Enikő Héja, Kata Gábor, László Simon, and Veronika Lipp

# GRAPH-BASED DETECTION OF HUNGARIAN ADJECTIVAL MEANING STRUCTURES VIA MONOLINGUAL STATIC EMBEDDINGS



**Abstract** The paper details the current state of an ongoing collaboration between Hungarian lexicographers and computational linguists. Our goal is to provide a comprehensive and consistent description of Hungarian adjectives, benefiting lexical semantics, lexicography and NLP. This thread of research focuses on identifying systematic semantic patterns of Hungarian adjectives and their typical subcategorization frames, with a particular emphasis on polysemous meanings. The proposed methodology is entirely unsupervised, reducing reliance on human intuition. It is based on a graph representation derived from adjectival static embeddings. The algorithm models adjectival semantic domains by specific subgraphs, namely, connected graph components. In the next step, potential subcategorization frames for the detected adjectival semantic domains, so called meaning structures, are also derived from corpus data. Then, a sample of the meaning structures is compared to the entries of the Concise Dictionary of Hungarian, evaluating the pros and cons of the proposed algorithm. Finally, as a further improvement, the automatically derived subcategorization frames were generalized.

**Keywords** automatic sense induction; monolingual lexicography; polysemy; unsupervised graph-based approach; adjectives

## 1. Introduction

### 1.1 Objectives – Lexicography, Lexical Semantics and NLP

This ongoing research aims to develop a consistent, data-driven method for describing Hungarian adjectives, which could benefit lexicography, lexical semantics, and NLP most importantly by providing adjectival sense distinctions. Here the sense distinctions are assigned to adjectival semantic domains rather than to specific adjectives, allowing for the identification of systematic semantic shifts. The data-driven nature empowers us to process a wide variety of adjectives, leading to a set of semantic features based on which more motivated sense distinctions can be made, thus resulting in a more consistent characterization of adjectival senses.

From a lexicographic perspective, the method aids in producing a consistent headword list by identifying common productive derivational processes in the formation of adjectives. This ensures that only those headwords not covered by explicit rules are included in the dictionary, which helps to maintain a consistent adjectival macrostructure.

The algorithm also supports the comprehensive and coherent lexical semantic description of Hungarian adjectives, which is currently lacking, although there are some lexical semantic investigations on Hungarian adjectives (Kiefer, 2008).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Moreover, by extracting explicit contextual clues – semantic features – for sense distinctions, our approach is expected to improve the quality of benchmark datasets to evaluate semantic knowledge in language models, particularly by increasing inter-annotator agreement in word sense disambiguation tasks (Pilehvar & Camacho-Collados, 2019).

## 1.2 The Scope of the Research

Although our previous investigation has shown that the proposed representation is able to filter homonym-candidates as well (cf. 2.3), the present research is focusing on *polysemous sense* distinctions. According to Haber & Poesio, 2024, p. 3), polysemy is "far less well understood" than homonymy and "far less homogeneous than assumed in earlier literature." To limit the scope of our investigation further, the present paper focuses on systematic meaning shifts, which operate on adjectival semantic domains rather than on a single adjective. These are sometimes referred to as idiosyncratic polysemy in the literature. The proposed algorithm models the definition of Apresjan (1974, p. 16) of regular polysemy. According to him, "Polysemy of a word A with the meaning $a_i$ and $a_j$ is called regular if [...] there exists at least one other word *B* with the meaning $b_i$ and $b_j$, which are semantically distinguished from each other in exactly the same way as $a_i$ and $a_j$ and if $a_i$ and $b_i$, $a_j$ and $b_j$ are nonsynonymous."

| | | | |
|---|---|---|---|
| bencés 'Benedictine' | 182 | 98.14 | 87.6 |
| ciszterci 'Cistercian' | 49.29 | 26.43 | 331 |
| ciszterci 'Cistercian (fem)' | 20.83 | 8.5 | |
| dominikánus 'Dominican' | 32 | 11.83 | |
| domonkosrendi 'Dominican' | 1.75 | | |
| ferences 'Franciscan' | 233.67 | 192.71 | 76.75 |
| ferencrendi 'Franciscan' | 4.67 | 6.75 | |
| jezsuita 'Jesuit' | 49.8 | 172.86 | 88.6 |
| karmelita 'Carmelite' | 48.33 | 12.4 | |
| piarista 'Piarist' | 49.67 | 94.6 | 193 |
| premontrei 'Premonstratensian' | 36.88 | 20.17 | 73.5 |
| | rendház 'monastery' | szerzetes 'monk' | iskola 'school' |
| | monostor 'monastery' | apát 'abbott' | egyetem 'university' |
| | kolostor 'convent' | pap 'priest' | gimnázium 'high school' |

**Fig. 1:** Part of the meaning structure of the lexical set 'monastic orders'

| | | | | |
|---|---|---|---|---|
| esztelen 'senseless' | 3.20 | 26.50 | 13.00 | 7.22 |
| fékevesztett 'uncontrollable' | 2.00 | 1.50 | 6.83 | |
| fékezhetetlen 'unstoppable' | 2.33 | 2.00 | 6.62 | |
| féktelen 'unrestrained' | 21.60 | 12.00 | 7.50 | 21.64 |
| oktalan 'unwise' | 3.25 | 1.50 | 3.75 | 12.67 |
| szertelen 'excessive' | 6.00 | 2.00 | 9.88 | |
| zabolátlan 'unbridled' | 1.33 | 9.00 | | |
| | nevetés 'laughter' | költekezés 'spending' | rablás 'robbery' | vágy 'desire' |
| | száguldás 'speeding' | pazarlás 'wastefulness' | pusztítás 'destruction' | harag 'anger' |
| | zabálás 'binge eating' | ámokfutás 'rampage' | | düh 'rage' |
| | ivászat 'drinking' | | | gyűlölet 'hatred' |

**Fig. 2:** Part of the meaning structure of the near-synonyms of *féktelen* 'unrestrained'

As the automatically extracted matrices (Figure 1 and Figure 2) illustrate, according to this definition the adjectives in the semantic domains 'monastic orders' and 'unrestrained' equally exhibit regular polysemies characterized by the nominal sets at the end of each column, where each column corresponds to a specific sense candidate. For instance, let *A* be the word *bencés* 'Benedictine' and *B* be 'Jesuit'. As indicated by the nominal sets at the end of the columns in Figure 1, both 'Benedictine' and 'Jesuit' yield the senses of {'building', 'monastery', 'convent'} and {'clergy': 'monk', 'abbot', 'priest'} conforming to the criterion of regular polysemy as defined by Apresjan. However, we not only indicate whether a word *A* appears with one sense or not, but also assess the frequency of the given sense based on corpus data. These estimates appear in the cells of Figures 1. and 2. The algorithm is described in more detail in Section 2.

Following Atkins and Rundell (2008), we focus on two types of systematic meaning shifts: common word usage scenarios in lexical sets (semantic frames) and specific meaning shifts (systematic polysemy). Both types of meaning variations are important from a lexicographic perspective: lexical sets are defined by semantic frames linked to common sense knowledge (Figure 1), while near-synonyms are defined by systematic polysemies linked to lexical knowledge (Figure 2). The main advantage of our method is that it automatically extracts interpretable meaning patterns, ensuring consistent treatment of systematic sense variations across the vocabulary.

The research examines how effectively the proposed method aids in compiling the adjectival section of a monolingual Hungarian general purpose dictionary both at the levels of the macrostructure and the microstructure.

The paper is organised as follows: Section 2 introduces the workflow, Section 3 provides a brief overview on the Concise Dictionary of Hungarian, Section 4 evaluates the generated meaning structures from a lexicographic perspective, and Section 5 explores the possibility to generalize over the automatically attained adjectival subcategorization patterns.

## 2. Workflow

Our methodology extends the unsupervised graph-based approach described in Héja et al. (2022a; 2022b, and 2023), which was partially inspired by Ah-Pine and Jacquet (2009).

### 2.1 Extraction of Adjectival Semantic Domains From Corpus Data

In brief, adjectival meaning structure candidates were obtained from Hungarian monolingual data via the extraction of specific subgraphs from a graph of Hungarian adjectival word2vec embeddings.

Static embeddings, although more interpretable than LLMs, face criticism for potentially merging different senses (Camacho-Collados & Pilehvar, 2018). The conversion of embeddings into graphs is primarily aimed at addressing this issue of meaning conflation. Although the present paper focuses on polysemy, we claim that our approach is capable of filtering homonym-candidates as well, based on the presupposition that the various senses of a homonymous word form only accidentally coincide.

On the other hand, the graph representation has proven capable of capturing systematic semantic phenomena too, likely due to the tendency of words with similar usage scenarios or sense alternations to appear in more similar contexts compared to words with non-systematic sense relations.

This paper focuses on connected graph components, where any node can be reached from any other node by traversing edges. We found that connected components of the adjectival graph G correspond to clear-cut semantic domains.

### 2.2 Extraction of Connected Components From the Adjectival Graph

Let us briefly recap the technical details of the workflow:

1. 300-dimension static word2vec (Mikolov et al., 2013) representations were trained (Rehurek & Sojka, 2011) on the basis of 170 m sentences, part of the Hungarian Webcorpus 2.0 (Nemeskey, 2020).

2. An $A$ adjacency matrix of the size $N$ X $N$ - $N$ is the size of the adjectival vocabulary - was created where the $<i, j>$ cell of $A$ matrix was calculated using the usual cosine similarity between the vector representations of $adj_i$ and $adj_j$.

3. The weighted undirected graph $W$ was generated based on $A$ adjacency matrix. Graph $W$ is made up of $N$ nodes and all nodes are interconnected in $W$ with weights corresponding to the strength of semantic similarity between every pair of adjectives. Importantly, the induced graph's undirectedness is guaranteed by the symmetric nature of cosine similarity.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

4. Subsequently, an unweighted graph $G$ was created by binarizing $W$. A $K$ cut-off parameter was used 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, graph G consists only of edges of the same strength ($w$=1); edges with $w = 0$ were omitted. Based on the results of manual evaluation $K$ was set to 0.7 for the present experiments.

5. Finally, the connected graph components were extracted from graph $G$.

6. Below, we present two examples of the automatically extracted connected adjectival graph components: the first covers the lexical set of monastic orders, while the second contains adjectives from the semantic domain of *féktelen* 'unrestrained'.

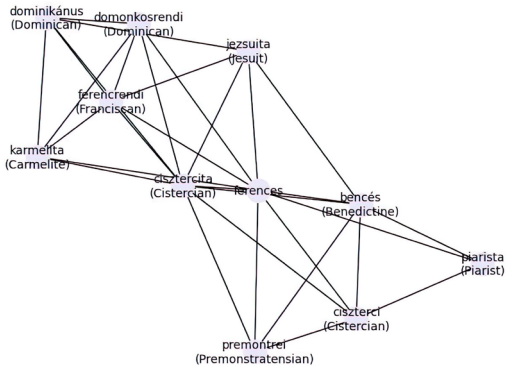


**Fig. 3:** Connected component of 'monastic orders'

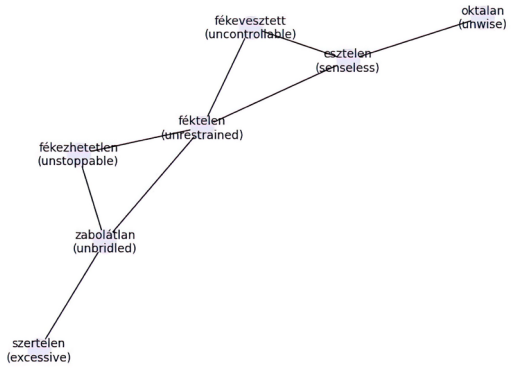


**Fig. 4:** Connected component for near-synonyms of 'unrestrained'

## 2.3 Characterization of the Results

Using a $K$=0.7 cut-off parameter, the 53,490 input adjectives were divided into 1,750 connected graph components, each with at least 2 nodes, encompassing 11,474 adjectives. Of these, 1,748 components were entirely meaningful. Manual evaluation of components resulting from higher K cut-off values, and manual inspection of so called ego-graphs (a sub-graph made up of an ego-node and its adjacent nodes) confirmed our hypothesis, namely, that the two largest components (4505 and 669 adjectives) resulted from merging smaller meaningful components. The third largest component (355 adjectives) comprised personal names with derivational suffixes *-i* and *-s*. The fourth largest component (74 adjectives) included names derived from Hungarian cities with the *-i* suffix. Another notable component (43 adjectives) represented the semantic field of ages. Note that only 11,474 of the original 53,490 adjectives formed part of connected components with at least two nodes. This raises the question of what happened to the remaining cc. 40,000 adjectives. The parameter setting yielded 42,016 isolated nodes: these were deemed dissimilar to everything else by the algorithm. The manual evaluation of the 30 most frequent isolated adjectives showed that homonyms tend to end up as isolated nodes. This observation perfectly fits Apresjan's definition of homonymy (Apresjan, 1974, p. 11), which describes

it as "a purely external coincidence of two or more words whose meanings have nothing in common." Since a homonym has two (or more) unrelated meanings, the different randomly related sets of contexts in which it appears are combined during the training process. This results in a unique vector that is specific to the homonym itself, reflecting its various meanings.

The detailed investigation of the isolated nodes forms part of our future research. As Figure 5. illustrates, the connected component covering the semantic domain of country names does not contain *lett* ('Latvian', also meaning 'became'), *észt* ('Estonian' and the accusative form of 'wit') and *ír* ('Irish' and 'writes'), which ended up as isolated nodes.

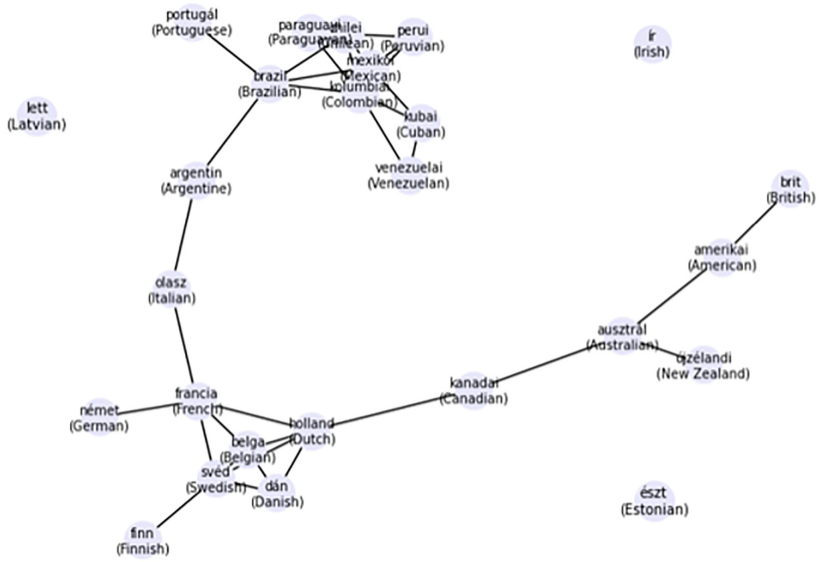


**Fig. 5:** The graph of country names with the three homonymic country names as isolated nodes

## 2.4 Creating the Meaning Structures

This study explores how automatically retrieved adjectival semantic domains can help characterise the meanings of adjectives, specifically identifying potential meaning structures of lexical sets or sets of near-synonyms (see Figures 1 and 2).

The workflow involves the following steps:

1. Manually select lexicographically interesting adjectival domains from the list of connected graph components (Figures 3 and 4).

2. Extract nouns that co-occur with any adjectives in the selected domains above a frequency threshold in the Hungarian National Corpus (Váradi, 2002) and cluster them using a hierarchical clustering algorithm (average linkage, cosine distance). The hierarchic clusters were flattened based on a cophenetic distance threshold within the clusters. The main advantage of hierarchical clustering is that the exact number of senses can vary, therefore,

do not have to be determined beforehand. The parameters were set based on the results of manual evaluation. The resulting nominal clusters serve as semantic feature candidates to characterise senses in the meaning structure of the adjectival semantic domains (columns in Figures 1 and 2).

3. Assess how well a nominal semantic category characterises a specific adjective by calculating the simple arithmetic mean of the number of adjective-noun co-occurrences for each nominal cluster (cells in Figures 1 and 2). This indicates the strength of association between each adjective and the corresponding semantic feature.

4. Automatically generate meaning structure matrices for the selected semantic domains (Figures 1 and 2).

## 3. The Concise Dictionary of Hungarian (CDH)

The Concise Dictionary of Hungarian (Pusztai, 2003) is a monolingual general-purpose dictionary containing 73,550 headwords, covering the vocabulary of both literary and colloquial Hungarian. To assess the extent to which the described method may contribute to the dictionary editing process, we first give a brief overview on CDH from multiple perspectives.

First, as for the adjectival vocabulary, the CDH comprises cc. 7,700 adjectival entries, comparable in size to that of the adjectival vocabulary assigned to connected components (11,474). However, the situation is not as clear as it may seem at first, since the part-of-speech categorization is not always consistent in CDH. In the case of adjectives the demarcation of nouns, adjectives and adjectival participles may raise some problems. We think that connected components can help in a more consistent PoS characterization, as due to the distributional similarity, adjectives belonging to the same connected component probably belong to the same PoS category/ies. For instance, although *megfelelő* 'appropriate' and *kellő* 'adequate' should both be considered adjectives, CDH assigns them falsely to different PoS-es.

The second problem is that although the CDH explicitly states (Pusztai, 2003, XII, XVII) that common derivational suffixes are enlisted and characterised as usual headwords, and thus, derivational forms are only enlisted if the suffixed form carries a new, distinct dictionary meaning compared to the base word, this principle is not adhered to consistently. At the same time, the semantic description of the derivational suffixes seems to be too loose and incomplete. Connected graph components offer a solution to provide a more explicit and clear-cut definition of derivational suffixes. In many cases, the extracted adjectives not only form tight semantic classes but also result from the same derivational process. This allows for an explicit and clear semantic description of the derivational process at both the input and output levels. For example, rather than using the vague and general definition of the derivational suffix *-s* in CDH as '<indicates that someone or something has someone or something, or that there is something in or on them>', it would be much clearer to list specific cases of the derivational process describing precise and explicit semantic relations.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

For instance, the semantic relations represented by the *-s* derivational suffix may differ in the contexts of <'course', 'meal'> and <'lane', 'road'>. This is evident in terms like *egyfogásos* 'one-course', *kétfogásos* 'two-course', or *háromfogásos* 'three-course' meals and *egysávos* 'one-lane' or *kétsávos* 'two-lane' roads.

The loose definition of derivational suffixes seems to cause problems at the level of the microstructure as well. For instance, although the derivational suffix *-s* is characterised as a headword, CDH comprises the transparently formed headword *kalapos* 'hatted': (adj) <*kalapot viselő* 'wearing a hat'>; <*aminek kalapja van* 'having a hat'>; <*amin kalapo(ka)t tartanak* 'on which hat(s) are kept'>. However, it is even more interesting that the third sub-sense of this meaning structure is basically non-existent in the light of synchronous corpus data.

Finally, the headword list is not consistent in the CDH either: for example, although in accordance with the edition principles the *-i* derivational suffix is neatly characterized as a headword, there are still cc. 1300 adjectival headwords with the *-i* suffix, several of them being completely transparent constructions.

## 4. Evaluating the Meaning Structures from a Lexicographic Point of View

During the evaluation phase, cc. 200 manually selected adjectives were examined representing various semantic relations, e.g., antonyms, synonyms, co-hyponyms. The automatically induced results were compared to the micro- and macrostructure of CDH.

### 4.1 Macrostructure

The analysis of the generated connected components revealed several further inconsistencies in the headword selection:

- Connected component I.: *csillagfényes* 'starlit', *csillagtalan* 'starless', *holdfényes* 'moonlit', *holdvilágos* 'moonlit', *teliholdas* 'full-mooned': neither *holdfényes* nor *teliholdas* were included as headwords in CDH.
- Connected component II.: *nullszaldós* 'break-even', *nyereséges* 'profitable', *rentábilis* 'profitable', *ráfizetéses* 'unprofitable', *veszteséges* 'loss-making': *ráfizetéses* was not included as a headword.
- Connected component III.: *gyakorlatlan* 'inexperienced', *gyakorlott* 'experienced', *rutinos* 'experienced', rutintalan 'inexperienced', *tapasztalatlan* 'inexperienced', *zöldfülű* 'rookie': *rutintalan* was not included as a headword.
- Connected component IV.: *bencés* 'Benedictine', *ciszterci* 'Cistercian', *cisztercita* 'Cistercian', *dominikánus* 'Dominican', *domonkosrend*i 'Dominican Order', *ferences* 'Franciscan', *ferencrendi* 'Franciscan Order', *jezsuita* 'Jesuit', *karmelita* 'Carmelite', *pannonhalmi* 'from Pannonhalma', *piarista* 'Piarist', *premontrei* 'Premonstratensian', *sarutlan* 'discalced': *domonkosrendi*,

*pannonhalmi*, and *ferencrendi* were not included as headwords. On the other hand, being only a collocate word, *sarutlan* should not be part of the connected component.

The method also works well with adjectives formed by productive derivation. For instance, adjectives with the derivational suffix *-nyi* form a semantically coherent adjectival class referring to distances, such as:

- Connected component V. *karnyújtásnyi* 'within arm's reach', *kéznyújtásnyi* 'within hand's reach', *kőhajításnyi* 'within a stone's throw', *nyíllövésnyi* 'within an arrow's shot', *puskalövésnyi* 'within a gunshot': based on Section 3, it is not obvious that these expressions should appear as headwords in CDH. But if one of them is included, every member of the component should be included, which is obviously not the case, as *kéznyújtásnyi* and *puskalövésnyi* are missing from CDH.

## 4.2 Microstructure

The method assists lexicographers by automatically finding the most important sub-senses, that is, meaning structures and the most frequent word combinations, which may serve as good examples in the dictionary entries. Referring back to Figre 1, we may say that the automatic methods reveal multiple subsenses (1. building, 2. position in the order, 3. educational institution).

Unfortunately, while the meaning structures demonstrate the potential to present these sub-senses coherently across the semantic domain, the senses are listed in the CDH inconsistently. For instance, when examining the entries for *sminkes* 'makeup artist', *pedikűrös* 'pedicurist' and *kozmetikus* 'beautician', we found that each is treated differently. *Sminkes* 'Makeup artist': I.(adj) <Related to or pertaining to makeup> II. (n) <A professional who performs makeup application in theatre or film studios.>; *Pedikűrös* 'Pedicurist' (n) <A foot care specialist.>; *Kozmetikus* 'Beautician' (adj and n) <A professional engaged in beauty care, particularly facial care.>

The PoS categories of all three entries should be uniform (both adjective and noun). Due to the strong semantic similarity, the definitions should also be standardised, for instance as follows: <A professional engaged in makeup application/foot care/facial care.>

Giving a closer look to another connected component comprising the adjectives *biszexuális* 'bisexual', *heteroszexuális* 'heterosexual', *homoszexuális* 'homosexual' and *leszbikus* 'lesbian', additional inconsistencies also were encountered in CDH: in three cases (*biszexuális*, *heteroszexuális* and *leszbikus*) the adjectival and nominal PoS categories are listed in a single definition, while in the remaining case (*homoszexuális*), the adjectival and nominal senses were told apart into separate definitions.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

| | | | | | | |
|---|---|---|---|---|---|---|
| biszexuális 'bisexual' | | | 13.2 | | | |
| heteroszexuális 'heterosexual' | | 9 | 21.57 | 3 | | |
| homoszexuális 'homosexual' | 14 | 16.4 | 54 | 22.67 | | 49 |
| leszbikus 'lesbian' | 16 | 24.6 | 47.5 | | 13 | |
| | fesztivál 'festival' | szex 'sex' | anya 'mother' | megrontás 'molestation' | érintés 'touch' | propaganda 'propaganda' |
| | felvonulás 'parade' | házasság 'marriage' | katona 'soldier' | zaklatás 'harassment' | | |
| | | szerelem 'love' | férfi 'man' | erőszak 'violence' | | |

**Fig. 6:** Part of the microstructure of the lexical set 'homosexual'

Furthermore, Figure 6 also shows that the method yields a more nuanced structure of senses including: (1) person, specifically a man, a soldier; (2) relationship/affair; (3) abusive behaviour and (4) propaganda.

## 4.3 Pitfalls of Automatically Extracted Meaning Structures

A closer inspection of the CDH has shown that some of the headword adjectives are missing from graph G, which necessitates the re-generation of meaning structures based on a more balanced corpus. Another possible disadvantage of the automatically retrieved meaning structures is that they might be too fragmented. For instance, in the case of the connected component *bódult* 'dazed', *delíriumos* 'delirious', *ittas* 'drunk', *kábult* 'stunned', and *zavarodott* 'confused', the extracted elements of subcategorization frames 1. driver, motorcyclist, cyclist 2. car driver, chauffeur, police officer, vehicle driver 3. defendant 4. person, individual 5. girl, boy, woman, youth, man should be probably merged under the sense '<person>' in a general-purpose dictionary.

Below, we propose an approach that can help to find the right grain-size of semantic features to characterize senses.

## 5. Generalizing the Adjectival Subcategorization Patterns

The first results showed that the extraction of the nominal contexts suffers from two limitations: the generic parameter setting and the suboptimal clustering of the context noun groups. The generic parameter yields an uneven distribution of context nouns: some of them occur with many or most adjectives in the group, while others are very frequent with only one of the adjectives and are a likely candidate for a collocation. The suboptimal clustering on the other hand results in arbitrary changes in granularity: some noun groups might be too generic, while others are subdivided into several groups.

The problems above arise eventually from the fact that corpus co-occurrences were directly extracted. As a next step, we seek to construct a more abstract representation of context nouns that can (1) bring us closer to a semantic feature set to distinguish between senses instead of finding direct collocates, (2) overcome the idiosyncrasies (e.g., collocations) by generalizing over observed contexts, (3) ideally, selects the right level of generalization for semantic features. Thus, we examined the use of

huBERT, the Hungarian version (Nemeskey 2020) of BERT (Devlin & al., 2019) to represent adjectival subcategorization both individually and at the level of semantic fields defined by connected components. For each adjective, we construct a pattern in the form of "adjective [MASK]".

We then extract the model's predictions for the word following the adjective, i.e., the logit layer of the model output at the masked token position. It is a vector the size of the vocabulary, and – after softmax transformation – can be interpreted as the probability of each word in the vocabulary to appear in the given position. However, we use the values only for ranking. An advantage is that since we don't extract a representation for the adjective itself, the method is insensitive to the tokenization of the adjective. From this vector we retrieve the first n elements of the vocabulary, those that are the most compatible with the preceding adjective. This will be the adjective's subcategorization profile. For instance, as for the adjectives *antidemokratikus* 'antidemocratic', *autokrata* 'autocratic', *despotikus* 'despotic' and *diktatórikus* 'dictatorial', the obtained results show that the subcategorization profile of the semantic frame is composed of the nouns *rendszer* 'system', *hatalom* 'power', *politika* 'politics', *szervezet* 'organization' – obtained by selecting the nouns that appear in the subcategorization of all the adjectives in the group. The only outlier element of this group is 'autocratic' which according to the LLM subcategorises exclusively persons.

## 6. Conclusion and Future Work

The paper outlines the current state of an ongoing research, the ultimate goal of which is to provide a comprehensive and consistent description of the lexical semantics of Hungarian adjectives. Using a graph-based distributional approach, adjectival meaning structures were induced from monolingual corpus data, and compared to the Concise Hungarian Dictionary. While the results are promising from a lexicographic perspective, improvements are still needed: (1) The cut-off parameter $K$ is currently based on manual evaluation and requires more mathematical grounding. (2) Over a third of the relevant adjectives are members of large connected components that merge various semantic fields, necessitating the dissection of the large connected components into more highly connected components. (3) Isolated nodes also may exhibit interesting semantic properties, most importantly homonymy and, therefore, need further investigation. (4) The coverage of results should be improved by retraining the algorithm on more balanced data. Finally, the potential for LLMs to complement the algorithm, particularly in generalizing subcategorization frames, should be further explored.

## References

Ah-Pine, J., & Jacquet, G. (2009). Clique-Based Clustering for Improving Named Entity Recognition Systems. In Lascarides, A. et al. (Eds.), *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 51–59).

Apresjan, J. D. (1974). Regular Polysemy. *Linguistics*, *12*(142), 5–32.

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

Camacho-Collados, J., & Pilehvar, M. T. (2018). *From word to sense embeddings: A survey on vector representations of meaning*. https://arxiv.org/abs/1805.04032

Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/abs/1810.04805

Haber, J., & Poesio, M. (2024). Polysemy—Evidence from Linguistics, Behavioral Science, and Contextualized Language Models. *Computational Linguistics*, *50*(1), 351–417. https://doi.org/10.1162/coli_a_00500

Héja, E., & Ligeti-Nagy, N. (2022a). A Clique-based Graphical Approach to Detect Interpretable Adjectival Senses in Hungarian. In Ustalov, D., Gao, Y., Panchenko, A. et al. (Eds.), *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing* (pp. 35–43).

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), 521–548. https://doi.org/10.1556/2062.2022.00579

Héja, E., Ligeti-Nagy, N., Simon, L., & Lipp, V. (2023). An unsupervised approach to characterize the adjectival microstructure in a Hungarian monolingual explanatory dictionary. In Medved et al.(Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century): Invisible lexicography* (pp. 150–167). Lexical Computing.

Kiefer, F. (2008). A melléknevek szemantikája. In F. Kiefer, (Ed.), *Strukturális magyar nyelvtan 4. A szótár szerkezete.* Akadémiai Kiadó.

Mikolov, T. et al. (2013). *Efficient estimation of word representations in vector space*. CoRR, abs/1301.3781.

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

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In Burnstein et al. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Associa- tion for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers) (pp. 1267–1273).

Pusztai, F. (Ed.) (2003). *Magyar* értelmező *kéziszótár* [Concise Dictionary of Hungarian], 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).

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

## Acknowledgements

This work was partially supported by the "PHC BALATON/HCB BALATON" programme (project number: 49855SF), funded by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and Research and the Hungarian Ministry for Higher Education.

Our special thanks go to the reviewers for their valuable and insightful remarks.

## Contact information

**Enikő Héja**
HUN-REN Hungarian Research Centre for Linguistics, Institute for Language Technologies and Applied Linguistics
heja.eniko@nytud.hun-ren.hu

**Kata Gábor**
INALCO, Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM)
kata.gabor@inalco.fr

**László Simon**
HUN-REN Hungarian Research Centre for Linguistics, Institute for Lexicology
simon.laszlo@nytud.hun-ren.hu

**Veronika Lipp**
HUN-REN Hungarian Research Centre for Linguistics, Institute for Lexicology
lipp.veronika@nytud.hun-ren.hu