VIRÁG NÁNDOR

Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
virag.nandor9910@gmail.com
https://orcid.org/0009-0000-6317-9755

Virág Nándor: A magyar WordNet alkalmazása főnevek szemantikai kategóriáinak definiálására
The Application of Hungarian WordNet for Defining Semantic Categories of Nouns
Alkalmazott Nyelvtudomány, Különszám, 2025/1. szám, 157–172.
doi: http://dx.doi.org/10.18460/ANY.K.2025.1.009

A magyar WordNet alkalmazása főnevek szemantikai kategóriáinak definiálására

The Application of Hungarian WordNet for Defining Semantic Categories of Nouns__ANY_szte_0009-0000-6317-9755

There are several factors that influence the formation of grammatical sentences. One of them is the underlying constraint imposed by predicates on the meanings of their arguments, called semantic selection. Semantic selection influences sentence formation to ensure that, in addition to being grammatically correct, sentences also have interpretable meanings (Komlósy, 2015).

The aim of the current research is to find a solution to automatically assigning category labels to nominal subjects that occur next to a verb. The basis of this comes from a previous work, where I attempted a clustering task on verbs based on their semantic selectional preferences using a fix set of five semantic category tags for their potential subjects, in which nouns were categorized manually into the five sets. Although the set of tags made it easy to identify patterns in the selectional preferences of verbs, it also limited the accuracy of this approach. Finding more flexible categories became the necessary next step.

A solution to more flexible noun categorization is the search for hypernymic labels. Hypernymy is a fundamental semantic relation that describes the hierarchical relationship between lexical elements based on how broad their meanings are (Bußmann és mtsai., 2006). This transitive relationship can be understood as if the lexical element with the more specific meaning were a subset of a set defined by an element with a broader meaning. Properties of hypernymy make it possible to build a recursive algorithm that can find the list of every hypernym a given word has. If we were to compare hypernym lists of the list of subjects a verb can have, we could find a shared “ancestor” that can act as their semantic category label.

The hypothesis of this research is that the common hypernym of words in a list can correspond to the selectional category governing their selection. Thus, the main question is whether an automatic system developed to assign appropriate semantic category labels to noun lists is well-suited to model semantic selection, and if so, how it should be made.

For this end, I used the Hungarian Wordnet (HuWN) (Miháltz és mtsai., 2008) as a knowledge base with over 40000 lexical elements. Based on the structure of HuWN, finding the optimal hypernymic label can be achieved through an algorithm using the lemmas, IDs, and hypernym IDs of words stored in well-structured synsets. By comparing the recursively gathered hypernym lists of all elements in a given word list, this algorithm identifies the first element that is a hypernym of all examined words (“the lowest common ancestor”), without yielding a hypernym that is either too narrow, and not all the examined words are hyponyms of, or too broad, and gives less information about the selectional constraint that was applied while constructing the word list.

In addition to performing the basic search function, the algorithm also allows searching for a label that fits only a certain percentage of the word list. By specifying what portion of the list must be matched, outlier subjects arising from verbs' metaphorical or idiomatic usage can be excluded. These outliers

often do not fit the core meaning of the verb or its selectional constraints, thus distorting the resulting label.

The algorithm constructed in this research can search the hypernymic label that optimally describes the semantic selectional constraint that was the driven force in the assembling of a list of words in many cases. Unfortunately, there are errors coming either from the hierarchical structure available or the lack of lexical elements in HuWN that prevents it from being a fully automatic annotation tool. However, it could serve as a semi-automatic annotation aid, thereby advancing the clustering task toward a more flexible system of semantic category labels and a more precise description of semantic selection.

Keywords: semantic selection, WordNet, hypernymy, automatic categorization

1. Bevezetés[1]

A természetes nyelv feldolgozása során gyakran keresünk olyan mintázatokat, csoportosítási lehetőségeket, amelyeknek a segítségével közelebb kerülhetünk egyes nyelvi jelenségek magyarázatához. Ilyen mintázat lehet az is, hogy az egyes igék vagy igék csoportjai milyen megszorításokat támasztanak azokkal az elemekkel szemben, amelyek argumentumként megjelennek mellettük. Az alaptag és az argumentumai közötti kapcsolat automatikus elemzésére morfológiai és szintaktikai szinten könnyen képesek vagyunk, azonban az ige által gyakorolt szemantikai szelekció automatikus vizsgálata nehezebb feladat. Ennek a problémának egy megoldási kísérletét ismerteti a jelen tanulmány.

Első lépésként olyan adatokra van szükség, amelyek elemzésével következtethetünk az igék által gyakorolt szelekciós korlátozásra. Erre össze kell gyűjteni az egyes igék mellett gyakran előforduló argumentumokat, amelyeknek vizsgálatával kereshető a szelekciós kategória. Ahhoz, hogy elindulhassunk egy ilyen keresés felé, szükség van egy olyan adatbázisra, amely tartalmazza a vizsgálandó elemek jelentésaspektusainak jellemzését, hogy azok között lehetséges legyen kapcsolatot találni. Erre alkalmas adatbázis a magyar WordNet (Miháltz és mtsai., 2008), amelynek alkalmazásával létrehozható egy olyan eszköz, amely az argumentumok közös hiperonimái alapján keresi a jelentésük alapján föléjük rendelhető szemantikai kategóriát.

Munkám során létrehoztam egy olyan félautomata eszközt, amely képes a szólistákhoz megtalálni a közös hiperonimát, amely a lista elemeinek szemantikai címkéjeként funkcionál. Ezt jelen kísérletben az igék mellett megjelenő főnévi alanyok kategorizálásának megoldása irányából közelítettem meg.

A tanulmány a következőképpen épül fel: a 2. fejezetben részletezem a kutatás elméleti hátterét, különös tekintettel a szemantikai szelekció szerepére és arra, hogy milyen módon segíthet ennek a felderítésében a hiperonímia. Szintén ebben a szakaszban ismertetem a kutatás során felállított hipotézisemet és kutatási kérdésemet. Ezt követően a 3. fejezetben a kutatás módszertanát írom le részletesen, a rendelkezésre álló eszközök bemutatásával. A 4. fejezetben a módszer konkrét megvalósulásaként létrejött programot mutatom be, majd az 5.

[1] Köszönöm a tanulmány két anonim lektorának értékes kérdéseit, hozzászólásait, észrevételeit és ajánlásait a tanulmányhoz, valamint témavezetőm, Szécsényi Tibor támogatását és iránymutatását a kutatás során.

fejezetben az eszköz létrehozása során felvetődő problémás esetekre, és azoknak potenciális megoldásaira térek ki. Végül a 6. fejezetben összegzem a munkámat.

2. Elméleti háttér

Az igei alaptagok változatos megszorításokat gyakorolnak a mellettük argumentumként előforduló elemeken. Chomsky (1965) szerint a mondatok generálása során vannak olyan mögöttes szabályok, amelyek megakadályozzák a teljesen szabad argumentumválasztást, azonban ezek nem a szintaxisban kódoltak, sőt, túlmutatnak a grammatikai szabályokon, ezért bár a mondatalkotás folyamatát befolyásolják, inkább a szemantika részét kell képezniük. Grimshaw (1979) a predikátumok által kifejtett különböző szelekciós megszorításokat két csoportra bontotta, az egyikben az argumentumok mondattani jellemzőit kontrolláló szintaktikai megszorítások vannak, a másikban pedig ezeknek az elemeknek a jelentését korlátozó szemantikai szelekció. A szemantikai szelekció azt befolyásolja, hogy a létrehozható grammatikus mondatok közül csak azok generálódjanak, amelyekben az alaptag és az argumentumok együttes jelentése is értelmezhető (Komlósy, 2015).

Amellett, hogy a szemantikai szelekciót a predikátum egy tulajdonságaként értelmezhetjük, tudjuk formalizálni is a hatását. Resnik (1997) alapján a szelekciót felfoghatjuk egy prior és egy poszterior eloszlás különbségeként, ahol a prior eloszlás az általános valószínűsége annak, hogy egy szó megjelenik egy tetszőleges pozícióban, a poszterior eloszlás pedig a feltételes valószínűsége annak, hogy ugyanez a szó megjelenik a vizsgált alaptag mellett. Amennyiben a kettő közötti különbség kicsi, az alaptag gyenge szelekciós megszorítást támaszt az argumentumaival szemben, ha viszont nagy, akkor a szelekció erős. Mivel ilyen módon megfigyelhetjük gyakorisági értékek alapján az adott alaptagok szelekciós preferenciáit, elő lehet állítani egy olyan eszközt, amely statisztikai alapon képes megkeresni a megfelelő szelekciós kategóriát az adott predikátumhoz.

Erre a tulajdonságra alapozva jelen kutatást egy korábbi, hasonló témájú munka előzte meg (Virág, 2023). Ebben a cél az igék csoportosítása volt klaszteranalízis segítségével a Magyar Nemzeti Szövegtárból (továbbiakban MNSZ2, Oravecz és mtsai., 2014) nyert nagyjából 70000 mondatos korpuszra alapozva. A klaszterezés alapja egyrészt az igék szemantikai szelekciója, másrészt a tematikusszerep-kiosztásuk volt a mellettük alanyi pozícióban megjelenő főnevek tekintetében: a csoportok olyan igéket voltak hívatottak összefogni, amelyek hasonló megszorításokat vetnek ki az argumentumukra. Ezen tulajdonságok vizsgálatához 100 igével és 54 alanyként előforduló főnévvel dolgoztam. Ezeknek a kiválasztása során a Szeged Dependency Treebank (Vincze és mtsai., 2010) adataiból készült gyakorisági listából választottam ki a kutatás során vizsgált igéket és főneveket, amelyek előreláthatólag az MNSZ2-ben is kellően gyakori elemek lesznek. Ezt követően minden igéhez 4000 elemű random mintát kértem le az MNSZ2-ből, ahol a mondatokban a választott főnevek valamelyike szerepelt alanyként. A

főneveket kézi elemzéssel öt szemantikai kategóriába soroltam az anyanyelvi beszélői intuícióm alapján. A kategóriák a következők voltak: humán, intézmény/csoport, nem élő (tárgy), absztrakt fogalom, élő, de nem humán.

Ezen a ponton merülhet fel az a probléma, amely életre hívta a jelenlegi kutatást. Egy alapvetően korpuszalapú, vagy akár korpuszvezérelt módszer megalapozásában nem szerencsés egy ilyen manuális folyamatra alapozni a munkát. Az anyanyelvi intuíció alapján történő megítélés ugyan általában megbízható, azonban nem mentes a hibáktól. Az elemzendő főnevek alacsony száma miatt abban a munkában ez még elfogadható metódus volt, azonban ha több főnév vizsgálatára akarjuk kiterjeszteni a kutatást, vagy esetleg nem szeretnénk behatárolni, hogy milyen főnevek állhatnak a vizsgált mondatokban alanyi pozícióban, már nem lenne megoldható a kézi annotálás. Ezen felül a kiválasztott főnevek szűrése sem egészen kielégítő, ugyanis ezeknél a gyakorisági listából azokra az elemekre szűrtem, amelyek a leggyakrabban szerepelnek alanyesetben. Ez a szűrési feltétel azonban nem biztosítja azt, hogy azokkal a főnevekkel lehet dolgozni, amelyek a választott igék preferált alanyai, így ez a probléma is megoldásra várt.

Emellett a kézi elemzés során megállapított kategóriák nem feltétlenül megfelelőek a szelekciós tulajdonságok vizsgálatára. Egyrészt amiatt, mert nem biztos, hogy az így megállapított kategóriák fedik az igék által valóban támasztott megszorításokat, másrészt pedig azért, mert a használt szemantikai kategóriák teljesen statikusak, kategorikusak voltak. Ez akkor lenne elfogadható megoldás, ha lenne egy pontosan definiált, széles körben elfogadott lista a szelekciós kategóriákról, amelyeket az igék alkalmaznak az alanyaikkal szemben, ilyen lista azonban nem áll rendelkezésre. Ennek következtében valamilyen flexibilisebb megoldást kell találnunk, amellyel úgy lehetünk képesek megtalálni azt a megszorítást, amelyet az ige alkalmaz a mellette megjelenő alanyok listájára, hogy az kevésbé mesterséges, és a kutatói intuíció helyett korpuszadatok vezérlik.

Ilyen rugalmasabb kategorizálási lehetőségként merült fel a hiperonímia lexikai relációján keresztül történő keresés. A hiperonímia egy alapvető szemantikai viszony egyes lexikai elemek között, amely esetben egy tágabb jelentéssel bíró lexikai elem fölérendelt szerepet tölt be más, szűkebb jelentésű elemekkel szemben, amelyek a fölérendelt elem jelentését részletezik vagy pontosítják, például a *gyümölcs* szó hiperonimája az *alma*, a *cseresznye* vagy a *narancs* szavaknak (Bußmann és mtsai., 2006). Két lexikai elem közötti hiperonim viszonyt felfoghatunk tehát egy IS-A, azaz *egyfajta* relációként, amellyel kifejezzük, hogy a szűkebb jelentésű elem a tágabb jelentésű elem által meghatározott halmaz részhalmaza, például *az alma egyfajta gyümölcs* (Resnik, 1993). Mindemellett a reláció tranzitív is, azaz nem csak a közvetlen hiperonima alapján meghatározott halmaznak része a szűkebb jelentésű elem, hanem az összes, fölötte közvetetten hiperonimaként megjelenő kifejezés által meghatározottaknak is, például *az alma egyfajta entitás*. A hiperonímia fenti

tulajdonságait figyelembe véve elképzelhetjük a relációt egy hierarchikus faként, ahol a levelek a specifikus jelentésű, hiponimával nem rendelkező elemek, és ezekből az egyes éleken felfelé haladva a csomópontokban találjuk a tágabb jelentésű elemeket, amelyek amellett, hogy önmaguk is saját jelentéssel bíró lexikai elemek, meghatározzák az alattuk elhelyezkedők halmazát is. Ezzel a felfelé irányuló mozgással pedig eljuthatunk a legáltalánosabb jelentésű, minden egyéb elemet magukba foglaló kategóriákig. Ha ezt a szelekciós megszorításokra értelmezzük, elmondhatjuk, hogy ha egy ige szemantikai szelekciós megszorítást tesz az egyik argumentumára, akkor az ige mellett az adott argumentumként megjelenő főnévi csoport feje hiponimája a szelekciós kategóriának. Így tehát a szavak egyes hiperonimái kvázi szemantikai címkeként működnek, amelyek segítségével be lehet azonosítani, hogy milyen csoportba kell tartozniuk. Elképzelhető, hogy az egy hiperonima alá besorolható kohiponimáknak nem mindegyikét preferálja az ige, így ha egy választott hiperonim címke felől közelítjük meg a kategóriát, akkor könnyen lehet, hogy ez túlgenerálna. Ezzel szemben a levelek felőli megközelítéssel elképzelve a túlgenerálás problémája kiküszöbölhető, ebben a felfogásban tehát nem azt mondjuk, hogy a potenciálisan előforduló elemek valamilyen meghatározott kategóriából kerülhetnek ki, hanem azt, hogy a ténylegesen megjelenő elemek kategorizálhatóak egy közös szemantikai csoportba. Annak elérésére, hogy a megtalált hiperonim kategória alól a realizálódott alanyok közül ne legyen kivétel, a legszűkebb jelentéssel bíró, de minden vizsgált elemet lefedő halmazt kell megkeresni.

Ahhoz, hogy egy erre alapozott automatikus eszköz létrehozható legyen, szükség van egy olyan tudásalapú adatbázisra, amely tartalmazza az egyes lexikai elemek közötti hiperonímiaviszonyokat. Erre alkalmas struktúrával kialakított adatbázisok lehetnek a WordNetek, amelyeket különböző nyelvészeti kutatásokban pontosabban illeszkedő kategóriák megállapítására már korábban is használtak (Ye, 2004). Jelen kutatáshoz a magyar WordNet (Miháltz és mtsai., 2008) (továbbiakban HuWN) állt rendelkezésre, amelynek az adatai alapján kíséreltem meg létrehozni egy automatikus szemantikai osztályozót az igék mellett alanyi pozícióban előforduló főnevek kategorizálására.

A kutatás hipotézise az, hogy egy szólista szavainak közös hiperonimája megfeleltethető a kiválasztásukat vezérlő szelekciós kategóriával. Ebből következően a következő kutatási kérdésre keresem a választ: egy olyan rendszerrel, amely főnévlistákhoz megfelelő szemantikai kategóriacímkét tud rendelni a HuWN felhasználásával, megfelelően modellálható-e az igei szemantikai szelekció az alanyokkal szemben, és ha igen, hogyan, milyen elvek mentén hozható létre ez a rendszer? A továbbiakban a létrehozott program működési módját ismertetem.

3. Kutatási módszer

A jelenlegi munkához hasonló kísérletre, amelyben más nyelveken az adott nyelvű WordNetre alapozva kerestek szelekciós kategóriaként működő hiperonimákat, már volt példa (l. Jurafsky & Martin, 2025), a cseh WordNetben például magának az adatbázisnak része egy ilyen jellegű kereső algoritmus (Pala és mtsai., 2011). Magyar nyelven a jelenlegi munkához részben kapcsolódó munkák például Héja és Gábor (2012) és Simon (2021), azonban ezek csak kisebb részben érintkeznek jelen témával, a szerző tudomása szerint pedig korábban nem készült még ilyen jellegű program a magyar nyelv vizsgálatára.

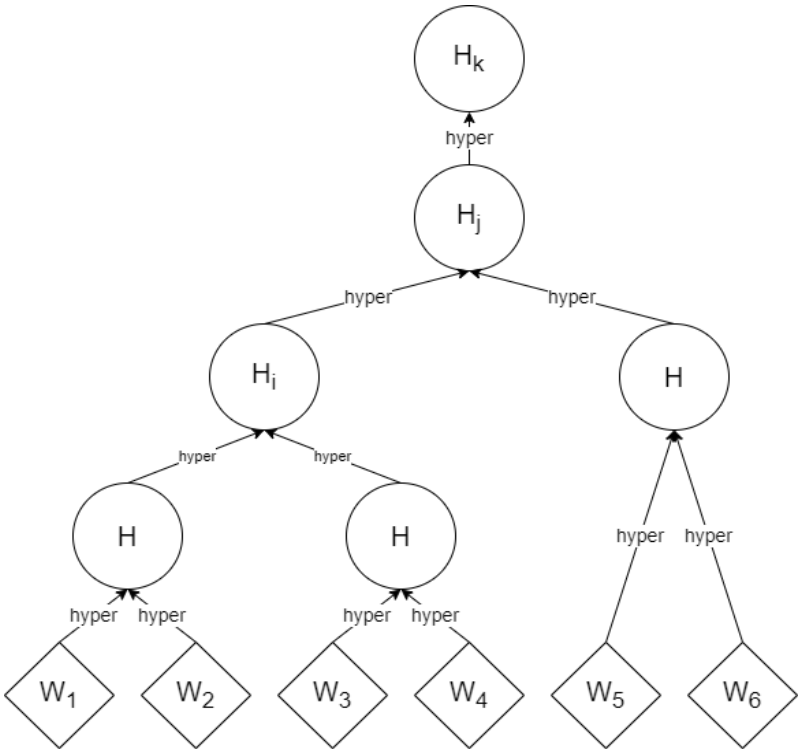
A rendelkezésre álló HuWN tehát alkalmas lehet a feladat elvégzésére. Ebben az egyes szavak jelentésére vonatkozó változatos információkat *synset*ekben tárolják, ilyen synsetből van az adatbázisban több mint 40000, melyek közül 33000 valamilyen főnévre vonatkozik (Vincze és mtsai., 2008). Az adatbázis építése során az eredeti angol nyelvű WordNet (Miller, 1995) konvencióit követték a készítők, miközben figyelembe vették a két nyelv között fennálló tipológiai különbségeket is (hasonlóan más, nem angol nyelvű WordNetek létrehozásához). A számos tárolt információk egyike minden szó esetében az is, hogy mely másik szó vagy szavak az adott lexikai elem hiperonimái. Ugyan a WordNetekben a hiperonímia reprezentációja nem tartalmazza inherensen a reláció tranzitivitását, ezzel kihívást állítva a nem közvetlen hiperonimák megtalálása elé (Cheng és mtsai., 2023), ez a tulajdonság a megfelelő algoritmus előállításával könnyen áthidalható.

A témában korábban végzett munkában főnevek fix listájához kézzel párosítottam kategóriák fix listájának tagjait, amely kategóriák az ige által kifejtett szemantikai szelekciós megszorításokat voltak hivatottak megjeleníteni. Így tehát főnevek listájának vagy halmazának közös felettes tulajdonsága alapján kerestem rájuk illő halmazcímkét. Jelen munkában nem a korábbi kutatáshoz összeállított szólistákat használtam, mivel ott korlátozva volt, hogy mely szavak fordulhatnak elő alanyként, és ez éppen az egyik olyan hiányosság, amelynek kiküszöbölésére jött létre a tanulmányban jellemzett algoritmus. A program itt bemutatott formája még csak tesztadatokon futott, ezek részben az MNSZ2-ből lekért random mondathalmazok alanylistái, részben speciálisan adott jelenségek (potenciális hibák, l. később) kipróbálására összeállított szólisták voltak. A kutatás egy későbbi fázisában szintén korpuszból kinyert adatokkal működne az elemzés, ezeknek a kinyerésére alkalmas eszköz lehet a Mazsola (Sass, 2018). A közös tulajdonságok keresése itt helyettesítődik a közös hiperonimák keresésével, itt a szólista szavainak, azaz adott ige mellett alanyként szereplő főnévi tövek összes hiperonimájának összehasonlításával megtalálható az a legszűkebb jelentésű kifejezés, amely minden listaelemnek közös hiperonimája, tehát a rájuk alkalmazott szemantikai szelekció kritériuma. Ahhoz, hogy ezeknek a hiperonim címkéknek a keresése lehetséges legyen, a synsetekben tárolt információk közül a listán szereplő szavak lemmáira (LITERAL), azonosítóira (ID) és a rájuk

vonatkozó hiperonima azonosítójára van szükség. Ezeknek az információknak a birtokában már létrehozható a kereső algoritmus.

A fenti információk tudatában könnyű elképzelni azt az eljárást, amellyel két szó közös hiperonimáját megtaláljuk egy rekurzív keresési módszerrel: először megkeressük minden szóra egyesével az összes hiperonimáját, egészen addig a csomópontig fellépkedve, amelynek már nincsen hiperonimája, majd az egyes szavakhoz tartozó hiperonimák listáit összehasonlítjuk. Ahol a listák először átfedésbe kerülnek, az jelöli azt a csomópontot, amely a két szóra vonatkoztatva a megfelelő szemantikai szelekciós címke lenne. Ezen az elgondoláson elindulva találhatunk megoldást a kereső algoritmusra: adott az egy ige mellett alanyként előforduló összes főnév listája, amely lista minden egyes eleméhez megkeressük az összes „ősét". Az így létrehozott listákat összehasonlítjuk, és megkeressük a legalacsonyabban lévő közös őst, amely elméletben az optimális címke a szólista kialakítását irányító szemantikai szelekció megértésére. Az 1. ábra ennek a keresési módnak a sematikus vizualizációját mutatja. Az algoritmus célja H_j megtalálása olyan módon, hogy elkerüljük a feltételeknek nem megfelelő H_i (túl kevés szót lefedő) és H_k (túl tág jelentésű) csomópontokat.

1. ábra. Sematikus faábrázolás az optimális hiperonim címke megtalálásához W_1-W_6 szavakra. A H_i címke nem fed le elegendő szót, a H_k címke túl tág jelentésű, a H_j címke optimális.



Felmerülhet olyan eset, amikor az ugyan elméletileg optimális H_j helyett mégis inkább a H_i-t kellene megkeresnünk. Ez akkor lehet hasznos, amikor a főnévlista egyes elemei egyértelműen kilógnak a többi közül, és ha mindegyik elem bevonásával keresnénk a szelekciós címkét, a talált csomópont vagy nem lenne

informatív, vagy akár az is megeshet, hogy nem is találna hozzájuk közös hiperonimát az algoritmus. Ilyen eset lehet például az idiomatikus, metaforikus, metonimikus kifejezésekben használt igék alanyainak bevonása a keresésbe. Normális körülmények között az az alany nem állhatna a vizsgált ige mellett, mivel nem felel meg a szelekciós megszorításoknak, azonban mégis ott szerepel. Például a *jön még kutyára dér* szólásban a *dér* nem tipikus alanya a *jön* igének, nem hasonlít más, a *jön* mellett gyakran előforduló főnevekhez, ezért ha keresnénk a H_j-t a *dér* bevonásával, nem kapnánk megfelelő eredményt. Természetesen metaforikus, metonimikus használat előfordulhat a hétköznapi nyelvhasználatban is, ez is problémát jelenthet, például ha azt mondjuk, hogy az *országgyűlés dönt*, akkor az *országgyűlés* valószínűleg nem illik bele a *dönt* alapvető szelekciós megszorításába. Ezzel további munkára lesz szükség, azonban előzetesen egy olyan helyzetet látok lehetségesnek, hogy valószínűleg van abban is valamilyen rendszerszerűség, hogy mely szavak fordulhatnak elő gyakori metaforikus használatban az ige mellett, például a *dönt* mellett valamilyen intézmények, csoportok, mint *egyesület*, *cég*, *vezetés*. Ha ezek valóban rendszerszerűen jelennek meg, akkor a prototipikus metaforikus használat ezzel megfejthető lenne.

Ezt a problémát olyan módszerrel lehet megoldani, hogy beállítjuk az algoritmust arra, hogy csak a szólista bizonyos százalékára keressen közös hiperonimát. Ezzel azt érjük el, hogy azokat a szavakat, amelyek nem érnek össze sehol a lista nagyobb részével, nem veszi figyelembe a keresésben a program. Ez a fent vázolt probléma megoldása felett további lehetőségek felé nyit utat. Megfelelő százalék megállapításával ugyanezen módszerrel a többjelentésű igék egyes jelentéseihez tartozó szemantikai szelekciós megszorításokat is megkereshetjük, és ezeket a szelekciós tulajdonságokat elválaszthatjuk egymástól. Ha maradunk az előző példánál, a *jön* ige jelenthet akár térbeli mozgást vagy időbeli bekövetkezést is, mindkét esetben eltérő alanyok állhatnak mellette, amelyeket ezzel a százalékos szűréssel el tudunk választani egymástól. Ehhez hasonló megközelítéseket a természetesnyelv-feldolgozás különböző területein széles körben alkalmaznak, például a jelentésegyértelműsítési (Word Sense Disambiguation) feladatokban elterjedt (Dhungana & Shakya, 2015; Resnik, 1995, 1997; Ye, 2004), bár az alkalmazhatóságával, hatásosságával, ismert hibáival kapcsolatban felmerülhetnek kritikák (Héja & Gábor, 2012). Emellett a HuWN fejlesztése során is használtak ehhez hasonló módszert (Miháltz és mtsai., 2013).

Mindezek alapján az algoritmus, amelyet létrehoztam a feladat megoldására, a következő elméleti lépéseket követi:

1. Bemenetként megkap egy főnévlistát, amely tartalmazza egy ige összes főnévi alanyát.
2. Megkeresi a synseteket, amelyeken a listán szereplő főnevek találhatóak.

3. Megkeresi az adott synsetben az ILR címke (az aktuális synset többi synsettel fennálló lexikai relációit tartalmazó adatok) alatt szereplő hiperonima azonosítóját.
4. Rekurzívan felfelé lépkedve összegyűjti az összes közvetett hiperonimát egészen a gyökérig.
5. Az összes főnévre elkészített hiperonimalistát egyesíti, és megkeresi ebben a listában az első közös csomópontot (vagy az első olyan csomópontot, amely a főnevek megadott százalékának közös őse).

A következő fejezetben bemutatom a megvalósult algoritmust, annak működését. Az érdeklődők számára elérhető a program a GitHubon[2].

4. Az algoritmus bemutatása, eredmények

A HuWN teljes adatbázisa egy XML formátumú fájlban érhető el, amelynek feldolgozása a megfelelő Python programmal nem okoz nehézséget. Első próbálkozásra az ElementTree (2024) használatával próbáltam kinyerni a megfelelő adatokat, azonban erről a módszerről hamar kiderült, hogy nem lesz tartható, mivel minden futtatás alkalmával újra és újra fel kellett dolgoznia az XML fájlt, ez pedig nagyon lelassította működését. Ahogy fentebb említettem, ennek a feladatnak az elvégzéséhez limitált információra van csak szükség a synsetekből, így ha el tudjuk érni, hogy a program azokat először nyerje ki, utána már nem kell a teljes fájlt feldolgoznia minden alkalommal, ami nagyban meggyorsítja a munkát. Ehhez az xmltodict (Blech, 2022) Python könyvtár segítségével hoztam létre a megfelelő adatokból szótárakat, azaz olyan rendezett struktúrájú adatsorokat, amelyekben egy kulcsot és egy értéket összepárosítva tárolhatjuk a szükséges információkat. Három ilyen szótár létrehozásával minden keresendő információt áthelyezhetünk a lassan elérhető XML fájlból belső objektumokba. Ezek a szótárak az id_lit_dict, lit_id_dict és id_hypo_dict voltak, ezek adatairól az 1. táblázatban található részletesebb információ. Az xmltodict könyvtár mellett még a numpy (Harris és mtsai., 2020) könyvtárat használtam a munka során elvégzendő különböző mérésekhez.

1. táblázat. A használt szótárak kulcsai és értékei

	id_lit_dict	**lit_id_dict**	**id_hypo_dict**
kulcs	a keresett szó azonosítója (ID)	a keresett szó lemmája (LITERAL)	a keresett szó azonosítója (ID)
érték	a keresett szó lemmája (LITERAL)	a keresett szó azonosítója (ID)	a keresett szó hiperonimájának azonosítója

[2] https://github.com/viragnandor/huwn_nounsem

A következőkben ismertetem a program egyes részeit, amelyeken keresztülmennek a bemeneti adatok, mire az algoritmus megtalálja a rájuk illeszthető hiperonim címkét. Az egyes szakaszok működésének illusztrálására példaként az MNSZ2-ből lekért, a *beszél* igét tartalmazó 100 mondat alanyaival történő futtatás adatait is megmutatom, hogy jobban értelmezhetőek legyenek a részletek. Ebben a mintában a *nő*, *barát*, *apa*, *anya*, *pénz*, *kormány*, *gyerek*, *tag*, *férfi*, *elnök* és *ember* főnevek voltak a *beszél* alanyai. Fontos megjegyezni, hogy az itt felhasznált mondathalmaz nem feltétlenül definitíven jósolja a megfelelő címkét az adott igére jelen esetben, csak a működés értelmezhetőségét szolgáló példaként szerepel jelen tanulmányban. A mondatok listája szintén megtalálható a fentebbi GitHub linken.

Egy lemma potenciálisan több synsetben is szerepelhet. A fenti szavak közül például a *barát* szó három különböző jelentéssel szerepel a HuWN-ben, három különböző azonosító alatt: ENG20-09242323-n (*barát* mint udvarló), ENG20-09247539-n (*barát* mint szerzetes) és ENG20-09463859-n (*barát* mint jó ismerős). Ez nem okoz problémát a program működésének szempontjából, mivel az minden egyes azonosítóval elkezdi keresni a szólista többi szavával közös hiperonimákat, és azt tekinti a megfelelő jelentésnek ilyen szempontból, amely beillik a többi szó jelentése közé, és talál velük közös hiperonimát. Ugyan a *barát* esetében mindegyik hasonló jelentéssel bírhat, jobban szemlélteti ezt a *kutya* szó, amely szerepel az adatbázisban háziállat és gazember jelentéssel is. Ezek közül a jelentések közül az adott szólista kontextusa segít megtalálni a programnak a megfelelőt.

Az első függvény, amely a hiperonim címke megkeresése felé vezet, a *findAllHyperonym*. Ennek a függvénynek egy synset azonosítója a bemenete, és megkeresi a synsetben foglalt szó összes hiperonimáját egészen a gyökérig, ezeket pedig egy listában tárolja el. Ha ezt a *barát* jó ismerős jelentést kódoló azonosítójával futtatjuk le, akkor a következő listát kapjuk: ENG20-00001740-n, ENG20-00003009-n, ENG20-00003226-n, ENG20-00005598-n, ENG20-00006026-n, ENG20-00016236-n. Ezek rendre a következő jelentéseket tartalmazzák: „entitás”, „élő vagy valaha élt entitás”, „szervezet, organizmus”, „kauzális ágens, kauzális erő, okozó, ok”, „egyén, halandó, személy, valaki, lélek”, valamint „dolog, fizikai dolog, objektum”. Következésképpen a szólistának ezen fogalmak közül valamelyik lesz a szelekciós tulajdonságot magyarázó hiperonim címkéje. Azonban ahhoz, hogy ezt megtudjuk, az összes szó összes hiperonimájára szükség van. Ehhez a fenti függvényt be kell ágyazni a következő függvénybe, amely egy szólista elemeinek közös őseit keresi.

A következő függvény az *ancestors* nevet kapta, ez keresi meg a szólista összes szavának hiperonimáit, ebbe ágyaztam be a fentebb tárgyalt függvényt. Bemenete lemmák listája (ellentétben az előzővel, ahol a keresett szó azonosítóját kellett megadni), kimenete pedig a megadott lemmák hiperonimáinak listáit tartalmazó lista. A példánál maradva tehát megadjuk bemenetként a főnevek listáját, majd

kimenetként azoknak a csomópontoknak az azonosítóit kapjuk listák formájában, amelyek rájuk igazak. Ezek különböző hosszúságúak lehetnek, a példa esetében a legrövidebb ilyen lista hét elemmel rendelkezik (*férfi*), a leghosszabb pedig hússzal (*tag*, *ember*). Értelemszerűen azoknak a szavaknak, amelyeknek lehet több jelentése, amely jelentések mentén külön-külön hiperonimaláncba kerülhetnek, több hiperonimát fognak tudni felsorakoztatni.

Ezen a ponton tehát rendelkezünk az összes szó összes hiperonimáival, meg kell keresni azokat, amelyek közösek. Ezt végzi el a *commonN* nevű függvény, amely lemmákhoz rendelt hiperonimalisták listáját és egy 0 és 1 közötti számot kér bemenetül. A listán tehát először lefuttatjuk az *ancestors* függvényt, ezt követően viszont ebből kiszűri azokat a csomópontokat, amelyek a lemmalista minden eleménél szerepeltek hiperonimaként. A 0 és 1 közötti számra azért van szükség, hogy tudjunk ne a teljes listához közös címkéket keresni, ahogy az előző fejezetben említettem. Itt tehát már lehetőség van arra, hogy kiszűrjünk olyan szavakat, amelyek nem illenek be a lemmalista elemei közé. A szám megadásával szabályozzuk, hogy a bemeneti lemmalista mekkora részére keresse a közös hiperonimákat, például ha 0,8-at adunk meg a programnak, akkor a lista 80%-ára kell, hogy illeszkedjen az a hiperonima, amelyet a függvény közösnek ítél. Amikor a függvény kiírja az eredményt, látunk egy azonosítót, amely az adott hiperonimát jelöli, illetve mellette azoknak a lemmáknak a listáját, amelyeknek az adott csomópont hiperonimája. A példaként használt szólistán, ha 0,8-as beállítással keresünk közös hiperonimákat, akkor hat ilyet talál a függvény. Ezek közül egyik például a „szervezet, organizmus” jelentésű ENG20-00003226-n, amelynek a szólista elemei közül hiponimái a *nő*, *barát*, *apa*, *anya*, *gyerek*, *tag*, *férfi*, *elnök* és *ember* szavak.

Itt már csak egy lépés választ el attól, hogy megtaláljuk a legalsó közös őst, mégpedig az, hogy a közös hiperonimákból ellenőrizzük, hogy melyik az, amelyik nem hiperonimája semelyik másik elemnek. Erre hivatott az utolsó függvény, amely a *lowestCommon* nevet viseli. Ugyanazokra a bemenetekre van szüksége, mint a *commonN*-nek, mivel ebbe beágyazva itt végzi el a közös hiperonimák keresését, és itt is megadhatjuk, hogy a szólista mekkora százalékára keresse a közös hiperonimát. Ha ez a függvény is sikeresen lefut, a *findAllHyperonym* függvény segítségével az algoritmus előállítja a közösnek ítélt címkék hiperonimáinak listáját. Ezután ellenőrzi, hogy a többi szó hiperonimalistáin szerepel-e a többi közös címke. Az lesz „a nyertes”, amelyik nem szerepel semelyik másik címkéhez tartozó listában, így azt tekinthetjük a legalsó közös ősnek, amely a szelekciós kategóriának feleltethető meg. A példa esetében 80%-os keresésnél ez az „egyén, személy” jelentésű ENG20-00006026-n lett, amely a 11 szavas listából 9 szónak hiperonimája. Ez megfelelő eredménynek tűnik, és ha belegondolunk, hogy a *beszél* ige alanyai általában emberek, megfelel az anyanyelvi intuíciónak is. Ideális esetben a függvények lefuttatása minden esetben talál közös hiperonimát a szólista elemeihez, ha nincsen benne olyan

elem, amely teljes kivétel lenne a többihez képest (l. fentebb az idiomatikus, metaforikus használat esetei). Ezeknek kiküszöbölésére alkalmas a beépített opció, hogy a szólistának bizonyos százalékára keressen csak közös őst az algoritmus. Ha ezen felül sem találna közös hiperonimát a program, akkor valamilyen problémás esettel állunk szemben, ezekről a következő szakaszban bővebben szólok.

Elképzelhető az is, hogy a függvény több közös őst talál a szólistához, leginkább akkor, ha nem 100%-ához keres közös hiperonimát, és az egyes elemek más csoportokba is beleilleszkedhetnek. Ennek a megoldása, a rendszer félautomata jellege miatt, az annotátorra marad, azonban ezekben az esetekben feltételezhető poliszémia a vizsgált ige esetén, így nem probléma, ha több potenciális hiperonim címke is felmerül. Erre alapozva működhet akár egy jelentésegyértelműsítési feladat is a program használatával, amely a különböző szelekciós megszorításokra alapozva választja szét a jelentéseket – ez azonban túlmutat a jelenlegi munkán, és további kutatásra lenne szükség a megvalósításához, de a jövőben felmerülhet jelen kutatás folytatásaként.

5. Problémás esetek és további lehetőségek

A működés és a példa ismeretében azt láthatjuk, hogy ez a program ebben a formában működőképesnek tűnik. Valóban, a tesztelés során sok esetben jól találta meg a hiperonimákat a megadott szólistákhoz, ezzel tehát megkönnyíthetné a kézi annotálást, és bizonyos esetekben ki is válthatná azt. Voltak azonban olyan problémás esetek, amelyeket érdemes megemlíteni, mert további feladatokat állítanak a folyamat automatizálása elé.

Az első ilyen probléma olyan esetekben merült fel, amelyben egymáshoz viszonylag közeli jelentésű szavakhoz kerestem hiperonim címkét. Vegyük például a *nyúl*, *kutya* és *ló* szavakat. Ha ezt adjuk meg bemenetnek, a program ennek a szólistának a legalacsonyabban elhelyezkedő őseként a „méhlepényes, méhlepényes emlős, valódi emlős” jelentésű csomóponthoz irányít. Ez a címke faktuálisan helyes információt közöl a szólistán szereplő szavak jelentésének közös tulajdonságairól, azonban intuícióellenesnek tűnik. Nem valószínű az, hogy az emberi elme szemantikai szelekciós folyamata során ez a kifejezetten tudományos, igen specifikus megszorítás merülne fel. Ennek a problémának az lehetne az orvoslása, ha lenne valamilyen előre definiált kategórialista, amelynek az elemeit mint végeredményt megtalálhatja a program – azonban ez nem egy jó megoldás. Egyrészt, ha megpróbálnánk intuíciónak megfelelően egy zárt listát összeállítani, amely tartalmazza az összes kategóriát, amely felmerülhet szemantikai szelekciós megszorításként, akkor voltaképpen semmissé tesszük az automatikusan található, flexibilis kategóriák keresését. Másrészt, ha nem a kutató saját anyanyelvi intuíciójára alapozva próbálnánk összeállítani ezt a listát, hanem például egy kérdőíves kutatás, vagy más, több anyanyelvi beszélő felől érkező input segítségével igyekeznénk megkeresni a kereshető kategóriákat, akkor

szintén az automatizálási feladat validitását csökkentenénk, illetve aránytalanul nagy befektetéssel járna ennek az egy problémának a megoldása. Ezeknél egyszerűbb megoldásnak kínálkozik, hogy ezen túl specifikus kategóriáknál ne a legközelebbi közös hiperonima legyen az alkalmazott kategória, hanem annak egy tágabb jelentésű hiperonimája (l. H_k az 1. ábrán). A félautomatikus annotálás során az annotátor által hozott döntés esetében ennek a problémának ez a megoldás gátat szabhat.

A másik említendő probléma még nagyobb nehézséget jelent. Ez az előzővel ellentétben olyan esetekben fordul elő, mikor egymástól viszonylag távoli jelentésű szavaknak keressük a közös hiperonimáit. Ilyen esetben előfordulhat, hogy a program nem talál egy közös hiperonimát sem, és egy üres listát kapunk eredményül. Ez már olyan szempontból is problémás, hogy az automatikus kategóriamegállapításnál előforduló ilyen esetben teljesen használhatatlanná válik a program. Ilyen esetre két példát említenék. Az első a program felépítésének és részeinek bemutatása során használt szólista. Ebben az esetben ugyanis, ha nem csökkentettük a közös hiperonimák keresését 80%-ra, nem talált volna semmilyen közös címkét nekik. Ellenőrizve a szólistát, a *pénz* szó az, amelynek a hiperonimalistája nem találkozik a többi elem listájával. Ehhez hasonló módon nem talál semmilyen közös címkét arra sem a rendszer, ha valamilyen tulajdonnevet próbálunk összekötni egy köznévvel, például a *Péter* és a *tanár* szavaknak semmilyen közös címkéje nincs, holott a valóságban könnyen lehet, hogy egyazon entitást jelöli ez a két elem. Mindezek mellett olyan esetben sem talál közös címkét a program, ha valamelyik keresett szó nem szerepel a HuWN adatbázisában. Ennek a problémának az oka végső soron a WordNet felépítése, a benne lévő adatok mennyisége és struktúrája. Erre a problémára sajnos jelenleg nem látok olyan megoldást, amely nem azzal végződik, hogy elhagyom a HuWN használatát, erről a továbbiakban írok bővebben.

A hiányzó közös hiperonimák esetei különböző megoldásokra szorulnak, amelyek komoly feladatokat jelentenek. A tulajdonnevek integrálásának problémáját egy névelem-felismerési lépéssel orvosolni lehet, ebben az esetben a konkrét tulajdonneveket helyettesíteni tudná az az információ, hogy milyen típusú névelemnek lehet őket beazonosítani (például *Péter = személy*). A nagyobb probléma az, ha egyes elemek hiányoznak a HuWN-ből. Ezeket vagy pótolni kellene, és bővíteni, javítgatni a HuWN adatbázisát, vagy elhagyni ezt az erőforrást mint tudásalapú adatbázist, és más módszerhez folyamodni.

Egy potenciális lehetőség a HuWN elhagyásával történő folytatásra az, ha nem keresünk helyettesítő tudásalapú adatbázist, hanem az ige mellett előforduló alanyok gyakorisági adataira alapozva „fuzzy” halmazokat alkotunk. Ehhez egy nagy méretű korpusz kialakítására lesz szükség, melynek megalapozásához potenciálisan használható lesz a Mazsola argumentumlisták kialakításához. Ezeket a halmazokat a központi, tipikus elemeik tulajdonságai határoznák meg, a tipikus elemek sok jellemzőben hasonlítanának egymásra, míg a perifériás esetek

néhány jellemzőben osztoznának csak a centrummal. Az így kialakuló mezők átfedései alapján találhatnánk bizonyos szemantikai szelekciós csoportokat, amelyekben a hasonló szelekciós preferenciájú igék hasonló szavakkal állnak gyakran, és ezekből a gyakori szavakból lehetne következtetni a szelekciós kategóriát jellemző kvázi címkére. Ez az ötlet azonban még nagyon korai stádiumban van, így jelenleg még csak egy a potenciális folytatási lehetőségek közül a jelenlegi struktúra javítása mellett.

Mindezekkel a problémákkal együtt értékelve a programot azt mondhatjuk, hogy bár automatikus eszközt ebből jelen formában még nem lehet létrehozni, nem sikertelen a kísérlet. Az így elkészült eszköz egy félautomata elemző, amely a kézi annotációt segíteni tudja, így a manuális elemzés hibalehetőségeit ilyen módon csökkenteni képes.

6. Összegzés

Jelen kutatásban egy automatikus eszközt kíséreltem meg létrehozni a HuWN felhasználásával, amelynek segítségével automatikusan meg lehetne állapítani egy vizsgált ige melletti összes főnévi alanyra vonatkozó szemantikai szelekciós kategóriát. A létrehozott eszköz teljesen működőképes, sok esetben valóban megtalálja a megfelelő hiperonim címkét a megadott szólistákhoz, azonban a korábban említett problémás esetek miatt (túl specifikus kategóriacímkék, hiányzó, a szólistákra nem illeszthető címkék) az automatikus működés még nem megoldható. Ettől függetlenül azonban egy használható félautomatikus eszköz született, amely megkönnyíti, sok esetben teljesen elvégzi a manuális címkézési feladatot, ami alapvetően a kiindulási pontja volt ennek a munkának.

Amellett, hogy ezt a kategorizálási feladatot javarészt megoldja az eszköz, utat nyitott egy másik, teljesen független, de azonos módszertanra alapozható jelentésegyértelműsítési próbálkozásnak. Ez nyilvánvalóan további kutatást és finomhangolást igényel, azonban egy, a jövőben megvalósítható projekthez remek kiindulási alapot szolgáltat jelen formájában is ez az algoritmus.

Jelen munka folytatásáról szólva, a legfontosabb dolog az automatizálhatóság elérése. Ehhez először is ki kell küszöbölni azokat a hibákat, amelyeket korábban ismertettem, valamelyik fent említett módszer segítségével. A közvetlen folytatás során igyekszem ezek közül többet is megkísérelni, akár a HuWN további használata mellett, akár annak elhagyásával.

Irodalom

Blech, M. (2022). *Xmltodict*. xmltodict. https://pypi.org/project/xmltodict/

Bußmann, H., Kazzazi, K., & Trauth, G. (2006). *Routledge dictionary of language and linguistics*. London: Routledge.

Cheng, X., Schlichtkrull, M., & Emerson, G. (2023). Are Embedded Potatoes Still Vegetables? On the Limitations of WordNet Embeddings for Lexical Semantics. In Bouamor, H., Pino, J. & Bali, K. (szerk.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (8763–8775). Szingapúr: Association for Computational Linguistics

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Dhungana, U. R., & Shakya, S. (2015). Hypernymy in WordNet, Its Role in WSD, and Its Limitations. *2015 7th International Conference on Computational Intelligence, Communication Systems and Networks* (15–19). Riga: Institute of Electrical and Electronics Engineers

Grimshaw, J. (1979). Complement Selection and the Lexicon. *Linguistic Inquiry*, *10(2),* 279–326. https://www.jstor.org/stable/4178109

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585(7825),* 357–362. https://doi.org/10.1038/s41586-020-2649-2

Héja Enikő & Gábor Kata (2012). Igék lexikai reprezentációja és a nyelvtechnológia. In Kenesei István, Prószéky Gábor & Váradi Tamás (szerk.), *Nyelvtechnológiai kutatások* (167–197). Budapest: Akadémiai Kiadó.

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3. kiadás, online kézirat). https://web.stanford.edu/~jurafsky/slp3

Komlósy András (2015). Régensek és vonzatok. In Kiefer Ferenc (szerk.), *Strukturális magyar nyelvtan I.* (274–486). Budapest: Akadémiai Kiadó.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., & Váradi, T. (2008). Methods and Results of the Hungarian WordNet Project. In Tanács, A. (szerk.), *GWC 2008: Proceedings* (311–320). Szeged: University of Szeged, Department of Informatics.

Miháltz, M., Sass, B., & Indig, B. (2013). What Do We Drink? Automatically Extending Hungarian WordNet With Selectional Preference Relations. In Popescu, O. & Lavelli, A. (szerk.), *Proceedings of the Joint Symposium on Semantic Processing : Textual Inference and Structures in Corpora* (105–109). Trento. https://aclanthology.org/W13-3828/

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38(11),* 39–41. https://doi.org/10.1145/219717.219748

Oravecz, C., Váradi, T., & Sass, B. (2014). The Hungarian Gigaword Corpus. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (szerk.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (1719–1723). Reykjavík: European Language Resources Association.

Pala, K., Čapek, T., Zajíčková, B., Bartůšková, D., Kulková, K., Hoffmannová, P., Bejček, E., Straňák, P., & Hajič, J. (2011). *Czech WordNet 1.9 PDT*. http://hdl.handle.net/11858/00-097C-0000-0001-4880-3

Python. (2024). *xml.etree.ElementTree–The ElementTree XML API*. https://docs.python.org/3/library/xml.etree.elementtree.html

Resnik, P. (1993). Semantic classes and syntactic ambiguity. In *Proceedings of the Workshop on Human Language Technology - HLT '93* (278–283). Princeton: Association for Computational Linguistics. https://doi.org/10.3115/1075671.1075733

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (448–453). San Francisco: Morgan Kaufmann Publishers Inc. https://doi.org/10.48550/ARXIV.CMP-LG/9511007

Resnik, P. (1997). Selectional Preference and Sense Disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?* (52–57). Washington D.C.: Association for Computational Linguistics. https://aclanthology.org/W97-0209.pdf

Sass Bálint (2018). Mazsola—Mindenkinek. In Vincze Veronika (szerk.), *XIV. Magyar számítógépes nyelvészeti konferencia* (16–24) Szeged: Szegedi Tudományegyetem, TTIK, Informatikai intézet.

Simon Gábor (2021). Egy keretalapú metaforaelemzés lehetőségei. Az azonosító konstrukció példája. *ARGUMENTUM*, *17*, 604–621. https://doi.org/10.34103/argumentum/2021/31

Vincze, V., Szarvas, G., & Csirik, J. (2008). Why are wordnets important? In *European Computing Conference. (ECC 08), WSEAS* (316–322).

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., & Csirik, J. (2010). Hungarian Dependency Treebank. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, &

D. Tapias (szerk.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC`10)*. European Language Resources Association (ELRA). https://aclanthology.org/L10-1321/

Virág Nándor (2023). *Igék szemantikai szelekciójának és tematikusszerep-kiosztásának korpuszalapú vizsgálata klaszteranalízis alkalmazásával*. Szakdolgozat. Szeged: Szegedi Tudományegyetem.

Ye, P. (2004). Selectional Preference Based Verb Sense Disambiguation Using WordNet. In *Proceedings of the Australasian Language Technology Workshop 2004* (155–162).