Acta Linguistica Academica 69 (2022) 4, 521–548
DOI: 10.1556/2062.2022.00579

A proof-of-concept meaning discrimination experiment to compile a word-in-context dataset for adjectives – A graph-based distributional approach

ENIKŐ HÉJA* and NOÉMI LIGETI-NAGY

Language Technology Research Group, Hungarian Research Centre for Linguistics, Hungary

Received: April 22, 2022 • Accepted: October 3, 2022

ABSTRACT

The Word-in-Context corpus, which forms part of the SuperGLUE benchmark dataset, focuses on a specific sense disambiguation task: it has to be decided whether two occurrences of a given target word in two different contexts convey the same meaning or not. Unfortunately, the WiC database exhibits a relatively low consistency in terms of inter-annotator agreement, which implies that the meaning discrimination task is not well defined even for humans. The present paper aims at tackling this problem through anchoring semantic information to observable surface data. For doing so, we have experimented with a graph-based distributional approach, where both sparse and dense adjectival vector representations served as input. According to our expectations the algorithm is able to anchor the semantic information to contextual data, and therefore it is able to provide clear and explicit criteria as to when the same meaning should be assigned to the occurrences. Moreover, since this method does not rely on any external knowledge base, it should be suitable for any low- or medium-resourced language.

1. INTRODUCTION

The last decade has brought about a huge improvement in natural language processing (NLP) covering most fields of the research area, such as machine translation, text generation and

[*] Corresponding author. E-mail: eniko.heja@gmail.com

question answering (Zhang et al. 2021). Surprisingly, a seemingly well defined, both linguistically and lexicographically relevant, much studied subfield – lexical semantics, the study of modeling lexical meaning – still remained unsolved. Many authors have pointed out both from a lexicographic and from an NLP perspective (e.g. Atkins & Rundell, 2008; Véronis 2003; Kuti, Héja & Sass 2010) that relying on surface observational data is indispensable when compiling sense inventories, such as monolingual explanatory dictionaries, wordnets, etc. This intuition is grasped by the shift of the editorial principles of monolingual dictionaries in the first place, insofar SOTA dictionary compiling methodologies make extensive use of corpus data, primarily in a corpus-based framework. However, the corpus-based methodology applies predominantly to languages with a rich lexicographic tradition, for instance, English, Dutch and French. Unfortunately, in the case of Hungarian there is still a huge gap in the availability of such resources (cf. Lipp & Simon 2021).

A further improvement in the compilation of these databases might be achieved by corpus-driven word sense induction (WSI) approaches. WSI is the task of automatically identifying different senses of a lexical unit from unstructured data, i.e. from corpora without sense labels. Throughout the decades several approaches have been proposed for that purpose, all of them have started from Harris' distributional hypothesis (1954), according to which similar senses tend to occur in similar contexts. The envisaged advantages of unsupervised WSI algorithms are twofold: (1) the editorial work could be hugely diminished compared to both intuition-based and corpus-based approaches (2) the corpus-driven methodology decreases the role of human intuition even further, and in addition, it may handle lexical units not salient for human perception, if they are present in the data. Thus, lexical databases can be complemented with lexical items that went previously unnoticed.

However, a serious consequence of the lack of corpus-based sense inventories is that WSI approaches cannot be evaluated on objective grounds.

To fill this gap we have decided to compile a dataset, the Hungarian Word-in-Context corpus, to enable the evaluation of competing WSI algorithms. The original version of this dataset, the Word-in-Context dataset (Pilehvar & Camacho-Collados 2019), forms part of the SuperGlue benchmark dataset (Wang et al. 2020).

The present paper is structured as follows: in Section 2 the English Word-in-Context corpus is introduced with a focus on the compilation methodology and on its quality. In Section 3 an alternative methodology is put forward to eliminate the drawbacks raised by the editing principles of the dataset, which seem to impose an upper bound even on human performance on the task. Our proposed graph-based distributional approach operates on unlabeled data, and it is completely corpus-driven, thus it does not rely on human intuition at all while coming up with meaning distinctions. The detailed description of the graph-based algorithm utilizes two types of adjectival representation, one dense and one sparse vector representation, both are described in Subsection 3.4. Then the two representations were converted into graphs, which served as the input for the meaning discrimination in Subsection 3.5. The automatic retrieval of the relevant nominal contexts specific to each meaning is described in Subsection 3.6. Section 4 is centered around the qualitative evaluation of the automatically generated meaning distinctions. This was performed from multiple perspectives. In Subsection 4.1 we list the most important research questions that emerged during the workflow. The rest of this section elaborates on these problems. Accordingly, Section 4.2 gives a brief outline on the conception of the relevant aspects of meaning, which underlie the meaning discrimination approach. Then the qualitative evaluation of the results follows in Section 4.3 with regard to the possible effects of the various parameter settings (4.3.1), with regard to a coarse-grained classification of the emerging meanings (4.3.2) and with regard to some practical rules that were applied during the validation of the distinct meanings (4.3.3). The results are then discussed from the perspective of our original goal, i.e. from the perspective of creating the Hungarian Word-in-Context corpus in Section 5, which is followed by the conclusions in Section 6.

2. THE ENGLISH WIC CORPUS

As opposed to the usual sense annotated corpora (e.g SemCor, Landes, Leacock & Fellbaum 1998; MASC-WSA, Ide et al. 2008; OMSTI, Taghipour & Ng 2015, etc.; for a detailed survey see Pasini & Camacho-Collados 2020), which are labeled with multiple sense tags from an existing sense inventory to evaluate word sense disambiguation (WSD) algorithms, the WiC dataset was compiled with a much simpler task in mind. Instead of utilizing a huge label set to assign concrete meanings to the word occurrences, the Word-in-Context corpus comprises only two labels reflecting whether a given target word in two different contexts convey the same meaning or not.

Therefore, instead of the difficult WSD task, the WiC benchmark corpus evaluates a more simple binary classification task. Fortunately, this simplification makes the corpus compilation process less complicated, because we do not need to be aware of the complete meaning characterization of the lexical unit.

2.1. The compilation methodology of WiC

Among the competing corpus building strategies Pilehvar & Camacho-Collados (2019) confined themselves to a more traditional one, namely, the corpus was "constructed using high quality annotations curated by experts". That is, they made use of existing sense-inventories by extracting the relevant example sentences from them, thus, the meaning distinction was based primarily on the content of these databases. Pilehvar & Camacho-Collados (2019) relied on the WordNet (Fellbaum 1998), on the VerbNet (Schuler 2006) and on the English Wiktionary[1] databases. The three resources were mapped through BabelNet (Navigli & Ponzetto 2012). The workflow was made up of three main stages: in the first step, all example sentences were extracted from all three sense-inventories forming positive and negative instances. By definition, in the case of positive instances the target word appears with the same sense, while in the case of negative instances the target word's occurrences convey two different meanings. Two constraints were also considered: "(1) not having more than three instances for the same target word, and (2) not having repeated contextual sentences across instances".

In the second step, WordNet meanings were coarsened by applying a simple pruning technique, that is, all pairs whose senses were first degree connections in the WordNet semantic graph were removed. Later, in the quality-check phase, pruning turned out to be an especially important step. The resulting dataset comprised cc. 7,500 instances with 3,040 various words as

[1] https://www.wiktionary.org/

target words in the training, development and test set altogether. About 52% of the target words belonged to the verbal POS category, while all the remaining target words were nouns.

2.2. Quality check of the WiC corpus

In the quality check phase 4 annotators with no lexicographic background were asked to evaluate 4 sets of 100 randomly sampled instances. ITA between two annotators was also calculated on 50 overlapping instances. This phase served to estimate the human-level performance ceiling, as well. Interestingly, the results showed a rather consistent accuracy score: individual scores were 79%, 79%, 80% and 82%. The ITA on the overlapping instances was 80% as well. Considering the fact that the instances were generated on the basis of manually build sense-inventories, we think that these results strongly imply that the word sense discrimination task is not well defined even for humans. Note, that the random samples from the original version of WordNet (without pruning) yielded much lower accuracy, which equaled only to 57% on average.

The results of the evaluation strongly correlate with the observation of Véronis (2003), who experimented with a very similar task: his research was concerned with the agreement on polysemy – that is, the extent to which coders agree that a word is polysemous or not. Six fourth-year linguistic students were asked to decide whether a word in the context of one paragraph has multiple meanings or only one single meaning in the case of 600 occurrences of 600 French words (200 nouns, 200 verbs, 200 adjectives). Besides, the answer ‘I don’t know’ was also available. Note, that if a context was underspecified regarding polysemy, the relevant answer would be ‘I don’t know’. The low proportion of such answers (4.05%) implies that the majority of contexts were specific enough to make a decision on polysemy. Interestingly, in spite of the low rate of ‘I don’t know’ answers there was a considerably low agreement regarding the polysemous nature of the words in contexts: 0.67 for adjectives, 0.36 for nouns and 0.37 for verbs in terms of the extended version of Cohen’s κ to multiple coders. According to Véronis (2003) these results show that “individual informants had no trouble making spontaneous judgments, but different informants tended to make different judgements”.

Taking both results into consideration, we can draw the conclusion that human intuition is not a reliable source of information in meaning discrimination tasks.

3. AN ALTERNATIVE METHOD TO BUILD THE HUWIC CORPUS

3.1. International and Hungarian background

To eliminate these difficulties we have decided to experiment with a data-driven approach, which aims at anchoring the various sub-senses to surface observational data. According to our expectations the retrieved senses and sub-senses along with their extracted contexts could serve as a basis to build the Hungarian version of the WiC corpus.

There is a wide range of available approaches aiming to extract meanings from unlabeled data in a corpus-driven way. A quite recent thread of research aims to tackle this challenge in the context of dense word representations by creating multi-sense word embeddings (MSEs) (e.g. Neelakantan et al. 2014; Li & Jurafsky 2015; Bartunov et al. 2016).

State-of-the-art meaning discrimination algorithms are based on contextualized embeddings instead of static ones. In their recent study Amrami & Goldberg (2019) showed that the substitute-based approach of Başkaya et al. (2013) transfers to the recently introduced BERT deep masked language model (Devlin et al. 2019) with a very significant improvement in WSI scores.

However, we have decided to experiment with a graph-based method, which – according to our expectations – is able to retrieve the separate meanings of a polysemous word along with the relevant contexts. The relevance of graphical approaches is proved by the fact that increasingly sophisticated graphical models dominated the state-of-the-art results of WSI up until recently (e.g. Lau, Cook & Baldwin 2013; Wang et al. 2015; Komninos & Manandhar 2016; Amplayo, Hwang & Song 2019; cf. Amrami & Goldberg 2019).

Although there is a wide variety of methods for WSI, as far as we know, investigations focusing on the Hungarian language are rather restricted in number. Earlier work comprises two small scale pilot projects. The first experiment concerns the automatic sense induction of verbs on the basis of their complementation patterns (Gábor & Héja 2007), where the main emphasis was laid on finding the relevant feature space. Another experiment (Héja & Takács 2010) aimed at detecting polysemous adjectival senses on the basis of a graph-based algorithm yielding promising results.

More recent approaches are centered around static word embeddings following two threads of research. The objective of Novák & Siklósi (2017) and Ficsor & Berend (2021) is to semantically interpret the embedding space. Because of conflating senses their results do not seem to be directly applicable to our purposes. As opposed to them, Borbély et al. (2016) and Makrai & Lipp (2017) investigated to what extent existing MSE models trained on Hungarian data are able to differentiate between senses of static word embeddings.

Based on the wide variety of available representations and algorithms in the international scientific scene, we think that there is much room for us to investigate word sense induction for Hungarian.

In our recent experiment we elaborate on the results of the pilot project performed by Héja & Takács (2010). Although the original WiC corpus comprised verbs and nouns as target words, we have decided to experiment with adjectives. The main motivation behind is that according to our presupposition adjectives can be represented with a much simpler set of features than the original POS categories.

Although their attention was centered around named entities, we basically followed the steps described in Ah-Pine & Jacquet (2009). Beside the investigated phenomenon, another big discrepancy was that we also experimented with static word embeddings. Throughout the workflow, one striking observation was that the static word embedding representation is more handy than that of the more traditional probability distribution approaches. In the case of the former, many preprocessing steps can be omitted from the workflow, moreover, recent embedding approaches do not require manually-selected features for word sense induction.

3.2. The representation of the investigated phenomena

3.2.1. Representing adjectives as graphs. In our graph-based representation of adjectives, vertex-labeled undirected graphs were generated. Vertices and their labels represent the adjectives, while the edges (or their lack) denote whether there is a semantic similarity relation between two adjectives (or not). This structure encodes some basic intuitions about meaning similarity:

1. 'Undirectedness' guarantees the symmetric nature of meaning similarity: if a meaning M is similar to meaning M', then the reverse is also true.
2. Since every adjective is similar to itself, there is a self-loop at every node of the graph.

3.2.2. Representing near-synonyms as cliques. Meaning is grasped through the notion of near-synonymy. Accordingly, two lexical units are near-synonyms if every occurrence of the one lexical unit can be substituted with an occurrence of the other lexical unit in a certain set of contexts so that the meaning of the utterance does not change significantly (Ploux & Victorri 1998). According to our hypothesis, near-synonyms exhibit "very similar" distributional behavior. Note that by the notion of near-synonymy we do not mean synonymy exclusively: that is, near-synonymy covers lexemes from tight semantic classes as well, such as names of days, colors, nationalities, numbers, etc. Besides, according to our expectation, there are sets of lexemes that are real synonyms in a restricted range of contexts. For instance, *lágy* ('soft') and *finom* ('smooth') are synonyms in the contexts of *átmenet* ('transition'), *árnyalat* ('shade') or *zene* ('music'), while *lágy* ('soft') and *puha* ('tender') are synonyms in the context of *margarin* ('margarine') and *fém* ('metal'). Our main objective in the present proof-of concept experiment is the detection of adjectival cliques of this type: cliques that are synonyms in a given set of contexts.

Following Ah-Pine & Jacquet (2009), near-synonyms which exhibit "very similar" distributional behavior, are grasped by cliques in the graph: that is, we search for those subgraphs that are maximally connected, i.e. where every vertex is connected to every other vertex in the subgraph. Now the nodes in the clique represent a set of adjectives with "very similar" distributional behavior. For example, as can be seen on Figure 1, the Hungarian adjectives *varázslatos* ('magical'), *mesés* ('fabulous'), *káprázatos* ('dazzling'), *gyönyörű* ('beautiful'), *csodálatos* ('marvelous'), *csodás* ('wonderful'), *fantasztikus* ('fantastic') all exhibit similar distributional behavior to one another, thus according to our hypothesis these adjectives belong to the same near-synonymy class. And indeed, these adjectives have closely related meanings.

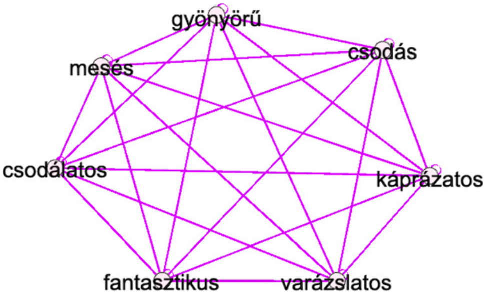

Figure 1. A maximally connected subgraph – a clique – representing a near-synonymy class for *varázslatos* 'magical'

3.2.3. Representing meaning-discriminations as shared cliques. This approach, on the one hand, makes it possible to detect multiple near-synonymy classes comprising a common adjectival lexeme, where the corresponding cliques represent differing sense candidates. In addition, ideally, it also enables meaning discrimination based on explicit surface data, inasmuch all the resulting cliques are anchored to the contexts in which each element of the adjectival clique may occur.

Therefore, according to our hypothesis, an adjective has multiple meanings if it belongs to multiple cliques, and the cliques are characterized by non-overlapping sets of context nouns.

Accordingly, the workflow conceptually comprises two main stages:

1. the detection of near-synonymy classes for a given adjective,
2. discriminating between the various meanings of the given adjective by the extraction of the relevant context nouns.

For instance, as Figure 2 illustrates, the Hungarian adjective *tárgyilagos* ‘objective’ belongs to two different cliques:[2] [*tárgyilagos* ‘objective’, *pártatlan* ‘impartial’, *elfogulatlan* ‘unbiased’] and [*tárgyilagos* ‘objective’, *tárgyszerű* ‘concise’, *tényszerű* ‘factual’].

According to our hypothesis these two cliques represent two different senses of *tárgyilagos* ‘objective’, where one sense is characterized by the adjectives *elfogulatlan* ‘unbiased’ and *pártatlan* ‘impartial’, while the other sense is characterized by the adjectives *tényszerű* ‘factual’ and *tárgyszerű* ‘concise’. These cliques correspond to the human intuition insofar the first meaning describes a situation where no participant was favored during the act described by the modified noun, and the second meaning refers to the fact that the denotatum of the modified noun corresponds to the reality. This meaning distinction has to be validated by the relevant context nouns in the next step: based on corpus data the first sense of *tárgyilagos* emerges in the context of nouns such as *vélemény* (‘opinion’), *mód* (‘manner’), *eljárás* (‘procedure’), *ítélkezés* (‘judgment’), *megfigyelő* (‘observer’), while the second sense appears before the nouns *leírás* (‘description’), *vita* (‘discussion’), *ismertetés* (‘review’).

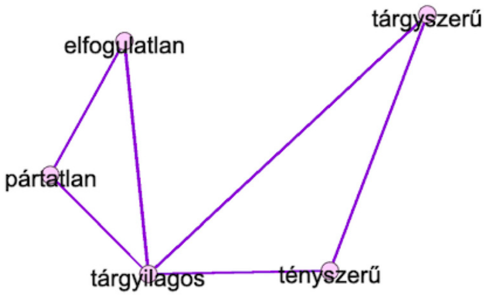

Figure 2. The Hungarian adjective *tárgyilagos* ‘objective’ belongs to two different cliques

[2] Note that for the sake of readability self-loops are omitted.

3.3. Methodology in brief

Our experiments can be told apart in two broad categories based on the representation of the adjectives. However, regardless of the representation of the adjectives, the basic steps of the methodologies are the same and are shortly described in what follows:

1. Selecting the set of relevant adjectives.
2. Creating the two types of adjectival representations.
   (a) At first static dense word embeddings were used to construct the distributional space.
   (b) In the second experiment adjectives were represented with sparse vectors based on the probability distributions of the following nouns estimated from corpus data.
3. Forming adjectival sub-senses by constructing multiple near-synonymy classes (cf. Figure 2).
   (a) In this phase first a similarity matrix was created (A_{sim}) containing adjectives as rows and columns. For doing so, a suitable similarity/distance measure was applied to fill in the cells of A_{sim}. That is, $A_{sim}\ (i, j) = sim\ (a_i, a_j)$, where a_i and a_j denote adjectives from the selected vocabulary.
   (b) In the second step A_{sim} similarity matrix was converted into an adjacency matrix A_a based on a suitable cutting heuristics. A_a matrix contains only 0 and 1 values indicating whether two given vertices in the graph are connected or not (0 or 1, respectively). More precisely, $a_{i,j} = 1$ denotes that a_i and a_j vertices of the graph are connected, while $a_{i,j} = 0$ denotes unconnected vertices in the graph. This in turn indicates whether the corresponding adjectives are semantically similar or not.
   (c) We search for undirected graphs due to the symmetric nature of meaning similarity. Undirected graphs are represented as symmetric adjacency matrices. Therefore, A_a adjacency matrix is symmetrized and, as a result, in this step A'_a symmetric square matrix is generated containing boolean values. A'_a adjacency matrix can be conceived of as a graph representation of the adjectives.

   Due to the reflexive nature of 'similarity' all the diagonal values of A'_a equal to 1.

   (d) In the last phase, maximally connected subgraphs, so-called cliques, were retrieved from the graph represented by the adjacency matrix to grasp adjectival near-synonymy classes.

Figure 3 depicts the neighborhood graph of the node *tárgyilagos* 'objective'. In this case we started from the dense embedding representation of adjectives. The edge weights are calculated as the similarity between the vector representations of the adjectives. The actual weights are the edge labels. The neighborhood graph contains all the directly connected nodes of the whole adjectival graph where the weights are greater or equal to 0.7. Thick lines represent such edges, which were taken into consideration while searching for cliques. Note that the edges between all the neighbor nodes are also present in Figure 3, regardless of their actual value: if their weights are less than 0.7, these edges are represented by thin lines, which denote edges that were discarded in the next step.

4. Validation of the results
   (a) The resulting adjectival cliques are validated by retrieving the set of nouns they may co-occur with. More precisely, we consider only those adjectival cliques to be potential candidates for adjectival meaning representation that have at least one co-occurring noun in common.

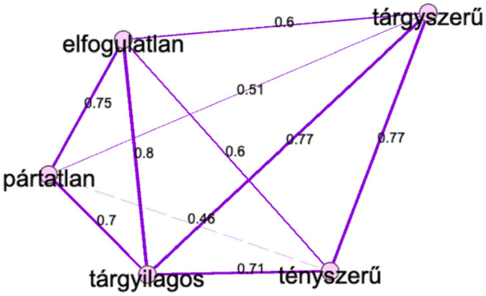

Figure 3. The neighborhood graph of *tárgyilagos* with edge weights above 0.7

(b) As the motivation for the research was to give observable criteria to meaning discrimination, the contexts specific to the given adjectival meaning candidate has to be specified. Thus, we keep only those subsequent nouns that are *specific* to the clique, that is, they did not occur with the elements of other competing cliques.

5. Evaluation of the results: Finally, the results were evaluated according to different parameter settings. Since, to our knowledge, there is no similar database available for Hungarian, a qualitative evaluation was performed. However, it is rather important to point out that throughout the evaluation process we stuck to our original objective: we investigated whether the automatically retrieved meaning candidates represent different meanings in the context of the automatically retrieved set of nouns.

3.4. The methodology in detail

3.4.1. Selection of the adjectives. Although the original word-in-context corpus contains meaning discrimination regarding nouns and verbs, we have decided to investigate adjectival meaning as the first stage of our thread of research. The main motivation behind was that according to our intuition the distributional representation of adjectival meaning is primarily determined by the following noun, so all we need to do is to extract adjective-noun pairs from a corpus, which is much more straightforward than to come up with the proper feature sets representing nouns and verbs. First, the adjectives of our interest were selected on the basis of the 180 million word Hungarian National Corpus (Váradi 2002). Although the frequency list contains adjectives with various case suffixes, we took only nominative adjectives into consideration, presuming that the adjective is always in nominative in the Adj + Noun constructions.

The resulting adjectival frequencies served as one parameter in our sense-extraction experiment.

3.4.2. The representation of the adjectives

3.4.2.1. Dense static word embeddings. The first class of experiments intended to grasp adjectival polysemy by relying on their word2vec representations. There are multiple threads of research in the Hungarian literature regarding static word vectors (Makrai 2015; Siklósi & Novák 2016; Siklósi 2016; Borbély et al. 2016; Makrai & Lipp 2017; Szántó, Vincze & Farkas 2018; Döbrössy et al. 2019), but as far as we know, no one has investigated to what extent these kinds of representations can be used to discriminate between multiple meanings. On the contrary, it is usually emphasized that, as opposed to contextual word embeddings, static vectors tend to lump together the various senses of a word, which is considered to be one of their main drawbacks.

Siklósi (2016) trained a CBoW model on two corpora: first on a tokenized, but otherwise raw corpus, where different vectors were derived for the different surface forms of the same word, and second, on a tokenized and morphologically annotated version of the corpus where each word form was represented by two tokens: a lemma, and the morphological analysis of the surface form as a tag. The embedding vectors were used for extracting coherent semantic groups from the corpus. Döbrössy et al. (2019) also experimented with different settings and found that using lemmas instead of words resulted in better semantic accuracy of the embeddings evaluated on the Hungarian translation of the Google analogy test set (Makrai 2015). Szántó, Vincze & Farkas (2018) trained a word level skip-gram model and a character level skip-gram model (based on Google's FastText algorithm, Bojanowski et al. 2017) on various Hungarian texts. Character level embeddings proved to be better on document classification tasks.

As opposed to Döbrössy et al. (2019), we presuppose that preserving morphosyntactic information contributes to the characterization of meanings. This assumption is supported by the findings of Levy & Goldberg (2014), according to whom dependency-based embeddings (where the contexts of a word are derived based on the syntactic relations it participates in) are less topical and exhibit more functional similarities of a co-hyponym nature than original skip-gram embeddings with their bag-of-words nature. In their experience set up, three training conditions were used: bag-of-words contexts (one with $k = 5$ and one with $k = 2$) and dependency-based syntactic contexts. Their qualitative evaluation shows that BoW contexts reflect the domain aspect of a word, finding words that associate with it, while the dependency-based contexts capture the semantic type of the target word, finding words that behave like it. They also mention Turney (2012) who described this distinction as domain similarity versus functional similarity.

We used word2vec word embeddings (Mikolov, Chen et al. 2013; Mikolov, Sutskever et al. 2013) trained on the first 999 file (21 GB raw texts) of the Webcorpus 2.0 (Nemeskey 2020) containing the normalized version of the original texts, cc. 170M sentences. It is important to note that the training was performed on word forms and not on lemmata, thus word forms were assigned vector representations.

300-dimension vectors were trained using the Gensim Python package (Rehurek & Sojka 2011) to perform CBoW training with a $6k$ window size. Since Hungarian is a highly inflected language and we trained embeddings on raw texts, this is not a pure bag-of-words model, as the abbreviation CBoW would imply. Roughly 8.5M word forms were assigned embeddings as the result of the training.

The choice of the LM was not unmotivated: multiple embeddings with the same hyperparameters were evaluated on the Hungarian version of the Google analogy test set (Makrai 2015), the only difference between them being the size of the training data. Accordingly, LMs were constructed on the first 9 (cc. 1M sentences), on the first 99 (cc. 13M sentences), and on the first 999 files of the Webcorpus 2.0. The whole Webcorpus 2.0 (cc. 590 M sentences) was used as training data as well. Based on the evaluation results, despite its size (cc. 20.7M word forms), the LM trained on the whole Webcorpus 2.0 yields only a slight improvement on the analogy task in comparison with the model trained on the first 170M sentences. Moreover, programs utilizing the latter model are much slower, due to the model size.

The trained LMs are available on GitHub.[3]

3.4.2.2. Sparse vector representations. Beside static embeddings, traditional probability distributions were also tested. For that purpose the steps described in Ah-Pine & Jacquet (2009) were followed. However, as it was mentioned above, their research was centered around named entities and not around adjectives.

In the first step the adjective-noun co-occurrence pairs had to be retrieved. This was performed on the basis of a 91.4 million-token subcorpus of the Hungarian Gigaword Corpus (Oravecz, Váradi & Sass 2014) compiled specifically for the present experiment.[4] During the compilation process we aimed at preserving the original proportion of the genres, thus, every domain of HGC was included in the new corpus: newspapers, literature, scientific, official, personal and spoken language. Accordingly, our corpus was made up of 30.5, 6.5, 11.6, 8.8, 28 and 6.6 m tokens, respectively.

In the second step, the distributional space was created for adjectives by creating an $A \times N$ matrix, where

$$\begin{aligned} A = \{ADJ|Freq_{ADJ} \geq threshold_1\} \\ \text{and} \\ N = \{NOUN|Freq_{<ADJ,NOUN>} \geq threshold_2\} \end{aligned} \tag{1}$$

Accordingly, adjectives from the original adjective list were kept with a frequency above a certain threshold, while only those Ns were considered that co-occurred with at least one of the adjectives more frequently than a certain threshold.

In the third step the maximum likelihood estimates (MLEs) for every adjective a_i were calculated as $P_{mle}(n_j|a_i) = \frac{C(a_i,n_j)}{C(a_i)}$, where $C\ (a_i,\ n_j)$ is the co-occurrence number of the given adjectives before the given noun, while $C\ (a_i)$ is the overall frequency of a_i in the whole corpus.

Unfortunately, as most of the nouns do not co-occur with a given adjective, this estimate of the probability distributions leads to sparse data, which is not suitable for measuring similarities. To overcome this difficulty we followed Ah-Pine & Jacquet (2009) as well, and applied Jelinek-Mercer smoothing to assess the probabilities of unseen events. Thus, Jelinek-Mercer smoothing changes values in the $A \times N$ matrix. The modified probability value $P'_{mle}(n_j|a_i)$ in the $(a_i,\ n_j)$ cell is calculated as follows:

$$\begin{aligned} P'_{mle}(n_j|a_i) = \lambda \times P_{mle}(n_j|a_i) + (1-\lambda) \times P_{mle}(n_j|CORP) \\ \text{where} \\ 0 \leq \lambda \leq 1 \quad \text{and} \quad P_{mle}(n_j|CORP) = \frac{\sum_i C(a_i, n_j)}{\sum_{i,j} C(a_i, n_j)} \end{aligned} \tag{2}$$

[3] https://github.com/nytud/w2v_models

[4] We decided to use different corpora for the different subtasks of this study to ensure the robustness of our proof-of-concept research.

The basic idea behind Jelinek-Mercer smoothing is that in the case of seen events it takes the original estimate with less weight into account by multiplying it with λ, while in the case of unseen events it assigns a probability on the basis of the context with a weight $1 - \lambda$. Jelinek-Mercer smoothing estimates the probability of the unseen event by estimating the probability of the context in which the event was unseen. In our case it means that we presume that an adjective has a better chance to show up in the context of a frequently occurring noun than in the context of a low-frequency noun. The frequency of the nominal context was estimated through the frequency of the overall occurrence of the noun with any of the investigated adjectives. This value was then normalized with a constant value: the total frequency count of all the adjective-noun pairs.

It is important to note that in this case the λ parameter of Jelinek-Mercer smoothing has an effect on the quality of the clique detection algorithm. That is why special attention has been paid to it during the evaluation. For instance, according to our observation, if λ is set too high the following nouns tend to become "too salient", forming collocational and thus incoherent cliques on the basis of the following noun (cf. 1.c in Section 4.3.1).

Ah-Pine & Jacquet (2009) set the value of λ to 0.5, giving the same weight to seen and unseen events. However, according to our intuition, seen data should be given more weight, especially if the size of the corpus is big.[5]

3.5. Constructing multiple near-synonymy classes

3.5.1. Generating the similarity matrix. In this step the similarity matrix is generated, that is, the $A \times 300$ matrix – in the case of dense representations – or the $A \times N$ matrix – in the case of sparse representations – is sent into an $A \times A$ matrix, where each cell a_{ij} consists of a value denoting the similarity between two adjectives, a_i and a_j. Below Figures 4 and 5 illustrate this step with some examples taken from Figure 1.

	DIM$_1$	DIM$_2$	DIM$_3$	DIM$_4$	...	DIM$_n$
mesés 'fabulous'	k$_{11}$	k$_{12}$	k$_{13}$	...	...	k$_{1n}$
káprázatos 'dazzling'	k$_{21}$	k$_{22}$	k$_{23}$	...	...	k$_{2n}$
varázslatos 'magical'	k$_{31}$	k$_{32}$	k$_{33}$	...	...	k$_{3n}$

Figure 4. The vector representation of three Hungarian adjectives

	mesés	*káprázatos*	*varázslatos*
mesés 'fabulous'	1	0.76	0.84
káprázatos 'dazzling'	0.76	1	0.77
varázslatos 'magical'	0.84	0.77	1

Figure 5. The $A \times A$ matrix, where the cells are filled in with the similarity values

[5] And, indeed, the size of the input used in Ah-Pine & Jacquet (2009) equals to about 300,000 tokens, which make the choice of λ comprehensible.

3.5.1.1. Static word embeddings. Two different approaches were used to measure similarity. In the case of the word2vec embeddings the usual cosine similarity was calculated. That is:

$$sim_{\cos}(v_1, v_2) = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|v_1\|\|v_2\|} \tag{3}$$

Cosine similarity measures the angle between two vectors: the more cos (v_1, v_2) is closer to 0 the more similar are the corresponding vectors. The diagonal values are 1s, since everything is very similar to itself.

3.5.1.2. Probability distributions. Following Ah-Pine & Jacquet (2009) cross entropy (CE) was used as a distance measure to compare probability distributions. Cross entropy is a standard way to compare two probability distributions. For discrete probability distributions it is calculated as follows:

$$\begin{gathered} CE(P,Q) = -\sum_{x} P(x) \times logQ(x), \text{that is,} \\ CE(a_i, a_i') = -\sum_{n_j \in N} D'(a_i, n_j) \times \log D'(a_{i'}, n_j) \end{gathered} \tag{4}$$

where D' stands for the smoothed distributional $A \times N$ matrix.

However, despite its wide usage as distributional distance, it is somewhat counter-intuitive to use CE as a similarity measure. First of all, if two probability distributions are the same, then the cross-entropy between them will be the entropy of the distribution, that is, as opposed to cosine similarity, it is not even a constant value. Secondly, CE is not symmetric, i.e. $CE\ (p, q)$ and $CE\ (q, p)$ are not equal.

We will see that both properties of CE have to be taken into consideration throughout the creation of the adjacency matrices.

3.5.2. Generating the adjacency matrix A_a.

In this stage the adjacency matrix A_a is generated from the similarity matrix A_{sim}. Again, we have two distinct approaches, just in the case of similarity matrices, reflecting the fact that the adjectives were represented with two different feature sets.

3.5.2.1. Static word embeddings. The symmetric nature of cosine similarity guarantees the expected symmetry of the adjacency matrix. Now we only have to make a decision regarding the cut-off value. Therefore, a threshold value is introduced at this step to tell apart the cosine similarity values into two classes: the similarity values in the $A \times A$ matrix below the threshold will be set to 0, while the values equal or above the threshold will be set to 1. Again, $a_{ij} = 1$ denotes an edge between a_i and a_j, while $a_{ij} = 0$ denotes that the two vertices are not connected in the corresponding graph. At the end of this step, the graph representing adjectival meanings is construed. Note that the cut-off value has a huge impact on the results both in terms of size and quality, thus, great attention should be paid to the optimal parameters.

3.5.2.2. Probability distributions. The A_{sim} similarity matrix was converted into the adjacency matrix A_a following the steps exactly described in Ah-Pine & Jacquet (2009). Namely, we take the matrix row-by-row.

In the first step all the adjectives in the row of a_i are ranked according to descending order based on their similarity value[6] with a_i: $<a^i_1, a^i_2, \ldots, a^i_{|A|}>$, where $|A|$ is the number of adjectives in the vocabulary.

Then a p-long list of the most significant a_j's is gathered by choosing the ones that bring the most relevant similarities according to the following criteria:

$$
\begin{aligned}
L(a_i) = \left\{ a^i_1, \ldots, a^i_p : \frac{\sum_{i'=1}^{p} CE(a_i, a^i_{i'})}{\sum_{i'=1}^{|A|} CE(a_i, a_{i'})} < q \quad \text{and} \quad p \leq b \right\}, \\
\text{where} \quad 0 \leq b \leq |A| \\
\text{and} \quad 0 \leq q \leq 1
\end{aligned}
\tag{5}
$$

Two different criteria are set to filter the best candidates: first, the sum of the potential nearest neighbor adjectives' CE values normalized by the sum of the overall CE values in the given row can be given an upper limit. Secondly, the number of possible candidates also can be constrained. At the present stage of investigation this parameter was as simple as possible, therefore we considered the 10 nearest neighbors, i.e. $p = 1$ and $b = 10$.

Since in most cases $CE\ (a_i, a_j) \neq CE\ (a_j, a_i)$, the resulting matrix had to be symmetrized as well. This was done as follows:

$$
A_a(a_i, a'_i) = \begin{cases} 1 & \text{if} \quad a_i \in L(a'_i) \quad \text{or} \quad a_{i'} \in L(a_i) \\ 0 & \text{if} \quad \text{otherwise} \end{cases}
\tag{6}
$$

3.5.3. Generating the cliques. In this step near-synonyms are searched for by detecting cliques in the adjacency matrices. Since the adjacency matrices represent distributional information from the two different sources in a uniform way irrespective of the representation of adjectives, the clique-detection step is the same in both cases.

However, not all detected cliques were preserved for evaluation. The cliques were validated in the next phase. The validation step was made up of two steps: (1) the near-synonymy class for a given adjective was accepted if there was at least one common context noun (2) the meaning discrimination between the various cliques of the given adjective was validated by non-overlapping sets of nouns, as described in the next section.

3.6. Validation: following nouns

As mentioned above, adjectival cliques are validated by retrieving the set of nouns they may co-occur with. According to our expectation, different senses of an adjective are characterized by the different sets of nouns they co-occur with. These non-overlapping sets provide explicit information on the context of meaning discrimination. A characteristic set of nouns is found as follows:

1. We collect all the nouns an adjective co-occurs with; we do this for all adjectives in a clique. This step is done on the 91.4 million-token subcorpus of the Hungarian Gigaword Corpus (described in section 3.4.2.2).

[6] Note that since CE is a distance measure, in reality CE values were sorted into ascending order.

2. We compute the intersection of the above sets: those are the nouns that co-occur with each adjective of a clique. If at least one such noun exist for a clique, then we consider the given clique as a potential meaning candidate.
3. We repeat step 1 and step 2 for each clique a given adjective belongs to. This results in a set of nouns for each clique.
4. Finally, we take these sets and omit the intersections: we keep only the nouns for a clique which are exlusive to the given clique; they do not appear in the sets of the other cliques. Example 1 shows the cliques of the adjective *cinikus* 'cynical'. The nouns listed below the cliques are those shared by all members of the clique. Nouns in bold are the ones specific to the clique. These are the nouns indicating the specific meanings therefore we kept them for further evaluation.

Ex. 1 *cinikus* 'cynical'
Clique 1: *ostoba* 'silly', *cinikus* 'cynical', *demagóg* 'demagogic'
Nouns: *dolog* 'thing', *kérdés* 'question', *lépés* 'move', *mód* 'way', ***szöveg*** 'text'
Clique 2: *ostoba* 'silly', *cinikus* 'cynical', *arcátlan* 'impudent'
Nouns: *dolog* 'thing', ***ember*** 'person', *kérdés* 'question', *lépés* 'move', *mód* 'way'

Our presumption is that the resulting sets of nouns are the ones specific to the given cliques: they capture the given sense of the adjective that is shared among the other adjectives of the clique.

4. EVALUATION

4.1. Research questions

The main objective of the evaluation phase was to verify our basic hypothesis, according to which the proposed techniques are able to provide solid methodological background to discriminate between meanings and, therefore, to compile the HuWiC benchmark. That is, the evaluation was carried out from this perspective: it was investigated to what extent the retrieved cliques enable us to make motivated binary decisions, instead of supplying full-fledged description of adjectival senses and sub-senses. However, our results may be suitable to give a more complete characterization of adjectives, which – as far as we know – is still missing from the Hungarian linguistic literature.[7]

Unfortunately, both representations of adjectives make use of a bunch of parameters, most importantly:

(i) The frequency of adjectives[8] ($Freq_{ADJ}$) and the co-occurrence frequency of the adjective-noun pairs ($Freq_{<ADJ,NOUN>}$) play an important role in both cases.

[7] The most complete description of the lexical semantics of adjectives is Kiefer (2008).

[8] This is in accordance with the fact that corpus-based lexicography has been unable to come up with a widely accepted agreement on the minimum frequency of the occurrences of an item to be characterized (e.g. Sinclair 1991, 1998 proposed a minimum of 150 corpus occurrences, while more recent works raised this threshold to 500, e.g. Atkins & Rundell 2008).

(ii) The K cut-off parameter has a great impact on the generation of A_a adjacency matrix in the case of the word embedding representations.
(iii) The situation is more complicated in the case of the probability distribution representations: here the λ parameter in the Jelinek-Mercer smoothing and the parameters constraining the conversion of A_{sim} similarity matrix into A_a adjacency matrix also had to be set.
(iv) The minimum frequency count of the nouns in the validation step was an additional parameter.

At the present stage of research only some of the parameters were experimented with. Various settings for $Freq_{ADJ}$, K and λ were tested, but more research needs to be done in this direction.

As the meaning discrimination task is conceptually made up of two steps – the detection of the near-synonyms and the discrimination between the corresponding meanings – we reflect on these steps separately by seeking answers for the following questions:

1. Which cliques are "ideal" from a meaning discrimination point of view?
2. What are the properties of an "ideal" clique?
3. What conditions should be met by the validator nouns to validate the cliques and to specify the corresponding meaning?

In the first attempt we set the threshold of the co-occurrence of the ADJ-N pair to 5, meaning that they are required to co-occur at least five times in the corpus. The first qualitative evaluation of the results showed, however, that many promising cliques were eliminated this way as they did not pass this filter. There are relatively few nouns appearing more than five times with a given adjective. To be able to go on with the qualitative analysis, we decided to set this threshold to 1: even one co-occurrence is enough to get a noun to the list. We are aware, of course, that this is a risky parameter setting: by letting in every possible noun appearing together with the adjective we may increase the noise in the list of nouns.

4.2. Conception of meaning from a sense discrimination perspective

Our primary objective is to detect distinct classes of attributive modification, where the adjective conveys different meanings. Distinct meanings may come from different sources. It is common to differentiate between collocational and more productive uses of an expression. In the present experiment we interpret productivity as a scale. On the one end of this scale there are collocations where both the adjective and the noun are fixed. In this case the meaning of the construction is yielded in a fully non-compositional way: neither component can be substituted with a near-synonym preserving the original meaning of the expression (e.g. *fehér zaj* 'white noise' or *fekete doboz* 'black box').

Albeit collocations are possible sources of additional meanings, we are more interested in 'semi-compositional' constructions in the present WSI task, where compositionality operates on a restricted set of adjectives or nouns. For example, *fehér/szürke/fekete gazdaság* (literally 'white/gray/black economy')[9] are not considered collocations in the strict sense, since the

[9] Here, as opposed to the meaning of the English expression ('health related goods and services'), the Hungarian counterpart of 'white economy' refers to the monitored and taxed sectors of economy.

restricted set of colors denotes a new dimension of meaning in the context of the noun *gazdaság* ‘economy’ (i.e. the extent to which a sector of economy is monitored and taxed). That is, one step further from collocations on the ‘productivity scale’ more interesting instances emerge, for example, *ékes* (‘ornate’) means *tipikus* (‘typical’) before a restricted set of nouns (*példa* ‘example’ and *képviselő* ‘representative’). And indeed, the most interesting cases are those where the nouns form one or more semantic classes allowing the adjectives in the cliques to be synonyms in those semantically restricted contexts. For example, the different meanings of *könnyű* (‘easy’), *komoly* (‘serious’), *szép* (‘nice’), *éles* (‘sharp’), *finom* (‘fine, delicate’), all can be discriminated on the basis of a set of synonym adjectives along with their semantically constrained nominal contexts. For example *könnyű* (‘easy’) has different meanings in the context of nouns referring to physical objects (‘a lightweight bag’), nouns referring to clothes (‘a light clothing’), foods (‘a light lunch’), and before nouns like ‘answer’, ‘task’, ‘solution’ (‘an easy answer/task/solution’).

The size of the semantically constrained nominal set may vary: on the other end of the scale there are really productive uses of adjectives that are still important for our purposes. For instance, the retrieved cliques imply that *vidám* ‘merry’ and *szomorú* ‘sad’ have different meanings when modifying nouns denoting humans and when modifying nouns referring to time periods. According to the cliques, we can say both *szomorú [időszak, év, nap]* (‘sad [period, year, day]’) and *gyászos [időszak, év, nap]* (‘mournful [period, year, day]’) but there is neither *bánatos [időszak, év, nap]* (‘sorrowful [period, year, day]’), nor *gyászos [lány, ember]* (‘mournful [girl, human]’).

Ex. 2 Clique 1: *szomorú* ‘sad’, *gyászos* ‘mournful’
időszak ‘period’, *year* ‘év’, *nap* ‘day’
Clique 2: *szomorú* ‘sad’, *bánatos* ‘sorrowful’
lány ‘girl’, *ember* ‘human’

The adjective *vidám* ‘merry’ exhibits rather similar behavior to *szomorú* ‘sad’ from this perspective.

Ex. 3 Clique 1: *vidám* ‘merry’, *derűs* ‘bright’
perc ‘minute’, *nap* ‘day’, *hétvége* ‘weekend’
Clique 2: *vidám* ‘merry’, *jókedvű* ‘cheerful’
fiú ‘boy’, *delfin* ‘dolphin’

As opposed to humans (and dolphins), periods of time cannot be *jókedvű*, and in tandem with this, *derűs fiú* and *derűs delfin* are not well-formed constructions in Hungarian.

4.3. Evaluation of the adjective cliques based on the word2vec representations

In what follows the first results yielded by the evaluation of the word2vec adjectival representations will be presented. This includes:

1. the investigation of some basic parameter settings,
2. the classification of emerging cliques,
3. finally, some clique validation rules are presented, which were formed during the evaluation.

4.3.1. The effects of basic parameter settings

1.a *The impact of K cut-off value*

Not surprisingly, during the evaluation of the word2vec representations we have found that the value of the K cut-off parameter has a serious impact on the number of the resulting cliques and also on the semantic field to which they belong to. For instance, in the case of adjectives occurring at least 200 times, $K = 0.9$ yielded only a handful of results: only 8 adjectives were assigned to more than one clique and only two cliques were validated by nouns. The retrieved cliques refer to numbers, months and days exclusively, therefore, they are not very interesting from a word-in-context point of view. On the other hand, with the same parameter settings, but with a lower similarity cut-off value ($K = 0.7$) we had 446 different adjectives belonging to multiple cliques, where all cliques are validated and discriminated by at least one following noun. This coverage may seem rather low as well, as from the 6,213 adjectives occurring at least 200 times in our input corpus 6,042 adjectives had word2vec representations. With $K = 0.7$ parameter setting there were 3,847 isolated adjectives and 1,085 adjectives belonging to sub-graphs with only two connected nodes, which obviously cannot form part of shared cliques. This means that with this parameter setting there is at most 1,110 adjectives to be assigned to multiple cliques.

1.b *The effect of the frequency count of the following noun*

The minimum frequency count of the validating nouns ($Freq_n$) also had to be taken into consideration. Two settings were tested ($Freq_{ADJ} = 200$, $K = 0.7$). In the first setting we considered a clique to be valid if there was at least one noun occurring at least 5 times with every element of the clique ($Freg_n \geq 5$). Validating only a handful of cliques, this threshold value was deemed to be too high. To keep the coverage as high as possible, the value of $Freq_n$ was set to 1. The overall frequency of the nouns in the corpus was set to 50. This change clearly improved the coverage, yielding 446 adjectives belonging to multiple cliques. In the rest of this section the results of the qualitative evaluation of these cliques will be presented, if not explicitly stated otherwise.

1.c *λ parameter of Jelinek-Mercer smoothing*

Although this detailed evaluation concentrates on the word2vec representation of adjectives, it needs to be mentioned that adjectival cliques derived with the second methodology were also evaluated,[10] the only difference between them being the λ parameter, 'the degree of smoothing' in Jelinek-Mercer smoothing. We used the original $\lambda = 0.5$ value (Ah-Pine & Jacquet 2009), which we found somewhat counter-intuitive, as they give the same weight to seen and unseen events, and also experimented with much higher values ($\lambda = 0.9$ and $\lambda = 0.99$). Contrary to our original expectation, we found that $\lambda = 0.5$ have yielded much more coherent cliques than $\lambda = 0.9$ and $\lambda = 0.99$. It seems that if λ is set too high the following nouns tend to become "too salient", forming collocational and thus incoherent cliques. For instance, *reménytelen* 'hopeless' is assigned to a clique made up of the following semantically

[10] Adjectives with more than 400 occurrences, adjective-noun co-occurrence frequency is greater than 100, the frequency of validating nouns equals to 1 and, as it was described above, the cut-off parameters fall back to the default value – the first 10 most similar adjectives were considered when building the adjacency matrix.

rather diverse adjectives: [*képtelen* 'nonsensical', *időjárási* 'meteorological', *reménytelen* 'hopeless', *kritikus* 'critical', *bizonytalan* 'insecure', *vagyoni* 'financial', *lehetetlen* 'impossible'] – the only thing they have in common is the subsequent noun, *helyzet* ('situation'). Considering the fact that this approach tends to detect more interesting examples–in the above defined sense–more effort needs to be devoted to the investigation of the representations based on probability distributions.

4.3.2. Coarse-grained classification of cliques

2.a *Narrow semantic classes*

One problem we had to face during the evaluation phase is that not all adjectives were equally interesting from a meaning discrimination perspective. For example, dates and measures did not exhibit any interesting properties in most cases, even if they were assigned to multiple cliques. Instead, adjectives from these tight semantic classes tended to belong to multiple cliques with the very same meaning. According to our hypothesis, due to their varying sizes and varying distances between the elements, these adjectives cannot be grouped into one clique in a coherent way, no matter what the parameter setting is. Another reason to disregard adjectives from these tight semantic classes is that their lexical meaning seems to be rather straightforward not allowing for polysemy, except for a handful of more complex ones (eg. *fekete* 'black', *fehér* 'white', *szürke* 'gray'). For instance, *hétfői* 'of.Monday' was grouped under two different cliques:

Ex. 4 Clique 1: *hétfői* 'of.Monday', *pénteki* 'of.Friday', *szombati* 'of.Saturday', *vasárnapi* 'of.Sunday'
Clique 2: *hétfői* 'of.Monday', *tegnapi* 'of.yesterday', *keddi* 'of.Tuesday', *csütörtöki* 'of.Thursday', *szerdai* 'of.Wednesday', *szombati* 'of.Saturday', *pénteki* 'of.Friday'

The following nouns did not supply enough evidence to accept the meaning discrimination indicted by the cliques: numerals, dates, names of colors, units of measurements and various national currencies belonged to this category.

2.b *Named entities*

Another class of adjectives was made up of named entities, primarily countries, cities and surnames. In spite of the rather striking results, they were not considered in the present investigation, since our main focus is lexical meaning here, while the clique-membership of NEs tend to reflect factual knowledge rather than lexical meaning. For instance, *egri* (related to the city of Eger) was assigned to two cliques [*egri*, *soproni*, *veszprémi*] (related to the cities of Eger, Sopron and Veszprém, respectively) indicating viticultural areas, whereas the other clique [*egri*, *esztergomi*] (related to the cities of Eger and Esztergom, respectively) are referring to archdioceses.

One interesting finding of the manual evaluation was that the 6*k* window size word2vec representation was rather efficient in the detection of tight semantic classes and cliques of named entities: out of the 446 adjectives 99 belonged to some types of named entities, 28 adjectives were terms of measurements, while 11 adjectives were assigned to at least two cliques referred to numerals.

2.c *Emotive intensifiers*

We found that emotive intensifiers tend to group in cliques not conveying separate meanings. For example:

Ex. 5 Clique 1: *borzalmas* ‘terrible’, *iszonyatos* ‘terrific’, *rettenetes* ‘awful’
szenvedés ‘suffering’, *kép* ‘picture’, *körülmény* ‘circumstance’
Clique 2: *borzalmas* ‘terrible’, *félelmetes* ‘dreadful’, *rettenetes* ‘awful’, *szörnyű* ‘horrible’
látvány ‘spectacle’, *nap* ‘day’, *érzés* ‘feeling’
Clique 3: *borzalmas* ‘terrible’, *borzasztó* ‘terrifying’, *rettenetes* ‘awful’, *szörnyű* ‘horrible’, *rémes* ‘fearful’
emlék ‘memory’, *élmény* ‘experience’
Clique 4: *borzalmas* ‘terrible’, *szörnyűséges* ‘eldritch’, *rettenetes* ‘awful’, *szörnyű* ‘horrible’, *rémes* ‘fearful’
történet ‘story’

While the cliques imply that negative emotive intensifiers form a coherent semantic class among adjectives, neither the cliques nor the following nouns do not supply enough evidence to discriminate between the meaning of cliques.

2.d *nagy* ‘great’

The adjective *nagy* ‘great’ and related notions, such as *óriási* ‘huge’, *hatalmas* ‘large’, etc, are posing another problem: here the abstraction step is quite easy to make along the various dimensions, therefore, in this case, lumping the sub-meanings indicated by the cliques may be a motivated choice. For example, *óriási* belongs to two different cliques characterized by plenty of nouns:

Ex. 6 Clique 1: *óriasi* ‘huge’, *nagy* ‘great’, *hatalmas* ‘large’
mosoly ‘smile’, *oroszlán* ‘lion’, *roham* ‘attack’, *piramis* ‘piramid’, etc.
Clique 2: *óriási* ‘huge’, *komoly* ‘serious’
kaland ‘adventure’, *konkurencia* ‘concurrence’, *kérdés* ‘question’, *lemaradás* ‘lag’, *marketing* ‘marketing’, *infláció* ‘inflation’, etc.

However, although *komoly* ‘serious’ cannot be used as a synonym of ‘huge’ before the elements of the first clique (eg. *komoly mosoly* (‘a serious smile’) ≠ *óriási mosoly* (‘a huge smile’) and *komoly oroszlán* (‘a serious lion’) ≠ *óriási oroszlán* (‘a giant lion’)), someone may claim that – in certain contexts at least – *óriási* and *komoly* conveys the same meaning at a certain level of abstraction. We confine ourselves only to make a notice on this phenomenon in the present paper and do not want to take a definite stance on this question.

4.3.3. Clique validation rules

3.a If the list of nouns specific to a given clique consists of only one noun, it is not sufficient to characterize the given clique. Example 7 shows the cliques of *nívós* ‘of high standard’. The first clique, [*nívós* 'of high standard', *színvonalas* 'of high quality'] is represented by a long list of nouns. The second clique, [*nívós* 'of high quality', *rangos* 'prestigious'], on the other hand, has only one noun specific to it, *kiállítás* 'exhibition'. Disregarding these nouns, the two cliques seem to be ideal; there is a clear difference between *színvonalas* 'of high quality' and *rangos* 'prestigious': the former one states something about the quality of the given noun, it is excellent, fine, superior. The latter one, on the other hand, places the noun on a scale of quality where quality is measured and expressed by a rank, a prize, an honor or some other kind of accolade. However, *kiállítás* 'exhibition' in itself does not sufficiently circumscribe this latter clique; with more nouns specific to [*nívós* 'of high standard', *rangos* 'prestigious'] one could easily evaluate these cliques (and create sentence pairs for a Hungarian meaning discrimination corpus).

Ex. 7 Clique 1: *nívós* 'of high standard', *színvonalas* 'of high quality'
szolgáltatás 'service', *étterem* 'restaurant', *program* 'program', *szálloda* 'hotel', *műsor* 'show', *előadás* 'performance', *koncert* 'concert', *képzés* 'training', *dolog* 'thing'
Clique 2: *nívós* 'of high standard', *rangos* 'prestigious'
kiállítás 'exhibition'

3.b *Dolog* 'thing' proves to be a rather vague noun: it appears with many cliques but does not add anything to the characteristics of the groups. It appears in Example 7 with *nívós* as well, but Example 8 shows an even better example for its behavior. The two cliques seem to be valid, *illetlen* 'inappropriate' is a more general term with a less specific meaning than *trágár* 'swinish' or *vulgáris* 'vulgar', which have a more specific and well-definable meaning: they involve scurrility, the use of filthy words. However, the former group, [*illetlen* 'inappropriate', *obszcén* 'obscene'] are not well characterized based on their nouns in the corpus; only *dolog* 'thing' is specific to them, which definitely does not represent a semantically specific noun class.

Ex. 8 Clique 1: *obszcén* 'obscene', *illetlen* 'inappropriate'
dolog 'thing'
Clique 2: *obszcén* 'obscene', *trágár* 'swinish', *vulgáris* 'vulgar'
kifejezés 'expression', *szöveg* 'text'

3.c The qualitative evaluation involved the search for counter-examples: we checked whether the adjectives of a clique certainly do not fit with the nouns of another: Example 9 shows the case of *parádés* 'superb'. Although there seem to be some coherent noun classes there (sport-related nouns, performances and movies, etc.), when trying to find counter-examples, we see that *bravúros* 'brilliant' from the first clique goes quite well with *szereposztás* 'casting' of the third clique (we support our introspection with simple Google queries): it is only an accidental gap that they do not co-occur in our corpus. Therefore these groups of nouns are not specific to these cliques, and thus, they do not provide us with sufficient evidence to draw the meaning distinction between the corresponding cliques.

Ex. 9	Clique 1:	*parádés* ‘superb’, *bravúros* ‘brilliant’
		mozdulat ‘move’, *hajrá* ‘finish’
	Clique 2:	*parádés* ‘superb’, *szenzációs* ‘sensational’
		sorozat ‘series’, *ötlet* ‘idea’, *alakítás* ‘performance’, *a*[11] ‘the’, *akció* ‘action’
	Clique 3:	*parádés* ‘superb’, *pazar* ‘magnificent’
		szereposztás ‘casting’, *sarkazás* ‘backheel’

3.d In many cases, seemingly valid cliques are not supported by the nouns. Example 10 shows the cliques of *királyi* ‘royal’: the first clique represents its sense as ‘sovereign’, ‘ruler’ of a country; the second one is concentrated around nobility, ranks and titles of nobility. However, the nouns do not support the distinction of these senses: we find many counter-examples (*császári kastély* and *hercegi címer* are just two trivial examples), and no clear semantic group of nouns specific to either clique can be drawn based on these lists. Thus, just as in the case of Example 9 we do not have enough empirical background to tell the meanings apart.

Ex. 10	Clique 1:	*királyi* ‘royal’, *uralkodói* ‘sovereign’, *császári* ‘imperial/caesarean’
		központ ‘centre’, *jog* ‘right’, *címer* ‘coat of arms’
	Clique 2:	*királyi* ‘royal’, *grófi* ‘of.earl’, *hercegi* ‘of.prince’
		uradalom ‘lordship’, *kastély* ‘castle’, *korona* ‘crown’, *birtok* ‘estate’, *rang* ‘rank’

5. RESULTS IN THE LIGHT OF THE HUWIC CORPUS

The main objective of our research was to create sentence pairs for a Hungarian WiC corpus, to the maximum extent possible, automatically. This means that we would like to extract sentence pairs with adjective-noun pairs in them but with a strong presupposition of whether the adjective appears in the two sentences in the same sense or not. This presupposition is derived from the validated adjectival cliques described in the previous sections of this paper. In other words, if we choose two sentences for an adjective where the nouns following it are both specific to the same clique of that adjective, then our presupposed label for that sentence pair is 1: the two contexts convey the same meanings of that adjective. On the other hand, if we choose sentences with nouns from different cliques, then the adjective appears with the different senses characterized by the adjectives of the different cliques. Example 11 shows two possible instances

[11] The appearance of the definite article *a* is the consequence of an erroneous part-of-speech tagging in the corpus we use.

of the adjective *napfényes* 'sunny, sun-drenched'.[12] In the first instance, the nouns following the adjective in the example sentences are taken from the first clique of *napfényes*. In the second instance, the nouns are taken from different cliques, one from the first, and the other from the second clique.

Ex. 11 Clique 1: *napfényes* 'sunny', *napsütéses* 'sunshiny'
vasárnap 'Sunday', *nap* 'day'
Clique 2: *napfényes* 'sunny', *napsütötte* 'sunlit'
terület 'area', *sziget* 'island', *oldal* 'side', *terasz* 'terrace'

1st instance:

Sentence1: *Egy szép* ***napfényes vasárnapon*** *megérkezett Budára Mátyás király.*

'On a **sunny Sunday** King Matthias arrived at Buda.'

Sentence2: *Ragyogó,* ***napfényes napok*** *várnak ránk a Balatonnál.*

'Bright, **sunny days** await us at Lake Balaton.'

Label: 1 (same sense)

2nd instance:

Sentence1: *Egy-egy melegebb,* ***napfényes nap*** *kerti munkára csábít.*

'A warm, **sunny day** invites you to work in the garden.'

Sentence2: *A nappaliból és a szobákból a nagy,* ***napfényes teraszra*** *jutunk ki.*

'The living room and bedrooms open onto a large, **sunny terrace**.'

Label: 0 (different senses)

As these cliques are ideal in every way, we can randomly choose from the following nouns to extract the example sentences. In some cases, however, as mentioned before, we first may need some manual evaluation to narrow down the list of nouns and extract nouns that do not fit well in the given clique. Naturally, the instances will be annotated by human annotators, and their final label will be set based on the human decisions and our initial label. Therefore, if the cliques are erroneous, a false initial label will still be corrected by the decisions of the human annotators.

Pilehvar & Camacho-Collados (2019) applied some extra criteria when compiling the dataset: they did not sample more than three instances for the same target word, and they filtered the sentences so no example sentence is present in more instances. Similarly, we may apply the following constraints for our dataset:

1. We do not use more than five[13] instances of the same target word.
2. We do not use the same adjective-noun pair in more than one instance.
3. We do not use the same example sentence more than once.

[12] The extraction of good example sentences requires further discussion.

[13] We may decrease this number as more and more suitable adjectival cliques are gathered.

Based on the word2vec representation of adjectives occurring at least 200 times ($K = 0.7$, $Freq\_n = 1$) we found 55 adjectives having at least two validated cliques with at least two nouns in those cliques. This collection is sufficient to create a corpus of 250–300 instances where one target word does not appear more than 5 times, and for each instance we choose different example sentences. If we allow a target word to have more than 5 instances in the dataset then we can easily create 500–700 instances, as there are some cliques with numerous (validated) nouns specific to them.

It is important to note that in accordance with our basic objective, we concentrated on the lexical meaning of adjectives, therefore, adjectives with referential meanings were omitted from our list, such as named entities, measures, currencies, nationalities, names of colors, days, etc. In accordance with Section 4.3.2 we did not consider emotive intensifiers either. Thus, cc. 200 adjectives were examined during the evaluation. We found that the majority of cliques were deemed to be wrong due to the quality of the retrieved context nouns. However, this can be improved by relaxing the relevant parameters.

The manually validated cliques are available on GitHub.[14]

6. CONCLUSIONS AND FUTURE WORK

This first attempt to discriminate adjectival meanings based on surface distributional data has yielded rather promising results. Even though the number of retrieved senses is one-magnitude smaller than the targeted number, the discovered senses tend to be rather enlightening. The methodology, due to its corpus-driven nature, is able to shed light on meaning distinctions which are difficult to grasp based on human intuition and possibly even in a corpus-based framework, as they are not salient for human perception. Moreover, not only meaning discrimination is performed in an unsupervised way, but relevant contexts are also automatically retrieved. We think that anchoring meaning discrimination task to inter-subjective observable data is of primary importance, as this task is rather difficult even for humans, thus, related databases and benchmarks may not be consistent from this perspective.

Accordingly, our future research plan aims primarily at increasing the coverage of adjectival lexemes. For doing so, we have multiple opportunities:

1. To rely on less strict parameters concerning adjectival frequency, the K cut-off parameter and the overall frequency of the following noun.
2. We also intend to consider the cliques yielded by the probability distribution-based approaches, which seem to cover different types of adjectives than the $6k$ window size word2vec representation. At the present state of research this technique tends to provide less strict cliques and more cliques with empty sets of nouns, still the meaning distinctions seem to be more interesting for our purposes.
3. An additional possibility is to use word2vec representations of $2k$ context window. As Levy & Goldberg (2014) points out, while a window size of 5 captures the broad topical content of the target word, a smaller window size may result in more focused information about it. Their quantitative evaluation, based on the WordSim353 dataset (Agirre et al. 2009) showed

[14] https://github.com/nytud/HuWiC

that BoW2 (the parameter setting with a $2k$ window size) is more likely to capture similarity above relatedness than BoW5 (the parameter setting with an $5k$ window size).

4. Another thread of research is related to the detection of relevant nominal semantic classes among the validating nouns.

REFERENCES

Agirre, Eneko, Enrique, Alfonseca, Keith, Hall, Jana, Kravalova, Marius, Paşca and Aitor, Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09). 19–27.

Ah-Pine, Julien and Guillaume, Jacquet. 2009. Clique-based clustering for improving named entity recognition systems. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). 51–59.

Amplayo, Reinald Kim, Seung-won, Hwang and Min, Song. 2019. AutoSense model for word sense induction. Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19). 6212–6219.

Amrami, Asaf and Yoav, Goldberg. 2019. Towards better substitution-based word sense induction. ArXiv. abs/1905.12598.

Atkins, B. T. Sue and Michael, Rundell. 2008. The Oxford guide to practical lexicography. International Journal of Lexicography 22(1). 94–102.

Bartunov, Sergey, Dmitry, Kondrashkin, Anton, Osokin and Dmitry, Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In A. Gretton and C.C. Robert (eds.) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 51). Cadiz: PMLR. 130–138.

Başkaya, Osman, Enis, Sert, Volkan, Cirik and Deniz, Yuret. 2013. AI-KU: Using substitute vectors and Co-occurrence modeling for word sense induction and disambiguation. Second Joint Conference on Lexical and Computational Semantics (*SEM), Vol. 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). 300–306.

Bojanowski, Piotr, Edouard, Grave, Armand, Joulin and Tomas, Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5. 135–146.

Borbély, Gábor, Márton, Makrai, Dávid Márk, Nemeskey and András, Kornai. 2016. Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. 83–89.

Devlin, Jacob, Ming-Wei, Chang, Kenton, Lee and Kristina, Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). 4171–4186.

Döbrössy, Bálint, Márton, Makrai, Balázs, Tarján and György, Szaszák. 2019. Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian. Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). 187–193.

Fellbaum, Christiane (ed.). 1998. WordNet: An electronic lexical database (Language, Speech, and Communication). Cambridge, MA: MIT Press.

Ficsor, Tamás and Gábor, Berend. 2021. Analysing the semantic content of static Hungarian embedding spaces. XVII. Magyar Számítógépes Nyelvézeti Konferencia (MSZNY2010). 91–105.

Gábor, Kata and Enikő, Héja. 2007. Clustering Hungarian verbs on the basis of complementation patterns. Proceedings of the ACL 2007 Student Research Workshop. 91–96.

Harris, Zellig S. 1954. Distributional structure. WORD 10(2–3). 146–162.

Héja, Enikő and Dávid, Takács. 2010. Melléknevek szűk szemantikai osztályainak detekciója a Magyar nemzeti szövegtárban jelentésegyértelműsítés céljából [The detection of narrow semantic classes of adjectives in the Hungarian National Corpus for the purpose of sense disambiguation]. VII. Magyar Számítógépes Nyelvézeti Konferencia (MSZNY2010). 360–362.

Ide, Nancy, Collin, Baker, Christiane, Fellbaum, Charles, Fillmore and Rebecca, Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. 2455–2460.

Kiefer, Ferenc. 2008. A melléknevek szótári ábrázolásáról (On the lexical representation of adjectives). In: Kiefer F. (ed.) Strukturális magyar nyelvtan. 4. A szótár szerkezete. Budapest: Akadémiai Kiadó. 505–538.

Komninos, Alexandros and Suresh, Manandhar. 2016. Dependency based embeddings for sentence classification tasks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 1490–1500.

Kuti, Judit, Enikő, Héja and Bálint, Sass. 2010. Sense disambiguation – 'Ambiguous sensation'? Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). 23–30.

Landes, Shari, Claudia, Leacock and Christiane, Fellbaum. 1998. Building semantic concordances. In C. Fellbaum (ed.) WordNet: An electronic lexical database. Cambridge, MA: MIT Press. 199–216.

Lau, Jey Han, Paul, Cook and Timothy, Baldwin. 2013. unimelb: Topic modelling-based word sense induction for web snippet clustering. Second Joint Conference on Lexical and Computational Semantics (*SEM), Vol. 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). 217–221.

Levy, Omer and Yoav., Goldberg. 2014. Dependency-based word embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 2: Short Papers. 302–308.

Li, Jiwei and Dan, Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 1722–1732.

Lipp, Veronika and László, Simon. 2021. Towards a new monolingual Hungarian explanatory dictionary: Overview of the Hungarian explanatory dictionaries. Studia Lexicographica 15. 83–96.

Makrai, Márton. 2015. Comparison of distributed language models on medium-resourced languages. XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). 22–33.

Makrai, Márton and Veronika, Lipp. 2017. Do multi-sense word embeddings learn more senses? In K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS). 1–8.

Mikolov, Tomas, Ilya, Sutskever, Kai, Chen, G.s, Corrado and Jeffrey, Dean. 2013. Efficient estimation of word representations in vector space. CoRR. abs/1301.3781.

Mikolov, Tomas, Kai, Chen, Greg, Corrado and Jeffrey, Dean. 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26(10).

Navigli, Roberto and Simone Paolo, Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193. 217–250.

Neelakantan, Arvind, Jeevan, Shankar, Alexandre, Passos and Andrew, McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 1059–1069.

Nemeskey, Dávid Márk. 2020. Natural Language Processing Methods for Language Modeling. Doctoral dissertation. Eötvös Loránd University, Budapest.

Novák, Attila and Borbala, Novák. 2017. A dologfelismerő (The thing recognizer). In: Vincze, V. (Ed.), XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). Institute of Informatics, University of Szeged, Szeged, pp. 25–36.

Oravecz, Csaba, Tamás, Váradi and Bálint, Sass. 2014. The Hungarian Gigaword Corpus. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 1719–1723.

Pasini, Tommaso and Jose, Camacho-Collados. 2020. A short survey on sense-annotated corpora. Proceedings of the 12th Language Resources and Evaluation Conference. 5759–5765.

Pilehvar, Mohammad Taher and Jose, Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). 1267–1273.

Ploux, Sabine and Bernard, Victorri. 1998. Construction d'espaces semantiques a l'aide de dictionnaires de synonymes. Traitement automatique des langues 1(39), 146–162.

Rehurek, Radim and Petr, Sojka. 2011. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3(2).

Schuler, Karin Kipper. 2006. VerbNet: A broad-coverage, comprehensive verb lexicon. Doctoral dissertation. University of Pennsylvania, Philadelphia, PA.

Siklósi, Borbála. 2016. Using embedding models for lexical categorization in morphologically rich languages. Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016. 115–126.

Siklósi, Borbála and Attila, Novák. 2016. Beágyázási modellek alkalmazása lexikai kategorizációs feladatokra XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016). 3–14.

Sinclair, John. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, John. 1998. The lexical item. In E. Weigand (ed.) Contrastive lexical semantics. Amsterdam: Benjamins. 1–24.

Szántó, Zsolt, Veronika, Vincze and Richárd, Farkas. 2018. Magyar nyelvű szó- és karakterszintű szóbeágyazások [Hungarian word-level and character-level embeddings]. XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). Original document in Hungarian. 323–328.

Taghipour, Kaveh and Hwee Tou, Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. Proceedings of the Nineteenth Conference on Computational Natural Language Learning. 338–344.

Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. Journal of Artificial Intelligence Research 44. 533–585.

Váradi, Tamás. 2002. The Hungarian national corpus. Proceedings of the Second International Conference on Language Resources and Evaluation. 385–389.

Véronis, Jean. 2003. Sense tagging: Does it make sense? In A. Wilson, P. Rayson and T. McEnery (eds.) Corpus linguistics by the Lune: A Festschrift for Geoffrey Leech. Frankfurt: Peter Lang. 273–290.

Wang, Alex, Yada, Pruksachatkun, Nikita, Nangia, Amanpreet, Singh, Julian, Michael, Felix, Hill, Omer, Levy and Samuel R. Bowman. 2020. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

Wang, Jing, Mohit, Bansal, Kevin, Gimpel, Brian D, Ziebart and Clement T, Yu. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. Transactions of the Association for Computational Linguistics 3. 59–71.

Zhang, Daniel, Saurabh, Mishra, Erik, Brynjolfsson, John, Etchemendy, Deep, Ganguli, Barbara, Grosz, Terah, Lyons, James, Manyika, Juan Carlos, Niebles, Michael, Sellitto, Yoav, Shoham, Jack, Clark and Raymond, Perrault. 2021. The AI index 2021 annual report. CoRR. abs/2103.06312.