DODÉ RÉKA – YANG ZIJIAN GYŐZŐ

HUN-REN Nyelvtudományi Kutatóközpont
dode.reka@nytud.hun-ren.hu
https://orcid.org/0009-0008-3880-2750

yang.zijian.gyozo@nytud.hun-ren.hu
https://orcid.org/0000-0001-9955-860X

Dodé Réka – Yang Zijian Győző: Generálás vs. kinyerés: nagy nyelvi modellel kapott kulcsszavak vizsgálata
Alkalmazott Nyelvtudomány, XXV. évfolyam, 2025/1. szám, 67–83.
doi:http://dx.doi.org/10.18460/ANY.2025.1.004

Generálás vs. kinyerés: nagy nyelvi modellel kapott kulcsszavak vizsgálata

Two types of keyword extraction methods can be distinguished: extractive and abstractive. Extractive keyword extraction relies on author-provided keywords, selecting the most appropriate ones for the given text. Abstractive keyword generation, on the other hand, is more similar to how a human would generate keywords. This method can generate keywords that may not have appeared in the author-provided list.

In our research, we employ the abstractive method to generate keywords by fine-tuning the PULI LlumiX 32K language model, which processes inputs of up to 32,000 tokens. We examine whether the keywords generated by the language model match the author-provided keywords and whether they appear in the text. If they differ from the author's keywords, we analyze the extent and nature of these differences—whether they are more general, more specific, synonyms, or other variations—evaluate their relevance to the text's topic, and determine if they accurately represent the text.

The purpose of keywords is to represent the text, provide insight into its content, and facilitate easier navigation among technical texts. However, the strategy behind author-provided keywords differs from a simple frequency-based approach.

The fine-tuning of the PULI LlumiX 32K language model, with a 32,000-token input, was based on author-provided keywords extracted from studies in the REAL repository. We briefly present the training data and process used for fine-tuning based on an already published study.

We further examined 133 manually reviewed results. The resulting keywords were classified into four groups: 1) entirely new (neither in the text nor among the keywords), 2) learned (among the keywords but not in the text), 3) extracted (not among the keywords but present in the text), and 4) accurate/matched (present both in the keywords and the text). In the first part of the research, we quantify these groups. We then take a closer look at the first group and the third group: how they relate to the author-provided keywords (more general, more specific, synonyms, or other differences), their relevance to the text's topic, and whether they accurately represent the text.

The analysis indicates that the language model effectively generates an appropriate number of keywords, with 44.2% matching those provided by authors. Additionally, 21.8% of the model-generated keywords do not appear directly in the text, suggesting creative inference. However, categorization challenges arise, particularly with synonyms and domain-specific terms, due to the model's limited conceptual understanding.

Compared to author-assigned keywords, the model's generated keywords differ in distribution, often leaning toward more general concepts. While some keywords align with indexing principles, others contradict specificity guidelines by being overly broad. Multi-word keywords, which are more thematically relevant, mirror patterns observed in human-generated keywords, making them more effective in representing text content.

Zero-shot prompting yields inconsistencies, resulting in generic keywords that weaken both search engine efficiency and text representation. Fine-tuning the model with additional data or modifying the prompting approach could enhance its performance, encouraging more precise and contextually relevant keyword generation.

This research contributes to the testing and refinement of keyword generation models, offering insights for improving future fine-tuning strategies. Adopting reinforcement learning based on human feedback could further optimize the model, fostering the generation of creative, multi-word, and highly specific keywords.

Keywords: PULI LlumiX 32K LLM, fine-tuning, zero-shot prompting, generated keyword, extractive keyword extraction

1. Bevezetés

Aki a tudományos életben tevékenykedik és publikál, jól ismeri a kulcsszavakat és azok célját. Számos területen bevett gyakorlat, hogy a szerzők kulcsszavakat adnak meg tudományos publikációikban, hogy azok segítsék az olvasókat. A kulcsszavak több, fontos funkciót töltenek be: betekintést nyújtanak a szöveg tartalmába, képviselik a fő témákat, megkönnyítik az akadémiai szövegek közötti navigációt, emellett növelik a láthatóságot a keresőmotorok által. Adódik azonban a kérdés, hogy milyen kulcsszavakat érdemes ehhez megadni?

Az UNESCO (1975) meghatározta azokat az alapelveket, amelyeket a kulcsszavaknak követniük kell, függetlenül attól, hogy manuálisan vagy automatikusan kerültek-e meghatározásra. A két fő alapelv a teljesség és a specifikusság. Firoozeh és munkatársai (2020) pedig további elvekkel egészítik ki ezeket: minimalitás, pártatlanság és reprezentativitás.

A teljesség azt jelenti, hogy a kulcsszavaknak le kell fedniük a dokumentum minden olyan témáját, amely potenciális információs értékkel bír, beleértve az implicit, a dokumentumban nem kifejezetten megjelenő témákat is. A specifikusság elve szerint olyan nyelvi elemeket kell kiválasztani, amelyek kifejezetten az adott dokumentumra jellemzők, és nem alkalmazhatók általánosan hasonló dokumentumokra (UNESCO, 1975). Firoozeh és társai (2020) szerint a minimalitás elve megköveteli, hogy a kulcsszavak egyedi jelentésűek legyenek – ez növeli a konzisztenciát és az érthetőséget is. A pártatlanság elve az objektivitást hangsúlyozza, míg a reprezentativitás elve szerint a kulcsszavaknak a dokumentum lényeges elemeit kell tükrözniük.

Dodé (2024) egyik tanulmányában a szerzői kulcsszavak szövegbeli előfordulását vizsgálta. Rámutatott, hogy az általa vizsgált korpuszban a kulcsszavak 37%-a kevesebb mint kétszer fordul elő a szövegekben. Ennek oka azonban nem csupán elírások vagy morfoszintaktikai különbségek, hanem az is, hogy a szerzők szándékosan általánosabb vagy specifikusabb kulcsszavakat választanak, figyelembe véve azt, hogy a szöveget a szakmai közönség olvassa majd.

A kulcsszóadás tehát egy átgondolt stratégia, amelyet automatikus kinyeréssel vagy épp generálással próbálunk reprodukálni, automatizálni.

A kulcsszókinyerés alapvetően egy szövegelemzési technika, amelyet a természetesnyelv-feldolgozás, az információkeresés és a szövegbányászat alkalmazásaiban használnak (Firoozeh et al., 2020). A kulcsszókinyerés hosszú múltra tekint vissza. Namoto (2023) az elmúlt 50 év kulcsszókinyerési módszereit foglalta össze tanulmányában, amelyben helyet kapnak az olyan módszerek, mint az automatikus terminusindexelés, a nyelvtaniszabály-alapú kinyerés, a gráfalapú rangsorolási modell, a külső információt alkalmazó modell, a szövegosztályozás, a különböző statisztikák, majd a generatív modellek. Nomoto (2023) végül javasolja a generatív rendszerekre való áttérést is, mivel ezek lehetőséget biztosítanak kulcsszavak belső és külső építésére. A generatív rendszerekkel pedig elérkeztünk a neurális nyelvi modellekhez és a generáláshoz.

Egy másik szempontból közelítve, kétféle módszert különböztethetünk meg: extraktív és absztraktív. Extraktív módszernek azt hívjuk, amikor a modellnek meglévő címkék közül kell választania. A fent említett megoldások közül például a terminusindexelés vagy a szövegosztályozás tartozik ebbe a kategóriába. A másik módszer az absztraktív, amikor a modell képes a megadott címkéken kívül teljesen új elemet létrehozni. Az utóbbi években a generatív nyelvi modellek lehetővé tették az absztraktív kulcsszógenerálást.

1.1. A nyelvmodellek

Az elmúlt évtizedben a neurális nyelvi modellek megjelenése alapvető változásokat hozott a nyelvtechnológia területén. Ez eleinte statikus szóbeágyazásokat jelentett, azonban mára számos transzformeralapú és generatív kontextuális nyelvmodell is rendelkezésünkre áll.

A nyelvi modellek a közelgő szavak előrejelzését adják a megelőző szókontextusból (Jurafsky–Martin, 2023). A korábbi nyelvi modellekkel szemben a neurális (nagy) nyelvi modellek „can handle much longer histories, can generalize better over contexts of similar words, and are more accurate at word-prediction” (Jurafsky–Martin, 2023: 147).[1] A *nagy* jelző a betanítóadatok hatalmas mennyiségére utal.

A mély neurális hálózatok elterjedésével lehetőség nyílt a vektoriális szemantika világából ismert szóbeágyazás (word embedding) (Mikolov et al., 2013) módszertanát alkalmazni. Egy beágyazás minden szóhoz egy sokdimenziós, folytonos vektort rendel. Ezek a vektorok egy szemantikus teret feszítenek ki, ahol a hasonló jelentésű szavak vektorai közel esnek egymáshoz. Ezek a modellek a szemantikai jegyek mellett szintaktikai jellemzőket is képesek megtanulni (Nemeskey, 2020).

A kezdeti statikus szóbeágyazáson alapuló nyelvmodelleket idővel felváltották a kontextuális transzformeralapú nyelvmodellek. A transzformer modell (Vaswani et al., 2017) az „önfigyelmi” (self-attention) mechanizmuson alapszik.

[1] „sokkal hosszabb előzményeket tudnak kezelni, jobban általánosíthatók a hasonló szavak kontextusában, és pontosabbak a szó előrejelzésében” (ford.: a szerzők)

A modell képes környezetfüggő szövegreprezentációt képezni, amelynek segítségével pontosabban tudja megtanulni a nyelvi jelenségeket. Az eredeti transzformer modellt egy enkóder és egy dekóder architektúrával tervezték meg, ahol az enkóder rész felel azért, hogy bemeneti szövegből egy vektor reprezentációt készítsen és a dekóder felel a szöveg generálásáért. A Transzformer modellen alapuló modellek megjelenése után szignifikánsan felülmúlta az addigi legjobb eredményeket (Barrault et al., 2020). A transzformer architektúráját tekintve, mind az enkóder, mind a dekóder külön-külön is használhatók és tanítható belőlük nyelvmodell. Tipikusan egy osztályozó feladatot, mint amilyen a kulcsszókinyerő feladat is lehet egy csak enkóder alapú nyelvmodellel, például egy BERT-modellel (Devlin et al., 2019) tanítani. Azonban az enkóderalapú modellekkel nem vagyunk képesek új tartalom generálására, ezért ha a feladat célja az, hogy új kulcsszavakat is tudjon generálni, akkor egy dekóderalapú modellt érdemes választani.

A csak dekóderalapú modellek családjába tartoznak az utóbbi évek legnépszerűbb modelljei; a nagy nyelvi modellek, mint a GPT-modellek (Brown et al., 2020), illetve az azokra épülő alkalmazások, mint a ChatGPT[2]. Az utóbbi hónapok kutatásai egyértelműen azt bizonyítják, hogy ha kellően nagy egy ilyen csak dekóderalapú nyelvi modell, és kellően sok adattal tanították, akkor képesek rendkívül széles körben megoldani nyelvtechnológiai feladatokat, köztük például a kulcsszókinyerést. Azonban az ilyen nyersen előtanított nagy nyelvi modellt az úgynevezett „következő szó prediktálása” feladattal tanították elő, ahol azt tanulja meg, hogy egy adott bemeneti szöveget hogyan kell folytatni. Így, ahhoz hogy kulcsszókinyerésre vagy kulcsszógenerálásra használhassuk, akkor azt tovább kell finomhangolni az adott feladatra.

Egy ilyen jellegű feladathoz, mint a kulcsszó generálása, többféle nyelvmodellt lehet választani. Mivel magyar nyelvre szeretnénk készíteni modellt, ezért olyan modellek jöhetnek szóba, amelyek magyarnyelv-tudással rendelkeznek. A nagy nyelvi modellek korszaka előtt több magyar nyelvű modell is létrejött, mint a huBERT (Nemeskey, 2021), vagy PULI BERT-Large vagy PULI GPT-2 (Yang et al., 2023a), azonban egyrészt ezek a modellek korlátozott bemeneti hosszt képesek kezelni, másrészt az új trendeket követve a nagy nyelvi modellekkel optimálisabb eredményeket lehet elérni.

Jelenleg négy nagy nyelvi modell érhető el magyar nyelvre. Az egyik a HILANCO GPTX[3] (Feldmann et al., 2021), egy angol-magyar kétnyelvű GPT-NeoX típusú modell (Andonian et al., 2023). Ez sajnos csak korlátozottan elérhető. Emellett rendelkezésre állnak a PULI-család[4] GPT-NeoX (Antonian et al., 2023) típusú nagy nyelvi modelljei, köztük a magyar nyelvű PULI 3SX (Yang et al., 2023a), valamint a háromnyelvű (magyar-angol-kínai) PULI Trio modell

[2] https://chat.openai.com
[3] https://hilanco.github.io/home.html
[4] https://puli.nytud.hu/

(Yang et al., 2023b). Végül a jelen kutatás során használt többnyelvű Llama-2 (Touvron et al., 2023) magyar nyelvre továbbtanított modellje, a PULI LlumiX 32K (Yang et al., 2024).

Az általános nyelvmodellek óriási mennyiségű szöveges adaton tanulnak, és az abból megszerzett tudásból építkeznek, így olyan problémákat is képesek megoldani, amelyeket a régebbi alkalmazások nem vagy nem olyan hatékonyan (pl. utasításkövetés).

1.2. Kapcsolódó irodalom

A kapcsolódó irodalmak között fontos megemlíteni ennek a kutatásnak az alapját képező, ezt a kutatást motiváló kutatást; Dodé (2024) a szerzői kulcsszavakat és azok szövegbeni előfordulását vizsgálta, majd összevetette őket a terminusok tulajdonságaival. Egy másik kutatás során Dodé (2023) különböző promptokkal kulcsszó- és terminuskinyerési kísérleteket végzett a ChatGPT-t segítségével. A prompt egy olyan utasítás, kérés vagy kérdés, amelyre válaszként kimenetet kapunk a finomhangolt modelltől (Jalalov–Gaszcz, 2023). A promptok fontosságát hangsúlyozza Reynolds és McDonell is a nyelvmodellek hatékony kiaknázása kapcsán (Reynolds–McDonell, 2021).

A promptkutatást követően magyar nyelven továbbtanított nyelvmodellel kísérletezett Dodé és Yang (2024a, Dodé–Yang, 2024b). Kutatásukban statikus szóbeágyazáson alapuló nyelvmodellt tanítottak be a fastText (Joulin et al., 2017) segítségével, valamint finomhangoltak transzformeralapú nagy nyelvmodelleket, mint a PULI 3SX és Llama-2. Egyértelműen alátámasztják azt, hogy a transzformeralapú nyelvmodell szignifikánsan jobb eredményt ér el, mint a statikus szóbeágyazáson alapuló modell, valamint nem elhanyagolható szempont az is, hogy a kutatásban használt cikkek rendkívül hosszú dokumentumok, így az is fontos szempont, hogy olyan Llama-2 modellvariánst alkalmaztak, ami megnövelt bemeneti hosszal rendelkezett.

Továbbá érdemes megemlíteni, hogy a kulcsszógenerálási feladat kétféleképpen is megoldható a nyelvmodellek segítségével. Az egyik megoldás lehet az úgynevezett extraktív megközelítés, amikor az adott szöveg alapján a meglévő címkék közül választ ki a modell releváns kulcsszavakat. A statikus szóbeágyazáson alapuló nyelvmodell csak extraktív módon képes kulcsszókinyerést végezni. A másik megoldás az absztraktív kulcsszógenerálás, amikor a modell képes az adott szöveg alapján az előzőleg megadott kulcsszavakhoz képest teljesen új kulcsszavakat generálni. Ehhez a feladathoz mindenképpen szükséges egy dekóderalapú architektúrájú nyelvmodellt használni. A korábbi kutatásunkban mindkét irányt kipróbáltuk, de egyrészt a nagy nyelvi modellek szignifikánsan jobb eredményt értek el, másrészt szeretnénk, ha a modell képes lenne kreatív, új kulcsszavakat generálni, ezért végül az absztraktív kulcsszógenerálást választottuk végső megoldásnak.

Yang és munkatársai (Yang et al., 2020) az extraktív módszert alkalmazva fejlesztett magyar nyelvű kulcsszó- és címkekinyerő rendszer, amelyet sajtó szövegekkel finomhangoltak. Kutatásukban szintén a fastText implementációt alkalmazták.

Az elmúlt évek során többen is foglalkoztak azzal, hogy a nagy nyelvi modelleket terminológiai feladatokra, például terminuskivonásra és kulcsszókinyésre finomhangolják (például Sun et al., 2020; Sammet–Krestel, 2023, Liu et al., 2021).

2. Módszertan

A kutatáshoz használt anyag egy már meglévő kutatásból származó eredmény további vizsgálata.

2.1. A tanítóanyag és a PULI LlumiX 32K nyelvi modell betanítása

Ez a fejezet Dodé–Yang (2024a) tanulmánya alapján készült, annak röviden bemutatott eredménye.

A PULI LlumiX 32K modellt a HUN REN Nyelvtudományi Kutatóközpont fejlesztette (Yang et al., 2024). A modell egy 32 ezres kontextus ablakú, magyar nyelvre továbbtanított Llama-2 modell. Az eredeti Llama-2 modellt a Together[5] továbbtanította, és bemeneti hosszát a finomhangolás során megnövelték 32.768 hosszúra (Llama-2-7B-32K). Ezzel lehetőség nyílt hosszú dokumentumok feldolgozására. Ezután Yang és munkatársai (2024) ezt a megnövelt bemenetű Llama-2-7B-32K modellt tanították tovább előtanítási módszerrel nyers magyar korpuszon. Ezzel gyakorlatilag magyar nyelvre adaptáltak egy Llama-2 modellt. A tovább előtanításhoz körülbelül 8 milliárd magyar szót használtak. A továbbtanítás tudásfelejtéssel járhat, ezért a 8 milliárd magyar szó mellé körül belül 2 milliárd szónyi angol szöveget is kevertek. Majd az így elkészült modellt finomhangolták kulcsszógenerálás feladatára.

A finomhangolás folyamata során egy előzetesen betanított modellt továbbképeznek egy adott feladatra (Jurafsky–Martin, 2023). Ilyen feladat például a koreferenciafeloldás. Az előtanítás során a modell hatalmas mennyiségű szöveg feldolgozásával megtanulja a szavak és mondatok jelentésének bizonyos reprezentációit (Jurafsky–Martin, 2023).

A kutatás a REAL-repozitórium (tudományos tanulmányok tára) egy részének felhasználásával készült. A PDF-anyagokat először txt formátumba konvertálták, végül 1146 szövegből nyerték ki a szerző által megadott kulcsszavakat, amivel a PULI LlumiX 32K modelljét finomhangolták. Az összehasonlítás kedvéért, az eredeti Llama-2-7B-32K modellt is finomhangolták és egyértelműen kimutatták, hogy a magyarra adaptált változat jobb eredményt ért el, vagyis a továbbtanítás magyar nyelvre egyértelműen értékes magyar tudást adott az eredeti modellhez.

[5] https://www.together.ai/

Így mi a továbbiakban már csak a PULI LlumiX 32K modellt fogjuk tanítani és elemezni.

A kutatásukhoz a korpuszt felbontották tanító- és tesztanyagra – 1000 cikket tanításra és 145 cikket tesztelésre. A tanításhoz a Stanford Alpaca promptsablont és az OpenChatKit implementációját használták (lásd az 1. ábrán).

1. ábra. Stanford Alpaca promptsablon és az OpenChatKit implementáció

Stanford Alpaca
Az alábbiakban egy utasítást találsz, amely leír egy feladatot, amelyhez egy bemenetet is mellékelünk, hogy további összefüggéseket adjon. Írj egy választ, amely megfelelően teljesíti a feladatot! (Below you'll find an instruction describing a task, along with an input provided to offer further context. Write a response that adequately fulfills the task!) ### Instruct: Generálj kulcsszavakat az alábbi szöveg alapján! (Generate keywords based on the provided text!) ### Input: [content of the article] ### Answer: [keywords]
OpenChatKit
[content of the article] <Q>: Generálj kulcsszavakat a megadott szöveg alapján! (Generate keywords based on the provided text!) <A>: [keywords]

A jelen kutatáshoz azután a 145 cikket tartalmazó teszthalmaz eredményeit használtuk. A 145 tételt végignéztük, és kiszűrtük belőle a biztosan hibásan felismert vagy hibásan konvertált szövegeket (például azok az esetek, ahol a kulcsszavak nem vesszővel, hanem valamilyen más szimbólummal voltak elválasztva, és ezt nem ismerte fel az OCR), így kaptuk meg a végleges 133 db tételt. Ezekhez a következő adatok álltak rendelkezésre (1. táblázat).

1. táblázat. A PULI LlumiX 32K modell teszthalmazához tartozó adatok

A referenciakorpusz (a szerző által megadott kulcsszavak)	**A PULI LlumiX 32K modell eredményei**
A szövegben előforduló kulcsszavak.	
A szövegben előforduló kulcsszavak száma.	
A szövegben nem talált kulcsszavak.	
A szövegben nem talált kulcsszavak száma.	

A szerzők által megadott kulcsszavak száma összesen a 133 szövegnél 621 db, míg a PULI LlumiX 32K modell eredményeinél 658 db, átlagosan tehát 4,67 kulcsszó volt szövegenként a refereciaanyagban és 4,95 a modell anyagában.

2.2. A kutatás kérdései és a vizsgálat módszerei: a kapott kulcsszavak kategorizálásai

A kutatás során arra voltunk kíváncsiak, hogy a nyelvmodell által kapott kulcsszavak mennyiben újak; mennyiben újak a szöveghez képest – tehát szerepelnek-e a szövegben – és mennyiben újak a szerzői kulcssszavakhoz képest. Kíváncsiak voltunk továbbá, hogy az eltérések milyen típusúak, és hogy relevánsnak tekinthetők-e a szöveg témáját és a szerzői kulcsszavakat tekintve.

Ezeknek a kérdéseknek a megválaszolásához a modell által kapott kulcsszavakat négy csoportba soroltuk:

1. **Teljesen új**: a kapott kulcsszó nincs a szövegben és nem egyezik a szerzői kulcsszóval.
2. **Tanult**: a kapott kulcsszó nincs a szövegben, de megegyezik a szerzői kulcsszóval.
3. **Kinyert**: a kapott kulcsszó szerepel a szövegben, de nem egyezik a szerzői kulcsszóval.
4. **Megegyező**: a kapott kulcsszó szerepel a szövegben és megegyezik a szerzői kulcsszóval.

A következőkben az első és a harmadik (tehát, amelyek eltértek a szerzői kulcsszavaktól) csoport kulcsszavait vizsgáltuk. Ekkor arra voltunk kíváncsiak, hogy az új kulcsszavak milyen típusúak, hogyan térnek el a szerző által adott kulcsszótól.

A kategóriák ez esetben a következők voltak:

1. Eltérő végződés (képző és rag is).
2. Eltérő helyesírás/írásmód (jelölt-jelöletlen szerkezetek is).
3. Fordítás: mindkét irányú.
4. Speciálisabb kulcsszó a szerzőjéhez képest: a kulcsszó kiegészül egy plusz tulajdonsággal, szerkezetként, összetételként szerepel (pl. *kiképzés – katonai kiképzés*).
5. Általánosabb kulcsszó a szerzőjéhez képest: az előző folyamat fordítottja (pl. *digitális kompetencia profil – digitális kompetencia*)
6. Rövidítés/potenciális szinonima.
7. Eltérő/nem kategorizálható: látszólag nem kapcsolódó, vagy nem közvetlenül kapcsolódó kulcsszavak, illetve asszociatív viszonyban állnak egymással.
8. Szerkezet: mellérendelést tartalmazó szerkezetek, illetve tagmondatjellegű szerkezetek.

A kategóriák után néhány példát részletesebben is megvizsgáltunk, illetve a szerző által adott, a szövegben nem szereplő esetek közül is bemutatunk néhányat.

2.3. Kulcsszavak a fogalmi struktúrában

Kétségtelen, hogy a hierarchikus rendszerezés az, amely a leginkább megfelel a dolgok rendszerré szervezéséhez.

A kulcsszavak tágabb értelemben véve terminusoknak tekinthetők (vö. Dodé, 2024), így vizsgálhatjuk őket a fogalmak rendszerében, ahogy azt a terminológiatudomány teszi. A „fogalmak nem elszigetelt tudásegységként léteznek, hanem mindig egymáshoz viszonyítva" (ISO 704: 2022)[6].

A fogalmi rendszer asszociatív és hierarchikus kapcsolatokból áll (ISO 704: 2022). A hierarchikuson belül beszélhetünk általános (generikus-specifikus) és partitív (rész-egész) viszonyokról. Egy specifikus fogalom feltételez egy általános vagy általánosabb fogalmat is, amelybe/amelyhez tartozik.

Szövegszinten az általános és a specifikus kapcsolatot lehet tetten érni a kollokációkon keresztül. Egy-egy kulcsszó összes alakjának kollokációira rákeresve, fel tudunk térképezni egy-egy fogalmi részrendszert (Kis–Pohl, 2005: 224).

A fogalmak nemcsak hierarchikus kapcsolatokban állhatnak egymással, hanem asszociatív viszonyokban is. Ez azt jelenti, hogy a fogalmak között szorosabb, de nem hierarchikus kapcsolatok is létrejöhetnek. Ennek számos fajtája van a *Terminology work — Principles and methods* című nemzetközi ISO-szabvány alapján (ISO 704: 2022):

1. sequential relation: szekvenciális kapcsolat – az objektumok valamilyen rendezésén alapul (pl. idő, következmény);
2. activity relation: tevékenységen alapuló kapcsolat (pl. ágens kapcsolat, célkapcsolat);
3. origination relation: az objektum eredetére vonatkozó kapcsolat (eredetkapcsolat, hozzávalókapcsolat);
4. instrumental relation: eszközkapcsolat (objektum-eszköz kapcsolat);
5. interactional relation: kétirányú kapcsolat az objektumok között (függőségi kapcsolat);
6. transmission relation: feladó–címzett elv alapján (feladókapcsolat, fogadókapcsolat);
7. opposite relation: ellentétes kapcsolat (ellentmondásos kapcsolat).

Ezeket a típusú kapcsolatokat azonban a szövegből, szaktudás nélkül, manuális úton nehéz detektálni, illetve ezeknek a kapcsolatoknak, relációknak a feltérképezésére az ontológia hivatott (Szakadát–Szőts–Szaszkó, 2006).

2.4. A vizsgálat lehetséges hibái

A kutatásnak vannak bizonyos hibafaktorai. Az egyik a szövegfelismerésből fakad. Bár az eredményeket átnéztük és a kulcsszavaknál detektált hibásan felismert szövegeket kiszűrtük, azonban a tanulmányok szövegében szintén előfordulhatnak felismerési hibák, amik ronthatják a keresés eredményét. Ilyen

[6] „Concepts do not exist as isolated units of knowledge but always in relation to each other."

például az elválasztás, vagy akár a nem felismert szimbólumok esete, illetőleg azokban a tanulmányokban, amelyek hasábos elrendezésben jelentek meg, különböző 'feldarabolt' szavak.

A másik hibája, torzítása az eredményeknek abból fakadhat, hogy a szövegbeli előfordulást egy egyszerű kereséssel végeztük – szó szerinti egyezéseket számoltunk, miután a szöveget kisbetűssé alakítottuk –, amely így a módosult végződésű, illetve bizonyos összetételekben előforduló kulcsszavakat valószínűleg nem találta meg.

2.5. Zero-shot prompting GPT-4o mini modellel

Kutatásunkban arra is kíváncsiak voltunk, hogy egy népszerű, sokak által használt nagy nyelvmodell – a GPT-4o mini[7] – milyen kulcsszavakat generál a szerzői kulcsszavakhoz (ilyen jellegű kutatás: Dodé, 2023) és a PULI LlumiX 32K modellhez képest. Az eredmények összehasonlításánál és értékelésénél fontos megjegyezni, hogy míg a PULI LlumiX 32K modell finomhangolva lett a feladatra, tehát előre megkapott 1000 szöveget és a hozzájuk tartozó szerzői kulcsszavakat, addig a GPT-4o mini modell csupán egy promptot kapott, amiben ugyanazt kértük tőle, mint a PULI LlumiX 32K modelltől:

> Instruct: „Generálj kulcsszavakat az alábbi szöveg alapján! (Generate keywords based on the provided text!)"
> Imput: [content of the article]

Ennek a technikának a neve a zero-shot prompting. Ez egy olyan promptolási technika, ahol a nyelvmodell nem lett előzetesen betanítva és példát sem lát egy adott feladatra, csupán egy promptra adott válaszként reagál az instrukcióra (Syed–Gadesha, 2025). Ennek fényében tehát az összehasonlítás nem validálható, viszont bizonyos minták kirajzolódnak a finomhangolt nyelvmodell és a zero-shot prompting közti különbség kapcsán.

3. Eredmény

3.1. Az új kulcsszavak kvantitatív elemzése

A 133 szöveghez összesen 621 szerzői kulcsszó tartozott, amelyből 23%, azaz 143 nem szerepelt a szövegben. A nyelvmodell által kapott kulcsszavaknál az eredmény 21,9% volt, azaz a 658 darab kulcsszóból 21,9%, 144 darab nem volt megtalálható a szövegben a keresési metódusunkkal. A módszereknél vázolt négy csoport eredményeit a 2. táblázat tartalmazza.

[7] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

2. táblázat. Az új kulcsszavak négy csoport szerinti eredményei

Teljesen új		Tanult		Kinyert		Megegyező	
db	%	db	%	db	%	db	%
122	18,5	22	3,3	245	37,2	269	40,9

A legkevesebb, kicsivel több mint 3% volt az olyan kulcsszavak száma, amelyek a szerző által adott kulcsszavakkal megegyeztek, azonban a szövegben nem szerepeltek. A kinyert és a megegyező kategóriák mennyisége hasonló volt (37,2% és 40,9%), míg a teljesen új is majdnem 20%.

3.2. Az új kulcsszavak kvalitatív elemzése

A teljesen új és a kinyert kategóriákban használt további csoportosítás eredménye a következő táblázatban (3. táblázat) látható.

3. táblázat. A teljesen új és a kinyert kulcsszavak további kategorizálása szerinti eredmények

	Teljesen új (db)	Teljesen új (%)	Kinyert (db)	Kinyert (%)
Eltérő végződés	1	0,8%	12	4,9%
Eltérő (helyes)írás	12	9,8%	17	6,9%
Fordítás: mindkét irányú	6	4,9%	7	2,9%
Speciálisabb kulcsszó a szerzői kulcsszóhoz képest	9	7,4%	14	5,7%
Általánosabb kulcsszó a szerzői kulcsszóhoz képest	5	4,1%	37	15,1%
Rövidítés/potenciális szinonima	7	5,7%	12	4,9%
Eltérő/nem kategorizálható	67	54,9%	141	57,6%
Szerkezet	15	12,3%	5	2%

A következőkben néhány példát mutatunk be kategóriánként.

4. táblázat. Példák a különböző kategóriák kulcsszavaira

	Referencia-kulcsszó (szerzői)	Eredménykulcsszó (nyelvmodell)
Eltérő végződés	*rehidrálás*	*rehidráció*
Eltérő (helyes)írás	*magia naturalis*	*mágia naturalis*
	egyetemi integráció	*egyetemintegráció*
Fordítás: mindkét irányú	*családkutatás*	*familysearch*
Speciálisabb kulcsszó a szerzői kulcsszóhoz képest	*kiképzés*	*katonai kiképzés*
	egyháztörténet	*egyháztörténet-írás*
Általánosabb kulcsszó a szerző által adott kulcsszóhoz képest	*gravitációs modell jel-kód*	*gravitációs modell*
	infrastruktúra mint szolgáltatás	*infrastruktúra*
	környezetvédelem	*környezet*
Rövidítés/potenciális szinonima	*szexualitás/szerelem*	*szerelem/szexualitás*
	büntetés-végrehajtási reintegrációs célok	*bv. reintegrációs célok*
Eltérő/nem kategorizálható	*református mikroközösség*	*református mikrotörténet*
	evangélikus iskoláztatás	*evangélikus oktatásügy*
	-	**szociokonómiai helyzet*
	-	*taxonómia*
Szerkezet	-	*bolygatott 10. századi temető*
	-	*minőségirányítás az oktatásban*

3.3. A GPT-4o mini által generált kulcsszavak

Mivel a GPT-4o mini modell nem lett előre betanítva – és az instrukcióban sem maximalizáltuk a kulcsszavak számát –, így nem tudott rátanulni a kulcsszavak számára, és minden esetben jóval több kulcsszót generált (a maximum 50 db volt). A modell válaszára többféle példát lehetett látni:

1. Bevezetővel kezdi a választ: pl.: *A szöveg alapján az alábbi kulcsszavakat lehet generálni:,* illetve ***Kulcsszavak:***.
2. Bevezetővel kezdi, majd le is zárja: *pl.*: *Íme néhány kulcsszó, amely a megadott szöveg alapján generálható: … Ezek a kulcsszavak segítenek összefoglalni a cikk főbb témáit és koncepcióit.*
3. Csak felsorolja (számozva vagy nem számozva).
4. Valamilyen további szimbólummal, írásjellel számozva sorolja fel (pl.: **).

Továbbá az esetek jórészében nagybetűvel kezdődnek az általa generált kulcsszavak.

A GPT-4o mini modell által generált kulcsszavak összevetése a szerzői kulcsszavakkal, illetve a PULI LlumiX 32K modell által generált kulcsszavakkal az alábbi táblázatban (5. táblázat) látható.

5. táblázat. A teljesen új és a kinyert kulcsszavak további kategorizálása szerinti eredmények

Szerzői kulcsszóval megegyezik (db, %)	PULI LlumiX 32K modell által generált kulcsszóval megegyezik (db, %)	Egyikkel sem egyezik (db, %)
129	42	450
20,77%	6,76%	72,46%

A GPT-4o mini modell által generált kulcsszavak esetében annyit vettünk vizsgálat alá, amennyi szerzői kulcsszó is szerepelt a szövegnél. Az összesen 621 db kulcsszóból 72,46% eltért valamilyen módon a szerzői és a PULI LlumiX 32K modell által generált kulcsszavaktól. 20,77% megegyezett a szerzői kulcsszavakkal és 6,76% volt, amely csak a másik modell által generált kulcsszavakkal egyezett meg.

A GPT-4o mini modell eredetisége nem csupán kreativitásból fakadt, előfordult, hogy a folyóirat címét adta meg kulcsszóként. Jellemző volt a túl általános kulcsszavak megléte, illetve bizonyos esetekben nehéz volt detektálni, hogy ugyanahhoz a szöveghez adott-e meg kulcsszavakat, vagy csak a gyakori szavakat vette kulcsszónak. A néhány PULI LlumiX 32K modellel való egyezés az eltérő írásmódból fakadt, illetve az általánosabb fogalom használatából.

A következőkben néhány példa látható (6. táblázat). Az első lista minden esetben a szerzői kulcsszó, a második a PULI LlumiX 32K modell eredménye, míg a harmadik a GPT-4o mini modell által generált kulcsszólista.

6. táblázat. Példák az eltérő és túl általános kulcsszavakra

1.
páros összehasonlítás, előrejelzés, valószínűség számítási háttér, csoportok összefésülése
páros összehasonlítások, csoportok összefésülése, előrejelzés, valószínűségi háttér
Thurstone módszer, sporteredmények, női kézilabda, Bajnokok ligája
2.
gamifikáció, oktatás, game engine-ek, szoftverfejlesztés
gamifikáció, oktatás, programozás, játékmotor
Online konferencia, Networkshop 2022, Debreceni Egyetem, HUNGARNET Egyesület
3.
tanárképzési reform, szaktanárképzés, gyakorlóiskolák, képzési kimeneti követelmények
tanárképzési reform, szaktanárképzés, gyakorlóiskolák, képzési kimeneti követelmények, tanárképzés, tanárképző központ, szaktanár, osztatlan képzés
Tanárképzés, Megújulás, Dilemmák, Lehetőségek

Következzen néhány példa, amikor kulcsszóként a terminusok helyett tulajdonnevek jelentek meg.

7. táblázat. Példák az eltérő és túl általános kulcsszavakra

1.
önkéntes katonai szolgálat, felkészítés, kiképzés
haderőfejlesztés, önkéntes tartalékos szolgálat, katonai kiképzés
Önkéntes katonai szolgálat (ÖKSZ), Honvédelmi Minisztérium, Magyar Honvédség
2.
alkotmányosság, értelmiség, jogtörténet, demokrácia
alkotmányosság, értelmiség, jogtörténet, demokrácia
Polgári Szemle, Recenzió, István Kukorelli, Három István

4. Konklúzió

Az elemzés eredményei azt mutatják, hogy a nyelvmodell sikeresen megtanulta, hogy megfelelő számú kulcsszót generáljon az egyes szövegekhez, hiszen átlagosan 4,95 kulcsszót rendel minden szöveghez, míg a szerzők átlagosan 4,67-et. A modell által generált kulcsszavak 40,9%-a és 3,3%-a, tehát összesen 44,2% megegyezik a szerző által megadott kulcsszavakkal. A kulcsszavak 21,8%-a olyan, amely nem található meg közvetlenül a szövegben, ami arra utal, hogy a modell ezek közül sokat saját kreatív módon generált. A teljesen újak között sok a releváns és találó kifejezés.

Bizonyos esetekben a kulcsszavak kategorizálása kihívást jelentett, például a potenciális szinonimák és fordítások esetén, mint pl. *etimológia–szótörténet*. Emellett pedig mindenképp nehézséget okoz a szakterületi háttértudás hiánya, mivel a fogalmi struktúrát is annak ismeretében lehet átlátni.

A szerzői kulcsszavakhoz képest az új kulcsszavak típuseloszlása különbséget mutat az első, teljesen új és a harmadik, kinyert kategória között. Az eltérő/nem kategorizálható kulcsszavak aránya mindkét kategóriában közel azonos (54,9% és 57,6%), bár a kinyert kategóriában a kulcsszavak általánosabb fogalmakat reprezentálnak a fogalmi rendszerben, mint például *középkor*, *fejlődés* vagy *pedagógusok*. Az ilyen általánosabb kulcsszavak inkább egy magasabb absztrakciós szintet tükröznek, és ezzel szembe mennek az UNESCO által megfogalmazott specifikusság elvével, ami arra vonatkozik, hogy egy kulcsszónak kifejezetten az adott dokumentumra jellemzőnek kell lennie, és nem lehet alkalmazható általánosan hasonló dokumentumokra. A szerzői motiváció természetesen diktálhat más stratégiát.

A modell által generált teljesen új kulcsszavak specifikusabbak és kreatívabbak, mint azok, amelyeket a nyelvmodell közvetlenül a szövegből nyert ki. Ezek általában többelemű kulcsszavak (vö. összetett terminusok), melyek túlnyomórészt nominális szintagmák és szerkezetek. Ez a tendencia látszik az általánosabb kulcsszó a szerzőjéhez képest kategória arányainál is, mivel 15,1% az arány a teljesen új kulcsszavak 4,1%-ához képest. Ezen túlmenően a teljesen

új kulcsszavak esetében a szerkezetek és a speciálisabb kulcsszó a szerzőjéhez képest kategóriák aránya magasabb (12,3% vs. 2%, illetve 7,4% vs. 5,7%), ami arra utal, hogy a modell generálásából származó kulcsszavak kreatívabbak és speciálisabbak, mint azok, amelyeket kinyert a szövegből a nyelvmodell.

A szerkezeteket azért fontos kiemelni továbbá, mert ilyen jellegű szerkezeteket a szerzői kulcsszavak között is látunk. Például: *emberi és isteni ítélet*, *katona-egészségügy története*, *idősoros vs panelregresszió*. Ezek a szerkezetek a tanulmány tematikájának esszenciáját képesek adni, így hatékonyan reprezentálják a szöveget, még ha így nem is tekinthetők terminusnak.

A szerkezetek továbbá a specifikusság elvét követik. A teljesség elve kapcsán, miszerint a kulcsszavaknak le kell fedniük a dokumentum minden olyan témáját, amely potenciális információs értékkel bír, beleértve az implicit, a dokumentumban nem kifejezetten megjelenő témákat is, megjegyzendő, hogy sok tanulmánynál maximalizálva van az adható kulcsszavak száma.

A zero-shot prompting eredményei rámutattak arra, hogy a prompton érdemes lenne módosítani, mivel a kimenet nem egységes, és így közvetlenül nem felhasználható az eredmény. Ezenkívül látható, hogy a GPT-4o mini eredménye bizonytalan, hibázhat. Sokszor nagyon általános kulcsszavakat generál, amelyek nem segítik elő sem a keresőmotorok hatékonyságát, sem a szöveg tartalmának reprezentálását.

A nyelvmodell szerzői kulcsszavakkal történő finomhangolása tehát mindenképpen hasznos, viszont a modellt érdemes további adatokkal finomhangolni annak érdekében, hogy kevésbé operáljon a túl általános, szövegből kinyert kulcsszavakkal.

Az egyelemű kulcsszavak (vö. egyszerű terminusok) kevésbé jól reprezentálják a szöveget (pl. *attitűd*, *olvasás*, *értékelés*, *modellezés*), így érdemes lehet a későbbiekben a többelemű kulcsszavak generálására ösztönözni a nyelvmodellt. Követve a nagy nyelvi modellek fejlődési irányát érdemes lenne az emberi visszejelzésen alapuló megerősítéses tanulásos módszertanát (Ouyang et al., 2022) alkalmazni arra, hogy tovább finomítsuk a modellt és arra ösztönözni a nyelvmodellt, hogy absztraktív módon kreatív többelemű, esszenciális kulcsszót generáljon.

A kutatás kétféleképpen is hasznosítható. Egyrészt hozzájárulhat kulcsszógenerálási feladatokra finomhangolt vagy egyéb (például prompttal programozott) nyelvmodellek teszteléséhez, másrészt pedig segíthet a tanítóanyagok további finomhangolásában, hogy a modell a jövőben még jobb eredményeket érhessen el.

Irodalom

Andonian, A., Anthony, Q., Biderman, S., Black, S., Gali, P., Gao, L., Hallahan, E., Levy-Kramer, J., Leahy, C., Nestler, L., Parker, K., Pieler, M., Phang, J., Purohit, S., Schoelkopf, H., Stander, D., Songz, T., Tigges, C., Thérien, B. & Weinbach, S. (2023). *GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.* [Online]. Available: https://www.github.com/eleutherai/gpt-neox

Barrault L., Biesialska M., Bojar O., Costa-jussà M., Federmann C., Graham Y., Grundkiewicz R., Haddow B., Huck M., Joanis E., Kocmi T., Koehn P., Lo C., Ljubešić N., Monz C., Morishita M., Nagata M., Nakazawa T., Pal S., Post M. & Zampieri M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). In P*roceedings of the Fifth Conference on Machine Translation. Online: Association for Computational Linguistics*. (1–55).

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I. & Amodei D. (2020). Language Models are Few-Shot Learners. In *34th Conference on Neural Information Processing Systems* (1877–1901). Vancouver, Canada: Curran Associates, Inc.

Devlin J., Chang M.-W., Lee K. & Toutanova K. (2019). 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* (4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. 10.18653/v1/N19-1423

Dodé Réka (2023). Magyar kulcsszavak vizsgálata és kinyerésük eredményeinek összevetése – pilotkutatás. *Filológia.hu, 14*(1–4), 51–64. https://doi.org/10.59648/filologia.2023.1-4.3

Dodé Réka (2024). Kulcsszavak és terminusok vizsgálata a REAL repozitóriumának anyagán – pilotkutatás. In K. Fogarasi, D. Ittzés, & D. Mány (Eds.), *Tudásmegosztás, információkezelés, alkalmazhatóság: I. Nyelvhasználat* (58–60). Budapest: Akadémiai Kiadó. https://mersz.hu/dokumentum/m1262tia__60/#m1262tia_58

Dodé, R. & Yang Z. Gy. (2024a). Further Keyword Generation Experiment in Hungarian with Fine-tuning PULI LlumiX 32K Model. In *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) Proceedings* (pp. 20-24). Debrecen.

Dodé Réka & Yang Zijian Győző (2024b). Kulcsszógenerálás magyar nyelvű, hosszú szövegekből nagy nyelvi modellekkel. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XX. Magyar Számítógépes Nyelvészeti Konferencia* (257–267). Szeged: Szegedi Tudományegyetem.

Feldmann, Á., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang Z. Gy., Váradi T. (2021). HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In G. Berend, G. Gosztolya & V. Vincze (Eds.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2021* (29–36). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Firoozeh, N., Nazarenko, A., Alizon, F. & Daille, B. (2020). Keyword Extraction: Issues and Methods. *Natural Language Engineering, 26*(3), 259–291.

ISO. (2022). *ISO 704: 2022 Terminology work — Principles and methods*. https://www.iso.org/standard/79077.html

Jalalov, D. & Gaszcz, K. (2023). *Prompt Engineering Ultimate Guide 2023: Kezdőtől haladóig. Metaverse Post*. https://mpost.io/hu/prompt-engineering-ultimate-guide/

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers. 427–431. Valencia, Spain: Association for Computational Linguistics.

Jurafsky, D. & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.). (Letöltés: 2024. október 10.) https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Liu, R., Lin, Z., & Wang, W. (2021). Addressing extraction and generation separately: Keyphrase prediction with pre-trained language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3180–3191.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv. https://doi.org/10.48550/arXiv.1301.3781

Nemeskey, D. (2020). Egy emBERT próbáló feladat. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia*: *MSZNY 2020* (409–418). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Nemeskey, D. (2021). Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*: *MSZNY 2020* (3–14). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Nomoto, T. (2023). Keyword extraction: A modern perspective. *SN Computer Science, 4* (92), 1–15. https://doi.org/10.1007/s42979-022-01481-7

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J. & Lowe R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th Conference on Neural Information Processing Systems.* New Orleans, Louisiana.

Reynolds, L. & McDonell, K. (2021). *Prompt programming for large language models: Beyond the few-shot paradigm*. arXiv. https://doi.org/10.48550/arXiv.2102.07350

Sammet, J. & Krestel, R. (2023, September). Domain-Specific Keyword Extraction using BERT. In *Proceedings of the 4th Conference on Language, Data and Knowledge* (659–665).

Sun, Y., Qiu, H., Zheng, Y., Wang, Z. & Zhang, C. (2020). SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access, 8*, 10896–10906.

Syed, M. & Gadesha, V. (2025). *What is zero-shot prompting?* IBM, 29 January 2025. Megtekintve: 2025. március 30. https://www.ibm.com/think/topics/zero-shot-prompting

Szakadát István, Szőts Miklós & Szaszkó Sándor (2006). *Az ontológia fogalma, építése, kezelése*. Megtekintve: 2024. október 10. http://www.ontologia.hu/ontologies.pdf

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B. & Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.*

UNESCO. (1975). *UNISIST Indexing Principle* (SC. 75/WS/58).

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł. & Polosukhin I. (2017). Attention Is All You Need. In *Proceedings of the 31th International Conference on Neural Information Processing Systems.* Long Beach, CA, USA: Curran Associates Inc. 6000–6010.

Yang, Z. Gy., Novák, A., Laki, L. J. (2020). Automatic Tag Recommendation for News Articles. In *Proceedings of the 11th International Conference on Applied Informatics (ICAI 2020)* (442–451). Eger.

Yang, Z. Gy., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, Á., Laki, L. János, Ligeti-Nagy, N., Vadász, N., Váradi, T. (2023a). Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)* (247–262). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Yang, Z. Gy., Laki, L. J., Váradi, T. & Prószéky, G. (2023b). Mono- and multilingual GPT-3 models for Hungarian. In *Text, Speech, and Dialogue*, Lecture Notes in Computer Science (94–104). Plzeň, Czech Republic: Springer Nature Switzerland.

Yang, Z. Gy., Dodé, R., Ferenczi, G., Hatvani, P., Héja, E., Madarász, G., Ligeti-Nagy, N., Sárossy, B., Szaniszló, Zs., Váradi, T., Verebélyi, T. & Prószéky, G. (2024). The First Instruct-Following Large Language Models for Hungarian. In *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) Proceedings* (247–252). Debrecen: University of Debrecen.