An Unsupervised Approach to Characterize the Adjectival Microstructure in a Hungarian Monolingual Explanatory Dictionary

Enikő Héja[1], Noémi Ligeti-Nagy[1], László Simon[2], Veronika Lipp[2]

[1]Hungarian Research Centre for Linguistics, Language Technology Research Group, Budapest, Hungary

[2] Hungarian Research Centre for Linguistics, Lexical Knowledge Representation Research Group, Budapest, Hungary

E-mail: {surname.forename}@nytud.hu

Abstract

The present paper describes the initial phase of a collaboration between Hungarian lexicographers and computational linguists aimed at compiling the new version of The Explanatory Dictionary of the Hungarian Language. This research thread focuses on the automatic sense induction of Hungarian adjectives in attributive positions, and their salient nominal contexts, with a particular emphasis on polysemies. The proposed methodology is intended to facilitate lexicographers' work in characterizing both the micro- and macrostructure of adjectives in a monolingual setting. A corpus-driven, unsupervised graph-based approach was employed, which, as per our expectations, could potentially reduce the reliance on human intuition, especially in the ambiguous domain of polysemic sense distinctions. Initially, distributional criteria for meaning distinction were introduced, followed by the description of the employed algorithm. The algorithm models adjectival semantics using two unique subgraphs: connected graph components are used to model adjectival semantic domains, while maximally connected subgraphs, so called cliques, model polysemies. Automatically induced meaning distinctions were validated using salient nominal context candidates extracted from corpus data. We expect that while connected graph components aid in characterizing the adjectival macrostructure, cliques provide lexicographers with useful insights for establishing the adjectival microstructure. These hypotheses were also tested: we investigated the extent to which the proposed framework can assist expert lexicographers during the dictionary compilation process by comparing a sample of our automatically obtained results to the previous version of The Explanatory Dictionary of the Hungarian Language.

Keywords: automatic sense induction, monolingual lexicography, polysemy, unsupervised graph-based approach, adjectives

1. Introduction

Although corpus-based methodology is increasingly central in monolingual lexicography, complemented by a variety of software tools and detailed guidelines (cf. Atkins & Rundell, 2008), we are not aware of any lexicographic projects employing a corpus-driven approach. Such an approach could significantly contribute to the field: notably, it could expedite the workflow and reduce the reliance on human intuition during the lexicographic process. This can be particularly useful in the nebulous area of meaning distinctions, thus assisting in the formation of the microstructure, a well-established challenge in both bilingual and monolingual dictionaries (Adamska-Sałaciak, 2006; Hanks, 2012; Véronis, 2003). A corpus-driven technique should strive to leverage corpus data to the fullest extent, with minimal human intervention. Consequently, establishing operationalizable distributional criteria for sense distinction is crucial. Regrettably, to our knowledge, there is no widely accepted distributional definition of polysemy that would allow for more data-driven, and hence more objective meaning distinctions (cf. Geeraerts, 2010).

This challenge is even more pronounced in the case of adjectives. Adjectives pose a significant difficulty when attempting to divide them into distinct senses (Moon, 1987). It is hard to analyze them in isolation because they essentially constitute an aspect of the modified noun. Furthermore, adjectival lexical semantics represents a relatively under-researched area in linguistics. While several attempts have been made to identify different verbal structures and their associated meanings based on distributional properties (e.g. Levin, 1993; Kipper-Schuler, 2005 and Sass et al., 2010 for Hungarian), we are not aware of any similar initiatives concerning adjectives. This is even more so in the case of Hungarian adjectives: to our knowledge, only Kiefer (2003, 2008) provides a detailed examination of adjectival semantics.

Accordingly, our primary objectives are: (1) to provide sufficient criteria to grasp adjectival sense distinction, including polysemies; (2) to model these criteria and (3) to evaluate the extent to which this technique can aid expert lexicographers to develop the adjectival microstructure of the new version of *The Explanatory Dictionary of the Hungarian Language* (EDHL). The EDHL is an up-to-date online dictionary of contemporary Hungarian (covering 2001–2020) that is being compiled using corpus-driven methods (Lipp & Simon, 2021).

As per our expectations, the automatically extracted adjectival subsenses should provide lexicographers with a ready-to-use adjectival microstructure, significantly facilitating their work. This hypothesis was tested from two distinct angles: First, approximately 60 automatically extracted polysemies were compared to the relevant microstructures of a traditional explanatory dictionary from multiple perspectives, including coverage and, most importantly, the motivatedness of meaning distinctions. In relation to this, special attention was devoted to the nominal contexts of the adjectives. We expect that the detected subsenses subcategorize certain semantic classes. Secondly, approximately 6400 adjectives from the Hungarian Webcorpus 2.0 (Nemeskey, 2020) were partitioned into semantic domains fully automatically. This partition was then compared with the macrostructure of the EDHL to examine the extent to which it could streamline the headword selection process.

2. Motivation

2.1 Lexicographic background

In lexicography, three distinct paradigms are employed: traditional, corpus-based, and corpus-driven approaches (Atkins & Rundell, 2008; Svensén, 2009). Within the traditional approach, lexicographers heavily rely on their linguistic intuition, which results in an imbalanced description of the relevant linguistic phenomena.

The two Hungarian monolingual general-purpose dictionaries of the 20th century, *A magyar nyelv értelmező szótára* [The Explanatory Dictionary of the Hungarian Language; EDHL] (Bárczi & Országh, 1959–1962) and *Értelmező kéziszótár* [Concise Hungarian Explanatory Dictionary; CHDL^1, CHDL^2] (Juhász et al., 1972; Pusztai & Csábi, 2003), were compiled using the traditional method. The editors of EDHL relied on their own mental lexicon throughout the dictionary creation process. As the leading editor asserts, “Our own language knowledge and language sense, which we constantly verified through surveys, served as the natural basis for our work in recording word meaning, usage, and stylistic value” (Országh, 1953: 397). Work on *A magyar nyelv nagyszótára* [Comprehensive Dictionary of Hungarian; CDH] (Ittzés, 2006–2021) began in 1985 based on a historical corpus. However, the limited size of the corpus (30 million words) did not provide sufficient data for dictionary writing.

To modernize linguistic research and link Hungarian lexicography to ongoing European projects, a text database of significant size and quality is needed. Databases like the Hungarian National Corpus (Váradi, 2002) (HNC) and the Hungarian Gigaword Corpus (Oravecz et al., 2014), while comparable to prominent corpora like the British National Corpus (Burnard, 2007) and Deutches Referenzkorpus (Kupietz et al., 2010), are not suitable for lexicographic research due to various limitations. Similarly, web-scrapped databases, such as the Hungarian Web Corpus (Jakubíček et al., 2013) are also insufficient due to their inbalanced nature and the limited metadata they provide.

The corpus-based lexicography focuses on word usage patterns and relies on the contexts in which words typically occur (Hanks, 2010). Senses and subsenses are established based on such information, utilizing suitable corpus tools. Taking a step further, the corpus-driven methodology aims to explore the meaning space of a word through fully automatic means, further reducing the reliance on human intuition. One of the significant advantages of this technique is its ability to handle vast data sets. In 2021, the Hungarian Research Centre for Linguistics initiated a project to update the EDHL, originally created in the 1960s, using automatic methods applied to a new, extensive, and representative input corpus. The primary objective is to obtain an objective lexical profile for each dictionary entry, anticipating that this information will expedite the creation of a new explanatory dictionary (Lipp & Simon, 2021).

2.2 Consistent methodology

Our proposed method aligns perfectly with the envisioned framework for creating the new version of EDHL. It not only relies on data but also leverages unlabeled data, apart from the part-of-speech annotation. This means that the algorithm processes data with minimal presuppositions about meanings. Moreover, our methodology is based on a substantial amount of data, especially from a lexicographic standpoint. The adjectival meanings are distilled from a subset of 170 million sentences, extracted from the Webcorpus 2.0 (Nemeskey, 2020). Contextual information is retrieved from the 180-million-word HNC. Furthermore, if needed, the amount of data utilized can be expanded.

The data-driven technique we employ relies on distributional criteria for meaning distinction, which we consider a novel contribution to the field. These criteria, in contrast to previous definitions based on etymology or sense relatedness, offer a more intersubjective approach. Additionally, they can be easily modeled using a simple graph-based approach.

Hopefully, the corpus-driven method can be enhanced through a meticulous lexicographic post-editing phase. The close collaboration between different fields ideally leads to the development of data-oriented, explicit lexicographic editing principles that apply to both the macrostructure and microstructure of the dictionary.

In the next section, we will present the distributional criteria for meaning distinction, followed by an overview of the unsupervised word sense induction experiment conducted on Hungarian monolingual data. The workflow can be conceptually divided into two main stages: i) The detection of subsense candidates for a given adjective, ii) discrimination between the different meanings of the given adjective by extracting relevant context nouns.

3. Distributional criteria for meaning distinction

3.1 Near-synonymy

First, let us recall the notion of near-synonymy (cf. Ploux & Victorri, 1998), a relaxed version of synonymy (cf. Frege, 1892), which is heavily relied upon when formulating the distributional criteria for meaning distinction. That is, two expressions are *near-synonyms* if they are interchangeable in a restricted set of contexts, preserving the meaning of the original sentence. For instance, the Hungarian adjectives *finom* 'fine' and *lágy* 'soft' are synonyms before nouns related to music, such as the Hungarian counterparts of 'music,' 'rhythm,' 'melody,' etc., as *lágy zene* and *finom zene* convey the same meanings. For the sake of the present research, the notion of near-synonymy is further extended: we also consider the members of tight semantic classes to be near-synonyms, as they denote different senses of a word, even though they may not preserve the truth value. This extension aligns with our original purpose of meaning distinction.[1]

3.2 Criteria for meaning distinction

Accordingly, an adjective has multiple meanings if:

1. There is (at least) one near-synonym for each sense of the adjective.
2. There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.
3. The two sets of context nouns that characterize the different senses are non-overlapping.
4. The non-overlapping set of nouns forms a semantic category, reflecting the sub-selectional properties of adjectives (Pustejovsky, 1995).

Example 1 illustrates the four criteria using two automatically extracted senses of the adjective *napfényes* ('sunny'). As observed, there is a near-synonym for each sense: *napsütéses* ('sunshiny') for the first sense and *napsütötte* ('sunlit') for the second sense. The listed nouns below the adjectives are those that form grammatical constructions with the respective near-synonyms, such as *napfényes/napsütéses vasárnap* ('sunny/sunshiny Sunday'), *napfényes/napsütéses nap* ('sunny/sunshiny day'), and *napfényes/napsütötte terület* ('sunny/sunlit area'), *napfényes/napsütötte terasz* ('sunny/sunlit terrace').

Importantly, the two sets of nouns do not overlap; there are no instances like **napsütéses terasz* ('sunshiny terrace') or **napsütötte nap* ('sunlit day'), and the same holds true for all adjective-noun pairs where the noun comes from the context noun set of the other sense. Finally, the nouns that match the above criteria form a semantic category: time periods with the first sense, and areas, places with the second.

[1] For example, *fekete* 'black' may belong to two different near-synonymy sets: one containing surnames and the other containing names of colors.

(1) **Sense 1:** *napfényes* 'sunny', *napsütéses* 'sunshiny'
Nouns of sense 1: *vasárnap* 'Sunday', *nap* 'day'

Sense 2: *napfényes* 'sunny', *napsütötte* 'sunlit'
Nouns of sense 2: *terület* 'area', *sziget* 'island', *oldal* 'side', *terasz* 'terrace'

4. Representation of the investigated phenomena

The present discussion is confined to a brief overview of the algorithm, possibly from a lexicographic perspective – with only the necessary amount of technical details. For more detailed information, please refer to Héja & Ligeti-Nagy (2022a,b). First, the representation of the input categories will be described, followed by the presentation of the various adjectival meaning representations and the related simple graph-theoretic concepts. Finally, we discuss how the salient nominal contexts were detected. It is important to emphasize that at this stage, the meaning representations are induced *fully automatically* from corpus data.

4.1 Selection of input adjectives

The adjectives of interest were selected based on the 180-million-word HNC. Specifically, we considered all the adjectives that occurred at least 2 times in the HNC.

4.2 Representation of adjectives

In the subsequent step, static vector representations (Mikolov et al., 2013a,b) were generated for the selected adjectives using the first 999 files (21GB of raw texts) from the Webcorpus 2.0 (Nemeskey, 2020). The cc. 170-million sentence training corpus consists of the normalized version of the original texts. To create the vector representations, 300-dimensional vectors were trained using the Gensim Python package (Rehurek & Sojka, 2011). The training was performed using the Continuous Bag-of-Words (CBoW) algorithm with a window size of 6k and a minimum frequency of 3. Roughly 8.5 million word forms were assigned embeddings. The trained language model (LMs) can be accessed at the following link: https://nlp.nytud.hu/word2vec/cbow_3.tar.gz.

While we acknowledge that static word embeddings have become outdated in the field of natural language processing, they still offer several advantages over more recent contextual embeddings. They are easy to train and handle, and importantly, they provide interpretability, which is crucial for lexicography. However, one drawback of this approach is the "meaning conflation deficiency" as described in (Camacho-Collados & Pilehvar, 2018), which states that such representations conflate the various subsenses of a lemma into one point in the semantic space.

In the subsequent sections, we will demonstrate that the meaning conflation deficiency can be effectively addressed through graph representations, particularly in the case of adjectival polysemies. This approach yields highly interpretable results and mitigates the limitations associated with static word embeddings.

4.3 Graph-based representation of adjectival meanings

Our methodology is based on the graph representation of adjectives. A graph is a mathematical structure composed of nodes and edges. In this context, nodes represent adjectives, while edges connecting two nodes represent whether the two adjectives are semantically similar. As can be seen in Figure 1, the ego graph[2] of *érzékeny* ('sensitive') includes all the adjacent adjectives to *érzékeny*, along with the edges between those adjacent adjectives. It demonstrates that the Hungarian adjective *érzékeny* is semantically similar to *gyengéd* ('gentle'), *törékeny* ('fragile'), and *fogékony* ('receptive'). As these latter nodes are not interconnected, they likely belong to different subsenses of the central adjective *érzékeny*.

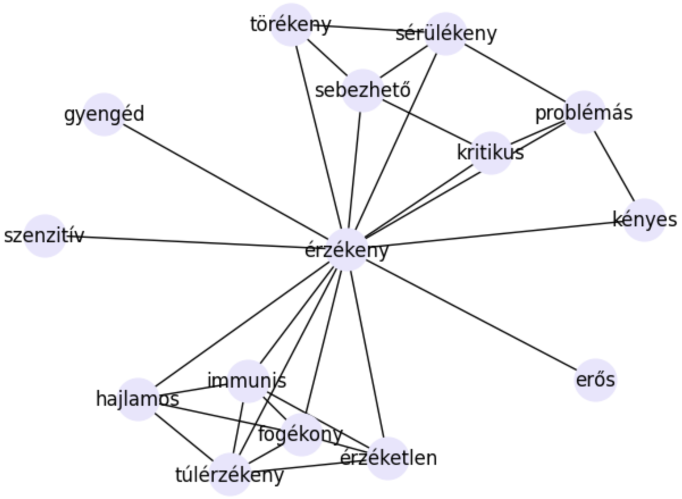


Figure 1: The ego graph of *érzékeny* 'sensitive'[3]

4.4 Representing near-synonymy classes as cliques

Following the generation of the graph representation of the adjectival semantic space, near-synonymy classes are modeled via maximally connected subgraphs, also known as *cliques*. A clique is a (sub)graph in which every node is connected to every other node in the (sub)graph (cf. Figure 2).

The basic premise of this representation is that in an adjectival clique, the meaning of each element is similar to that of every other element, thus, cliques are strong candidates for near-synonymy classes representing a (sub)sense of an adjective. Indeed, the meanings of *gyönyörű* 'beautiful', *csodaszép* 'stunning', *gyönyörűséges* 'gorgeous', *szépséges* 'lovely',

[2] The ego graph or ego network is a specialized type of graph consisting of a central node (the ego) and all other nodes directly connected to it (the alters). Edges between the alters also form part of the ego graph.

[3] *törékeny*: 'fragile', *sérülékeny*: 'vulnerable', *sebezhető*: 'susceptible', *kritikus*: 'critical', *problémás*: 'problematic', *kényes*: 'delicate', *erős*: 'strong', *immunis*: 'immune', *hajlamos*: 'prone', *fogékony*: 'receptive', *érzéketlen*: 'insensitive', *túlérzékeny*: 'oversensitive', *szenzitív*: 'sensitive', *gyengéd*: 'gentle'

[4] *gyönyörű*: 'beautiful', *csodaszép*: 'stunning', *gyönyörűséges*: 'gorgeous', *szépséges*: 'lovely', *meseszép*: 'fabulous', *tündéri*: 'adorable'

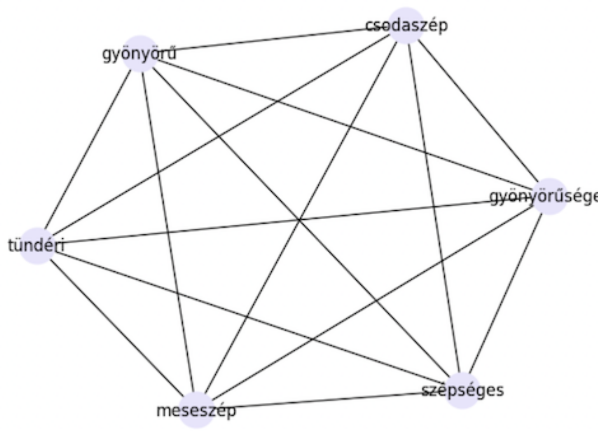


Figure 2: The clique modeling the near-synonymy class of *gyönyörű* 'beautiful'[4]

meseszép 'fabulous', and *tündéri* 'adorable' are highly similar, indicating that these adjectives belong to the very same meaning.

4.5 Meaning distinction: one adjective in multiple cliques

Consequently, in the next step, multiple subsenses of a lemma are to be modeled by multiple cliques. That is, an adjective may have multiple senses, if it belongs to multiple cliques (cf. criterion 1).

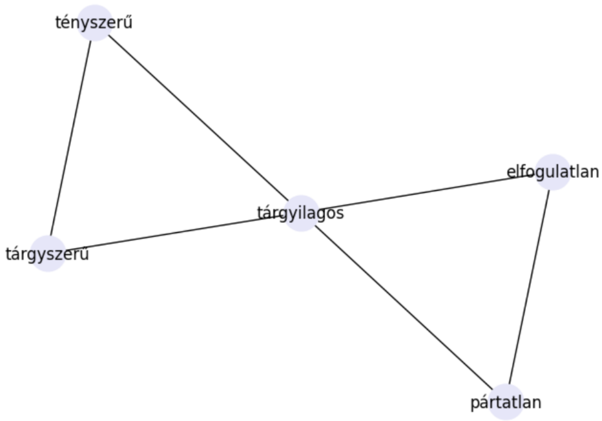


Figure 3: The Hungarian adjective *tárgyilagos* 'objective' belongs to two cliques[5]

For example, as illustrated in Figure 3, the Hungarian adjective *tárgyilagos* 'objective' belongs to two different cliques, indicating two distinct subsenses of the lexeme: clique 1 comprises *tárgyszerű* 'concise' and *tényszerű* 'factual' as near-synonym candidates, whereas clique 2 consists of *pártatlan* 'impartial' and *elfogulatlan* 'unbiased', representing a different subsense. Notably, this sense distinction is further underpinned by the following nouns (cf. criterion 3). The elements of clique 1 co-occur with nouns such as *leírás* 'description', *ismertetés* 'exposé', *vita* 'discussion', while adjectives in clique 2 co-occur with nouns

[5] *tárgyszerű*: 'concise', *tényszerű*: 'factual', *tárgyilagos*: 'objective', *elfogulatlan*: 'unbiased', *pártatlan* 'impartial'

like *megítélés* ’judgement’, *vélemény* ’opinion’, and *eljárás* ’procedure’. This outcome supports our intuition according to which the first sense of *tárgyilagos* is more objective corresponding to the facts, while the second sense is used rather in the sense of being impartial.

4.6 Clique validation via the following nouns

4.6.1 Extracting the nominal contexts

Three out of the four criteria for meaning distinction pertain to the nouns modified by the attribute adjectives: there should be (1) a set of (2) non-overlapping context nouns (3) that form coherent semantic classes, reflecting the sub-selectional properties of the adjectival near-synonymy sets. These three *clique validation steps* are vital to our workflow. They align with Levin (1993) and are predicated on the assumption that adjectives, similar to verbs, impose semantic selectional restrictions on their arguments. Consequently, tight nominal semantic classes are required to validate the adjectival subsense candidates. In cases of two meaning candidates, i.e., two shared cliques, criterion (2) and (3) can be expressed more formally as computing the symmetric difference of the nominal sets A and B, where A comprises nouns occurring after all adjectives in clique 1, and B includes nouns occurring after all adjectives in clique 2.

Let’s revisit example 1: *napfényes* ’sunny’ had two separate subsenses, *napsütéses* ’sunshiny’ and *napsütötte* ’sunlit’:

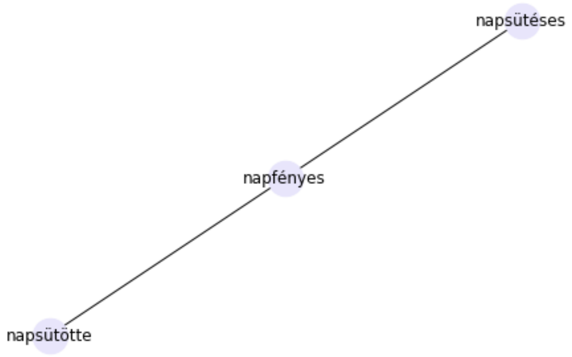


Figure 4: The Hungarian adjective *napfényes* ’sunny’ belongs to two cliques[6]

It was also claimed that the two separate submeanings are characterized by two distinct sets of nouns, as follows:

(2) **Sense 1:** *napfényes* ’sunny’, *napsütéses* ’sunshiny’
Nouns of sense 1: *vasárnap* ’Sunday’, *nap* ’day’

Sense 2: *napfényes* ’sunny’, *napsütötte* ’sunlit’
Nouns of sense 2: *terület* ’area’, *sziget* ’island’, *oldal* ’side’, *terasz* ’terrace’

[6] *napsütötte*: ’sunlit’, *napfényes*: ’sunny’, *napsütéses*: ’sunshiny’

It is noteworthy that these nouns form non-overlapping sets: nouns co-occurring with both senses were discarded. The resulting nominal sets were first checked for semantic coherence by automated means, then by meticulous lexicographic inspection.

4.6.2 Detecting the salient nominal contexts via binary trees (dendrograms)

In many cases, the set of retrieved nominal contexts was too large to interpret at a glance. In such instances, the word2vec representations (as described in Section 4.2) of the context nouns were clustered to yield salient semantic categories for the given subsense. The noun vectors were clustered using a hierarchical agglomerative algorithm with cosine distance and average linkage. For instance, *mindennapi* 'common' had been assigned two meanings: *hétköznapi* 'ordinary' and *mindennapos* 'everyday'. On one hand, the respective near-synonyms are rather enlightening with regard to the two senses of the adjective; one of them meaning 'normal' or 'ordinary', while the other refers to regular, everyday activities. However, we still need to know which nouns can induce the relevant meanings. For this purpose, dendrograms are created, yielding information that, for example, language-related things, such as *szóhasználat* 'word usage' and *nyelvhasználat* 'language use', along with *hős* 'hero', *figura* 'character', and *jelenet* 'scene', are more likely to be common or ordinary than periodical. On the other hand, *gyakorlás* 'practice' and *testmozgás* 'exercise' are regular, everyday activities and not necessarily common or ordinary ones. Therefore, the branches of the dendrogram indicate the semantic classes of nouns that the adjectival senses subcategorize.

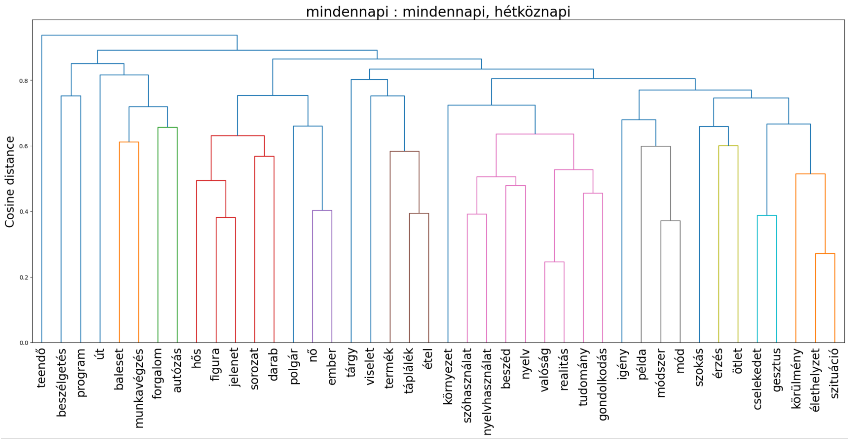


Figure 5: The dendrogram of the adjective *everyday* meaning 'ordinary' with its salient nominal contexts[7]

[7] *mindennapi*: 'everyday', *hétköznapi*: 'ordinary', *teendő*: 'task', *beszélgetés*: 'conversation', *program*: 'program', *út*: 'road', *baleset*: 'accident', *munkavégzés*: 'work', *forgalom*: 'traffic', *autózás*: 'driving', *hős*: 'hero', *figura*: 'figure', *jelenet*: 'scene', *sorozat*: 'series', *darab*: 'piece', *polgár*: 'citizen', *nő*: 'woman', *ember*: 'human', *tárgy*: 'object', *viselet*: 'clothing', *termék*: 'product', *táplálék*: 'nutrition', *étel*: 'food', *környezet*: 'environment', *szóhasználat*: 'word usage', *nyelvhasználat*: 'language usage', *beszéd*: 'speech', *nyelv*: 'language', *valóság*: 'reality', *realitás*: 'reality', *tudomány*: 'science', *gondolkodás*: 'thinking', *igény*: 'demand', *példa*: 'example', *módszer*: 'method', *mód*: 'way', *szokás*: 'habit', *érzés*: 'feeling', *ötlet*: 'idea', *cselekedet*: 'action', *gesztus*: 'gesture', *körülmény*: 'circumstance', *élethelyzet*: 'life situation', *szituáció*: 'situation'.

4.7 Representing semantic domains as connected components

A connected component is a subset of network nodes such that there is a *path* from each node in the subset to any other node in the same subset. As Zinoviev (2018: 129) notes, ”The property of connectedness is global and, while important for social and communication networks [...], may not be adequate for semantic, product, and other types of networks”. In the light of this assertion, it was quite unexpected that the connected components of the adjectival graph strictly corresponded to non-overlapping, semantically coherent components. The original adjectival graph, consisting of 10,153 adjectives, was dissected into 1,807 components using this technique, yielding a partition over 6,417 adjectives. Each component corresponds to a well-defined semantic domain. Note that one component of such networks is always a giant connected component (GCC), which comprises approximately one-third of the input adjectives (3,736) in this case. Unfortunately, the GCC merges multiple clear-cut semantic domains into one huge conglomerate, thus remaining uninformative about the meaning of the node adjectives as a whole.

Moreover, the adjectival graph components not only keep the various semantic domains separate but also reveal the relations between the inner node adjectives. These relations provide valuable information regarding polysemies and meaning shifts (Figure 6).

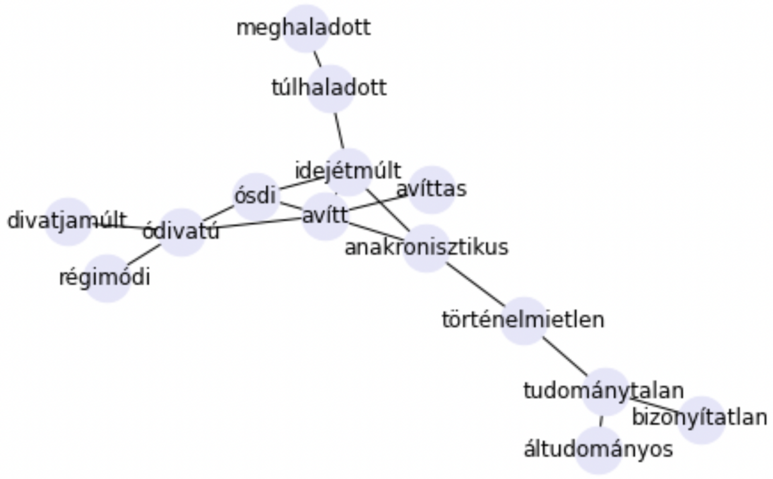


Figure 6: A connected component of the adjectival graph from the semantic domain *outdated*[8]

As Figure 6 indicates, there is an adjectival semantic field corresponding to *idejétmúlt* ’outdated’. There are three different edges from this node pointing to three different submeanings: *ósdi* ’shabby’, *túlhaladott* ’obsolete’, and *anakronisztikus* ’anachronistic’. The figure also shows that the next node after *anakronisztikus* is *történelmietlen* ’ahistorical’, which leads to *áltudományos* ’pseudoscientific’ in two steps.

[8] *divatjamúlt*: ’outdated’, *régimódi*: ’old-fashioned’, *ódivatú*: ’antiquated’, *ósdi*: ’shabby’, *meghaladott*: ’outmoded’, *túlhaladott*: ’obsolete’, *idejétmúlt*: ’outdated’, *avítt*: ’stale’, *avíttas*: ’musty’, *anakronisztikus*: ’anachronistic’, *történelmietlen*: ’ahistorical’, *tudománytalan*: ’unscientific’, *áltudományos*: ’pseudoscientific’, *bizonyítatlan*: ’unproven’

Consequently, connected components offer lexicographers a neatly categorized headword list, enabling a more thesaurus-like editing process, as opposed to the traditional alphabetical one, aligning with Stock (1984: 38).

5. Workflow: Unsupervised Extraction of Representations from Corpus Data

The methodology detailed here extends the fairly simple unsupervised graph-based approach described in Héja & Ligeti-Nagy (2022a,b), which was partially inspired by Ah-Pine & Jacquet (2009). Nevertheless, we introduced several significant changes. Firstly, we considered adjectives with lower frequency counts in the HNC to enhance coverage. Secondly, contrary to the previous experiment, we searched the entire HNC for salient noun candidates. Furthermore, our research didn't limit itself to polysemy: in addition to cliques, we generated and explored connected subgraphs from a lexicographic perspective. The key steps of the unsupervised graph induction process are recapped below:

1. Initially, we generated a weighted undirected graph, F, based on adjectival word2vec representations (cf. Subsection 4.2). In this graph, nodes represent adjectives, while edge weights indicate the strength of semantic similarity between every pair of adjectives. The weights were calculated using the standard cosine similarity measure. Importantly, the induced graph's undirectedness is guaranteed by the symmetric nature of cosine similarity.
2. Subsequently, we created an unweighted graph, G, by binarizing F. We used a K cut-off parameter to eliminate edges with low strength. Each edge weight w was set to 1 if $w \geq K$, and w was set to 0, if $w < K$. As a result, the graph G consists only of edges of the same strength ($w = 1$), where edges with $w = 0$ were omitted. During our experiments, K was set to 0.5 or 0.7.

However, in accordance with Zinoviev (2018: 80) we found that determining the optimal value for K presents a challenging task for future research. To illustrate the role of the K cut-off parameter, let us revisit the ego graph shown in Figure 1, where $K = 0.5$ was used. This graph consists of 15 nodes and 27 edges. By contrast, with $K = 0.7$, *érzékeny* becomes an isolated node, i.e., a subgraph containing no edges, since all adjacent nodes are connected with weights where $0.5 < w < 0.7$. Setting $K = 0.65$ results in an ego graph with 6 nodes and 5 edges (cf. Figure 7), indicating that a higher K cut-off value yields a smaller subgraph, both in terms of nodes and edges, likely possessing a less rich microstructure.

Moreover, the manual evaluation of the adjectival graph showed that the edge weights are characteristic of the semantic field to which the investigated adjectives belong. For example, a slicing threshold of $K = 0.9$ results in a graph where the components tend to correspond to referring adjectives with minimal lexical meaning components, such as names of days (cf. 8a), names of months (cf. 8b), or terminological expressions (e.g., *ősszájú* 'protostome', *újszájú* 'deuterostome').

As expressions with poor lexical meanings are less interesting from a lexicographic perspective, we must reduce the K cut-off value. As implied by Figure 7 and Figure 1, the

[9] *érzékeny*: 'sensitive', *sérülékeny*: 'vulnerable', *kényes*: 'delicate', *érzéketlen*: 'insensitive', *fogékony*: 'susceptible', *túlérzékeny*: 'hypersensitive'

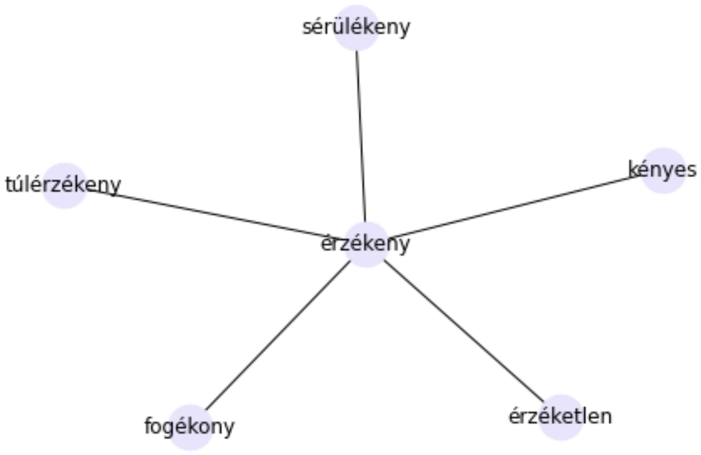
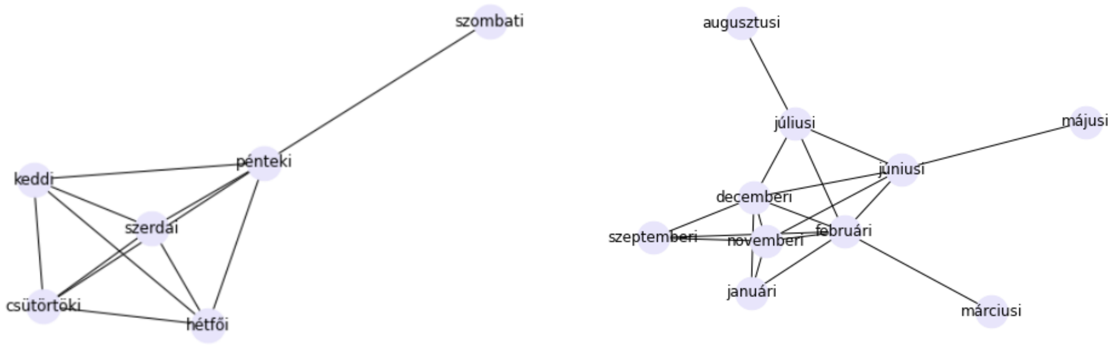


Figure 7: The ego graph of *érzékeny* 'sensitive' with $K = 0.65$ as cut-off parameter[9]

lower the K cut-off value, the richer the semantic content of the resulting microstructure candidate.

However, a lower K cut-off parameter may lead to more chaotic connected components and cliques, particularly in specific semantic domains. Thus, the precise parameter setting must be guided by meticulous lexicographic inspection, where both the semantic domains and the extent of the coverage need to be considered.



(a) Names of days[10]

(b) Names of months[11]

Figure 8: Graphs of referring expressions: $K \geq 0.9$

6. Lexicographic Perspective

In this section, we focus on the potential application of the proposed method for lexicographic purposes, specifically in the compilation of monolingual explanatory dictionaries. To this end, we will compare the automatically induced results with the micro- and

[11] *hétfői*: 'of-Monday', *keddi*: 'of-Tuesday', *szerdai*: 'of-Wednesday', *csütörtöki*: 'of-Thursday', *pénteki*: 'of-Friday', *szombati*: 'of-Saturday'

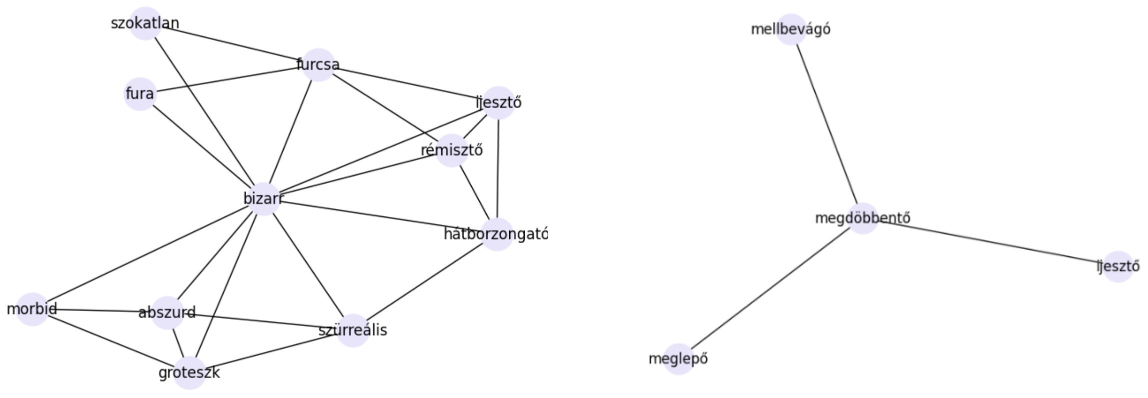
[11] *januári*: 'of-January', *februári*: 'of-February', *márciusi*: 'of-March', *májusi*: 'of-May', *júniusi*: 'of-June', *júliusi*: 'of-July', *augusztusi*: 'of-August', *szeptemberi*: 'of-September', *novemberi*: 'of-November', *decemberi*: 'of-December'

macrostructure of the EDHL. Unfortunately, EDHL does not offer any insight into the selection principles for its adjectival headword list. It merely states that the cataloged headwords, as curated by the editorial board, are “common, widely known, frequently used, and vital in communication and daily interaction in our language” (Bárczi & Országh, 1959–1962: VII).

From a lexicographic perspective, we tested five hypotheses:

1. The induced cliques can assist lexicographers in constructing the adjectival microstructure.
2. The automatically extracted and clustered nouns, modified by attributive adjectives and represented in the dendrograms, may aid lexicographers in supplementing the data used in EDHL for defining the adjectival microstructure.
3. The clusters of nouns might characterize the adjectival microstructure independently, indicating where distinctions in meaning need to be made, without relying on any pre-existing definitions.
4. We also investigated whether the automatically induced dendrograms can assist lexicographers in identifying inconsistencies in the EDHL, which may arise as a side effect of intuition-based methodologies.
5. The automatically extracted subgraphs, i.e., connected components, may also help in identifying missing headwords, thereby supplementing the macrostructure.

A detailed analysis of the ego graphs for 20 frequent adjectives, cut at a $K = 0.7$ threshold, revealed that in 8 instances, corresponding cliques included relevant adjectives not found in the EDHL. For example, the headword *bárgyú* ’silly’ does not include the subsense *bugyuta* ’foolish’. Similarly, the headword *bizarr* ’bizarre’ lacks *morbid* ’morbid’ and *szürreális* ’surreal’ (refer to Figure 9a), while the headword *megdöbbentő* ’shocking’ does not comprise the subsense *mellbevágó* ’gut-wrenching’ (see Figure 9b). When we lower the threshold to $K = 0.5$, the cliques become more granular, highlighting additional missing subsenses in the microstructure. For instance, the adjective *érzékeny* (refer to Figure 1) lacks references to subsenses *sebezhető* ’vulnerable’ and *problematikus* ’problematic’ in the EDHL.



(a) Subsenses of *bizarr* ’bizarre’[12]

(b) Subsenses of *megdöbbentő* ’shocking’[13]

Figure 9: Ego graphs compared to the microstructure of EDHL; $K \geq 0.7$

The second hypothesis proved to be completely correct based on the assessment of randomly selected dendrograms. This outcome isn't surprising, particularly considering that nodes near the terminals in the dendrogram align with cohesive, semantically related noun classes. For example, *fontos* 'important' can co-occur with military events such as *csata* 'battle', *hadművelet* 'military operation', and *küldetés* 'mission', or with various legal acts like *rendelet* 'order', *törvénytervezet* 'legislative proposal', *egyezmény* 'convention', and *szerződés* 'contract'. Similarly, *alacsony* 'low' often modifies financial terms related to money such as *áfakulcs* 'VAT rate', *alapanyagár* 'raw material price', *áramár* 'power tariff', and *adósságállomány* 'debt portfolio'.

The third hypothesis was partially validated. It was discovered that only nodes close to the terminals in the dendrogram—those with low cosine distances—indicate accurate meaning distinctions. For instance, as the red branch in Figure 10 demonstrates, military-related light weapons, including *kard* 'sword', *szablya* 'saber', *cirkáló* 'cruiser', *puska* 'rifle', and *ágyú* 'cannon', share the definition of 'a <smaller-sized weapon> that does not require much effort to carry, transport, and handle' in EDHL and group together convincingly. The data pertaining to the definition of 'a <military unit> equipped with such weapons' (*gyalogság* 'infantry', *tüzérség* 'artillery') is also well-differentiated.

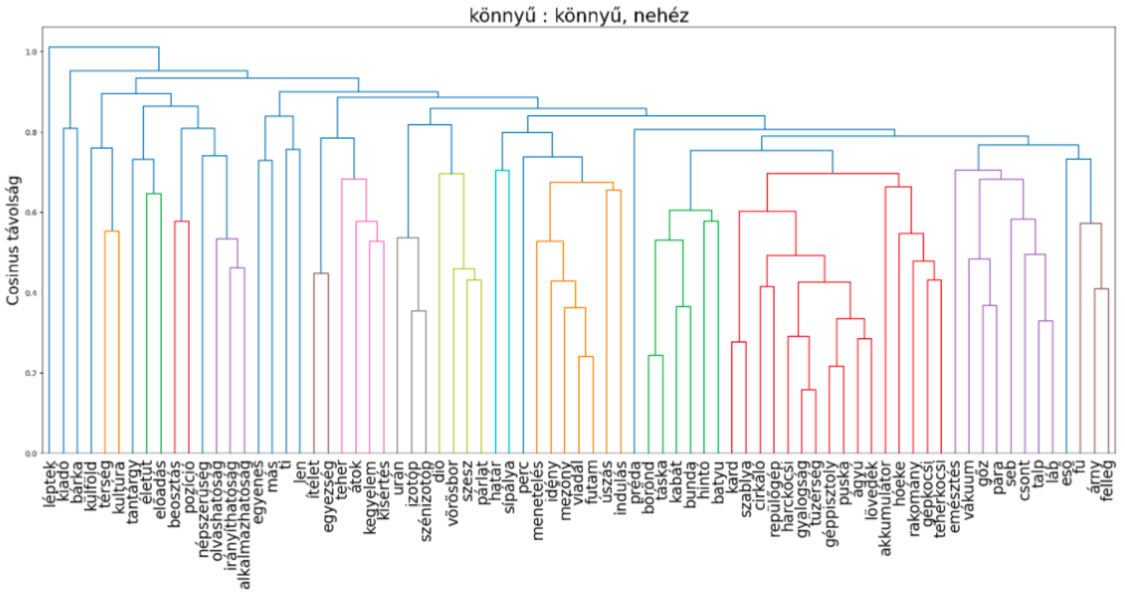


Figure 10: The dendrogram of the adjective *easy* with its antonym *difficult/heavy* and their salient nominal contexts[14]

[12] *bizarr*: 'bizarre', *szokatlan*: 'unusual', *fura*: 'strange', *furcsa*: 'peculiar', *ijesztő*: 'scary', *rémisztő*: 'frightening', *hátborzongató*: 'spine-chilling', *szürreális*: 'surreal', *abszurd*: 'absurd', *groteszk*: 'grotesque', *morbid*: 'morbid'

[13] *megdöbbentő*: 'shocking', *mellbevágó*: 'striking', *ijesztő*: 'scary', *meglepő*: 'surprising'

[14] *könnyű*: 'light' , *nehéz*: 'heavy' , *léptek*: 'steps' , *kiadó*: 'publisher' , *bárka*: 'boat' , *külföld*: 'foreign country' , *térség*: 'region' , *kultúra*: 'culture' , *tantárgy*: 'subject' , *életút*: 'life path' , *előadás*: 'lecture' , *beosztás*: 'schedule' , *pozíció*: 'position' , *népszerűség*: 'popularity' , *olvashatóság*: 'readability' , *irányíthatóság*: 'controllability' , *alkalmazhatóság*: 'applicability' , *egyenes*: 'straight' , *más*: 'other' , *ti*: 'you' , *len*: 'be' , *ítélet*: 'judgment' , *egyezség*: 'agreement' , *teher*: 'load' , *átok*: 'curse' , *kegyelem*: 'grace' , *kísértés*: 'temptation' , *urán*: 'uranium' , *izotóp*: 'isotope' , *szénizotóp*: 'carbon isotope' , *dió*: 'walnut' , *vörösbor*: 'red wine' , *szesz*: 'liquor' , *párlat*: 'spirit' , *határ*: 'border' , *sípálya*: 'ski slope'

Comparing manually the automatically induced results with the microstructures in EDHL revealed that without adequate context, it can often be challenging to determine the appropriate placement of the adjective-noun construction within the microstructure. While this doesn’t necessarily imply overlapping sense distinctions in EDHL’s microstructure, this potentiality should be considered in future evaluations.

In line with this, we encountered several issues during the disambiguation of the attributive adjectives *fontos* ‘important’ and *jelentős* ‘significant’ in EDHL due to strongly overlapping definitions in the microstructures. The correct interpretation of *fontos* ‘important’ was particularly problematic when the modified nouns were one of the following: *munkatárs* ’colleague’, *tisztség* ’position’, *bizottság* ’committee’, *jogintézmény* ’legal institution’, *rendelet* ’order’, *törvénytervezet* ’legislative proposal’, *egyezmény* ’convention’, *szerződés* ’contract’, etc. Indeed, the following two senses of *fontos* appear to overlap:

1. <jelentőségénél fogva különös gondot, figyelmet érdemlő, jelentős, lényeges>[15]
2. <vmely cél elérésében, ill. a gyakorlati élet vmely területén jelentős szerepet betöltő, alig nélkülözhető>[16]

Determining whether *jelentős diadal* ‘significant triumph’, *jelentős térnyerés* ‘significant expansion’, *jelentős fölény* ‘significant advantage’, *jelentős emberveszteség* ‘significant loss of life’, *jelentős jövedelemforrás* ‘significant source of income’, *jelentős kiegészítés* ‘significant supplement’, *jelentős ismeret* ‘significant knowledge’ can be subsumed under multiple senses in EDHL, such as <‘very important, of great significance’>, <‘above average, considerable, significant, noteworthy’>, or <‘playing an important role; significant, influential as a result of its effects or outcomes’> is also challenging. The overlapping senses suggest that providing more textual context would probably be insufficient to enable the lexicographer to find the correct meaning in this case.

Regarding the fifth hypothesis, the comparison of EDHL and the automatically retrieved semantically related adjectives, extracted via the connected graph components, was rather telling. For example, the graph-based algorithm cataloged 90 adjectives referring to quantities from the training corpus, of which only 8 are listed in EDHL (*ujjnyi* ’one/two inch’ or ’a finger-sized’, *arasznyi* ’5-6 inches’, *körömnyi* ’nail-sized’, *késhegynyi* ’knife edge-sized’, *tenyérnyi* ’palm-sized’, *mázsás* ’two hundred pounds heavy’, *mérföldes ’mile-long’*, *púpozott* ’rounded’ as in a ’rounded tablespoon of sg.’). Regrettably, important adjectives are missing from the headword list: the corpus data clearly indicate that *gyűszűnyi* ’thimble-sized’ and *ökölnyi* ’fist-sized’ should form headwords on their own, but they are only included in the microstructure of the corresponding nominal headwords (e.g., *gyűszű* ‘thimble’ and *ököl* ‘fist’).

, *perc*: ’minute’ , *menetelés*: ’marching’ , *idény*: ’season’ , *mezőny*: ’field’ , *viadal*: ’tournament’ , *futam*: ’race’ , *úszás*: ’swimming’ , *indulás*: ’departure’ , *préda*: ’prey’ , *bőrönd*: ’suitcase’ , *táska*: ’bag’ , *kabát*: ’coat’ , *bunda*: ’fur coat’ , *hintó*: ’carriage’ , *batyu*: ’sack’ , *kard*: ’sword’ , *szablya*: ’saber’ , *cirkáló*: ’cruiser’ , *repülőgép*: ’airplane’ , *harckocsi*: ’tank’ , *gyalogság*: ’infantry’ , *tüzérség*: ’artillery’, *géppisztoly*: ’machine gun’, *puska*: ’rifle’, *ágyú*: ’cannon’, *lövedék*: ’bullet’, *akkumulátor*: ’battery’, *hóeke*: ’snowplow’, *rakomány*: ’cargo’, *gépkocsi*: ’car’, *teherkocsi*: ’truck’, *emésztés*: ’digestion’, *vákuum*: ’vacuum’, *gőz*: ’steam’, *pára*: ’vapor’, *seb*: ’wound’, *csont*: ’bone’, *talp*: ’sole’, *láb*: ’foot’, *eső*: ’rain’, *fű*: ’grass’, *árny*: ’shadow’, *felleg*: ’cloud’.

[15] ’By virtue of its significance, it deserves special care, attention, and is significant and essential.’

[16] ’Playing a significant role in achieving a particular goal or in a certain area of practical life; being scarcely dispensable.’

Apart from the insufficient coverage of certain semantic fields, additional inconsistencies emerged during the random testing of certain headwords. For instance, both *kisbirtokos* 'smallholder' and *középbirtokos* 'medium-sized landowner' (lit. mediumholder) appeared in the headword list in their adjectival forms. However, the adjectival form *nagybirtokos* 'large landowner' (lit. largeholder) was absent: only the nominal form was cataloged as a headword: *'<Feudális v. kapitalista rendszerben> nagybirtokkal rendelkező, nagybirtokán mezőgazdasági (és állattenyésztési) munkát végeztető és dolgozóit kizsákmányoló személy.|| a. jelzői használat(ban) Ilyen személyekből álló <csoport>. Nagybirtokos arisztokrácia, család, kaszt.'*[17]. Another inconsistency is the absence of the adjective *kisméretű* 'small-sized', which should be a headword given that *nagyméretű* 'great-sized' is part of the headword list, and that *kisméretű* is used rather frequently in the definitions of other headwords.

7. Future work

An unsupervised graph-based methodology is described in this paper. The aim is to support the work of expert lexicographers in compiling the macro- and microstructure of a monolingual explanatory dictionary for Hungarian. Although the proposed framework seems promising, there are multiple issues that need to be addressed to fully realize the method's potential.

Most importantly, the optimal value of the slicing parameter K should be set so that the automatically obtained results best suit the specific objectives of the lexicographers. Determining the optimal parameter setting requires robust collaboration between lexicographers and computational linguists for several reasons.

First, the selection principles of the adjectives are significantly determined by the purpose and target audience of the dictionary. Secondly, further lexicographic inspection is needed to set the K cut-off parameter, which depends not only on the network topology and weight distribution but also on the specific semantic classes of the adjectives.

Thirdly, the editing principles of the planned dictionary should be explicitly stated: those morphologically or semantically productive cases that, due to their productivity, should not form part of the dictionary, should be cataloged. As the randomly sampled lexicographic observations indicated, the described algorithm seems to be useful for these purposes as well. Various types of subgraphs may yield information both on the morphology-semantics interface and on the systematic subcategorization patterns of adjectives. Again, careful lexicographic work is indispensable to compile a comprehensive list of these attributes.

Finally, the prototype algorithm should be implemented as a software tool to enhance the efficiency of lexicographers' work.

8. References

Adamska-Sałaciak, A. (2006). *Meaning and the Bilingual Dictionary. The Case of English and Polish. (Polish Studies in English Language and Literature 18).* Frankfurt am Main:

[17] '<In the era of a feudal or a capitalist system> a person who owns a large estate, employs agricultural (and livestock) workers, and exploits them.|| a. (attributive use) A <group> formed by such persons. Large landowner aristocracy, family, caste.'

Peter Lang.

Ah-Pine, J. & Jacquet, G. (2009). Clique-Based Clustering for Improving Named Entity Recognition Systems. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 51–59. URL https://aclanthology.org/E09-1007.

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Burnard, L. (ed.) (2007). *Reference Guide for the British National Corpus (XML Edition)*. URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Bárczi, G. & Országh, L. (eds.) (1959–1962). EDHL = *A magyar nyelv értelmező szótára I–VII. [The Explanatory Dictionary of the Hungarian Language]*. Budapest: Akadémiai Kiadó.

Camacho-Collados, J. & Pilehvar, M.T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. URL https://arxiv.org/abs/1805.04032.

Frege, G. (1892). Uber Sinn und Bedeutung. In M. Textor (ed.) *Funktion - Begriff - Bedeutung*, volume 4 of *Sammlung Philosophie*. Göttingen: Vandenhoeck & Ruprecht.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.

Hanks, P. (2010). Compiling a Monolingual Dictionary for Native Speakers. *Lexikos*, 20, pp. 580–598.

Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25, pp. 398–436.

Héja, E. & Ligeti-Nagy, N. (2022a). A Clique-based Graphical Approach to Detect Interpretable Adjectival Senses in Hungarian. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*. Gyeongju, Republic of Korea: Association for Computational Linguistics, pp. 35–43. URL https://aclanthology.org/2022.textgraphs-1.4.

Héja, E. & Ligeti-Nagy, N. (2022b). A proof-of-concept meaning discrimination experiment to compile a word-in-context dataset for adjectives – A graph-based distributional approach. *Acta Linguistica Academica*, 69(4), pp. 521 – 548. URL https://akjournals.com/view/journals/2062/69/4/article-p521.xml.

Ittzés, N., et al. (ed.) (2006–2021). *CDH = A magyar nyelv nagyszótára I-VIII. [Comprehensive Dictionary of Hungarian]*. Budapest: Nyelvtudományi Kutatóközpont.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P. & Suchomel, V. (2013). The TenTen corpus family. In *7th international corpus linguistics conference CL*. Lancaster University, pp. 125–127.

Juhász, J., Szőke, I., O. Nagy, G. & Kovalovszky, M. (eds.) (1972). ÉKSz^1 = *Magyar értelmező kéziszótár. [Concise Hungarian Explanatory Dictionary]*. Budapest: Akadémiai Kiadó.

Kiefer, F. (2003). How much information do adjectives need in the lexicon? In *Igék, főnevek, melléknevek [Verbs, nouns, adjectives]*. Tinta Könyvkiadó, pp. 32–43.

Kiefer, F. (2008). A melléknevek szótári ábrázolásáról. In *Stukturális magyar nyelvtan 4. A szótár szerkezete*. Akadémiai Kiadó, pp. 505–538.

Kipper-Schuler, K. (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari & et al. (eds.) *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 1848–1854.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press.

Lipp, V. & Simon, L. (2021). Towards a new monolingual Hungarian explanatory dictionary: overview of the Hungarian explanatory dictionaries. *Studia Lexicographica*, 15, pp. 83–96.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. URL https://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Moon, R. (1987). The Analysis of Meaning. In J. Sinclair (ed.) *Looking Up.* pp. 86–103.

Nemeskey, D.M. (2020). *Natural Language Processing Methods for Language Modeling.* Ph.D. thesis, Eötvös Loránd University.

Oravecz, C., Váradi, T. & Sass, B. (2014). The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 1719–1723. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf.

Országh, L. (1953). A magyar nyelv új szótáráról. [On the new dictionary of the Hungarian Language]. *Magyar Nyelvőr*, 77(5–6), pp. 387–407.

Ploux, S. & Victorri, B. (1998). Construction d'espaces sémantiques a l'aide de dictionnaires de synonymes. *Traitement automatique des langues*, 1(39), pp. 146–162.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, MA: MIT Press.

Pusztai, F. & Csábi, S. (eds.) (2003). ÉKSz^2 = *Magyar értelmező kéziszótár. [Concise Hungarian Explanatory Dictionary].* Budapest: Akadémiai Kiadó.

Rehurek, R. & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Sass, B., Váradi, T., Pajzs, J. & Kiss, M. (2010). *Magyar igei szerkezetek. A leggyakoribb vonzatok és szókapcsolatok szótára.* Budapest: Tinta Könyvkiadó.

Stock, P.F. (1984). Polysemy. In R.R.K. Hartman (ed.) *LEXeter '83: Proceedings. Papers from the International Conference on Lexicography at Exeter, 9–12 September 1983.* Tübingen: Max Niemeyer, pp. 131–140.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge: Cambridge University Press.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas.* pp. 385–389.

Véronis, J. (2003). Sense tagging: does it make sense? In A. Wilson, P. Rayson & T. McEnery (eds.) *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech.* Frankfurt: Peter Lang.

Zinoviev, D. (2018). *Complex Network Analysis in Python: Recognize - Construct - Visualize - Analyze - Interpret.* The Pragmatic Bookshelf.