetések

A tanulmány szerzői többször is kihangsúlyozzák, hogy további vizsgálódások szükségesek. A húsz közlemény eredményének elemzése alapján az emberi és gépi indexelés különbségeivel kapcsolatban az alábbi megállapításokat foglalják össze:

1. A gép több tárgyszót rendelt és pontosabban választott tárgyszavakat a magasabb impakt faktorú lapokban megjelent közlemények esetében.
2. A tárgyszavak rangsorolása során a gép nem használta következetesen az időskorra és a nemi különbségekre vonatkozó címkéket.
3. Lehet, hogy a gépi indexelés során több esetben megfelelőbb szinonima került kiválasztásra, azonban nem volt következetes a tárgyszóválasztás.

Az NLM több változtatást és fejlesztést is kilátásba helyezett. Mivel a közleményben csak egy régebbi, 2011-es verziót volt lehetőség használni, így a legfrissebb verzió talán már orvosolt pár indexelési problémát.

A közlemény szerzői szerint egyelőre az MTI-re, mint automatizált indexelő eszközre való támaszkodás veszélyezteti a MeSH teaurusz használatának integritását, pontosságát és használhatóságát.

Az információkeresés során sokkal több munkát és szűrést vár el a felhasználótól egy releváns találati lis-

ta kinyerése. A könyvtárosok jelenleg több rutinnal rendelkeznek és megbízhatóbbak a tárgyszavazás terén.

Az MTI viszont mindenképpen befolyásolni fogja a publikálások minőségét. Az elemzés igazolta, hogy a gép nagyban támaszkodik a közlemény címére, ez a publikáló szerzőket pontosabb és konkrétabb címválasztásra fogja készíteni. Ha a cím metaforát, vagy kétértelmű kifejezéseket tartalmaz, az félre fogja vezetni a gépi indexelést.

A közlemény szerzői javasolják, hogy az NLM jobban vonja be a könyvtáros szakembereket az MTI fej-

lesztésébe, ugyanakkor kérik, a legfrissebb verzió tesztelésének lehetőségét is.

Elismerik, hogy cikkük kevés közlemény tesztelése alapján született, ugyanakkor azt is be kell látni, hogy az emberi indexelés is változott az elmúlt 20 év során, ma már máshogy tárgyszavaznak, mint tették azt 2000-ben.

Végül egy idézet a szerzőktől: „Célunk ebben a tanulmányban az volt, hogy példák sorával illusztráljuk a Medline automatizált indexelésével kapcsolatos lehetséges problémákat, és tegyük ezt meg akkor, mielőtt még az NLM befejezi ezt a mérőkövető számítógépes átállást.” ■

The screenshot shows the NIH National Library of Medicine website. The header includes the NIH logo and the text "National Library of Medicine" and "Lister Hill National Center for Biomedical Communications". A search bar is visible on the right. The main navigation menu includes "RESEARCH/R&D", "PUBLICATIONS", "DOWNLOADS", "PERSONNEL", and "ABOUT". The "RESEARCH/R&D" section is active, showing a list of projects: "LHCBC Project List", "Current Projects", "Health Information Standards and Discovery", "Natural Language Processing", "Image Processing", and "Legacy Projects". The "Natural Language Processing" project is selected, showing details for the "Indexing Initiative". The page title is "Research Area: Natural Language Processing" and "Researchers: Jim Mork". The project link is "https://lhncbc.nlm.nih.gov/ii". A description of the Indexing Initiative (II) project is provided, along with a diagram of a book spine. The text describes the project's goal of facilitating the retrieval of biomedical information from textual databases such as MEDLINE. It also mentions the development of the Medical Text Indexer (MTI) system, based on two fundamental indexing methodologies.

<https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/NLP/indexing-initiative.html>