2021. *Proc Ling Soc Amer* 6(1). 54–68. https://doi.org/10.3765/plsa.v6i1.4941.

The role of verb semantics in Hungarian verb-object order

Dorottya Demszky*

Abstract. Hungarian is often referred to as a discourse-configurational language, since the structural position of constituents is determined by their logical function (topic or comment) rather than their grammatical function (e.g., subject or object). We build on work by Komlósy (1989) and argue that in addition to discourse context, the lexical semantics of the verb also plays a significant role in determining Hungarian word order. In order to investigate the role of lexical semantics in determining Hungarian word order, we conduct a large-scale, data-driven analysis on the ordering of 380 transitive verbs and their objects, as observed in hundreds of thousands of examples extracted from the Hungarian Gigaword Corpus. We test the effect of lexical semantics on the ordering of verbs and their objects by grouping verbs into 11 semantic classes. In addition to the semantic class of the verb, we also include two control features related to information structure, object definiteness and object NP weight, chosen to allow a comparison of their effect size to that of verb semantics. Our results suggest that all three features have a significant effect on verb-object ordering in Hungarian and among these features, the semantic class of the verb has the largest effect. Specifically, we find that stative verbs, such as *fed* 'cover', *jelent* 'mean' and *övez* 'surround', tend to be OV-preferring (with the exception of psych verbs which are strongly VO-preferring) and non-stative verbs, such as *bírál* 'judge', *csökkent* 'reduce' and *csókol* 'kiss', verbs tend to be VO-preferring. These findings support our hypothesis that lexical semantic factors influence word order in Hungarian.

Keywords. Hungarian; lexical semantics; computational linguistics; corpus analysis

1. Introduction. [1]

Hungarian is often referred to as a free word-order language, given that grammatical functions, such as subject and object, are not linked to invariant structural positions in the sentence (É. Kiss 2002; p. 2). For example, all six permutations of the transitive verb *szeret* 'love', subject *Józsi* 'Joe' and object *Sárit* 'Sarah-ACC' are possible in principle, as shown in (1).

(1) a. Józsi szereti Sárit.
Joe loves Sarah-ACC
'Joe loves Sarah.'
b. Józsi Sárit szereti.
c. Sárit szereti Józsi.
d. Sárit Józsi szereti.
e. Szereti Sárit Józsi.
f. Szereti Józsi Sárit.

* I would like to thank the three supervisors of this project: Beth Levin, Dan Jurafsky and László Kálmán, as well as András Komlósy and Peter Henderson for helpful conversations, and Iván Mittelholtz, Tamás Váradi and Gábor Prószéky for providing access to the Hungarian Gigaword Corpus. Author: Dorottya Demszky, Stanford University (ddemszky@stanford.edu).

[1] Please find the Supplementary Material at https://ddemszky.github.io/hungarian_verbs/supplement.pdf

Despite the apparent freedom of word order, it is widely accepted that information structure plays a major role in determining the word order of a Hungarian sentence (see É. Kiss 1978; É. Kiss 1981; É. Kiss 2002; Kálmán et al. 1989; Gecség & Kiefer 2009). For example, only orders (1b) and (1c), where 'Sarah' is in the preverbal focus position, can be used in felicitous replies to the question 'Who does Joe love?'. Examples such as these suggest that the discourse context is crucial for determining the felicity of particular word orders.

In this paper we argue that, in addition to information structure, the lexical semantics of the verb also plays a significant role in determining Hungarian word order. Our hypothesis builds on the work of Komlósy (1989), who presents evidence that certain verbs and their objects show systematic ordering preferences in Hungarian across discourse contexts. His findings, involving approximately 80 verbs, suggest that lexical semantic factors such as verb choice might also be at play in determining Hungarian word order. While the nature of this lexical influence has not yet been investigated, there are semantic similarities among verbs that according to Komlósy's study share the same ordering preference. For example, many verbs he identifies as having a preference to follow their objects (OV-preferring) are psych verbs (e.g., *utál* 'hate', *tud* 'know', *emlékszik* 'remember') and many verbs he identifies as OV-preferring express states or spatial configuration (e.g., *tartalmaz* 'contain', *marad* 'remain'). Thus, we hypothesize that the systematic ordering preference of verbs is driven by lexical semantic factors.

In order to investigate the role of lexical semantics in determining Hungarian word order, we conduct a large-scale, data-driven analysis on the ordering of 380 transitive verbs and their objects, as observed in hundreds of thousands of examples extracted from the Hungarian Gigaword Corpus (Oravecz et al. 2014). We use empirical methods, as they allow us to verify and estimate the significance of word order tendencies using a large corpus of natural language. Such methods have not been used previously to study Hungarian word order, so we draw on works in other languages (Bresnan et al. 2007; Benor & Levy 2006), which seek to estimate the influence of different factors on word order patterns via a logistic regression model trained on a large corpus. We restrict our analysis to transitive verbs and their objects to reduce the number of confounds arising from verbs with multiple arguments in the verb phrase.

We test the effect of lexical semantics on the ordering of verbs and their objects by grouping verbs into semantic classes. We identify 11 semantic classes based on salient semantic patterns in our training data as well as in previous literature on verb classes (Levin 1993), and we use these classes as features in our analysis. In addition to the semantic class of the verb, we also include two control features related to information structure, chosen to allow a comparison of their effect size to that of verb semantics. These features are object definiteness and object NP weight. Komlósy (1989) has noted the influence of definiteness on verb-argument ordering, while the influence of constituent weight on word order has been noted for other languages (Wasow 1997). We obtain object definiteness and object NP weight automatically, using natural language processing tools for Hungarian (Qi et al. 2018; Trón et al. 2005). We construct a logistic regression model to study the effect of these three features (semantic class of the verb, object definiteness and object NP weight) on the ordering of verbs and objects in our data.

Our results suggest that the features we used have a significant effect on verb-object ordering in Hungarian and among these features, the semantic class of the verb has the largest effect. Specifically, we find that stative verbs tend to be OV-preferring (with the exception of psych verbs which are strongly VO-preferring) and non-stative verbs verbs tend to be VO-preferring. These findings support our hypothesis that lexical semantic factors influence word

order in Hungarian.

The organization of the paper is as follows. In Section 2, we lay out our research questions, which we group into three topically related sets. In Section 3, we present background literature on Hungarian. We then describe our methods for answering the research questions in Section 4: the data and the pre-processing (Section 4.1), the verb classification (Section 4.2) and other, object-related features we used (Section 4.3) and finally, the logistic regression model (Section 4.4). We return to the research questions again in Section 5, where we present and discuss our results. In our conclusion (Section 6), we summarize our work and discuss avenues for future research.

2. Research Questions. In Section 1, we introduced our main motivation for focusing on the role of three features on Hungarian verb-object order. Here we provide our three sets of topically-related research questions and a glance at the methods we use to address them.

1. *Ordering Preference of Verbs*
   How well can we predict the ordering of verbs and their objects in our data based on the verb exclusively? Which transitive verbs have an OV preference, which ones have a VO preference, and which show no preference?

   Methods: We extract verb-object pairs from our corpus and run a logistic regression model to predict their ordering based on the lemma of the verb exclusively.

2. *Verb Classes*
   Can we identify a small set of semantic classes for the verbs, which are salient based on previous literature and based on the semantic features that seem relevant for the verbs' ordering preference? How well can we predict the ordering of verbs and objects based on the semantic class that the verb belongs to? Which verb classes have a VO preference, which ones have an OV preference and which none?

   Methods: We identify a set of semantic classes and assign verbs to them. We then run a logistic regression model to predict the ordering of verbs and their objects based on the verb's semantic class. During prediction, we test on previously unseen data that we did not look at while identifying the verb classes.

3. *Importance of Object-Related Features*
   Besides the verb's semantic class, how well can object-related features, and specifically the definiteness of the object and the weight of the object NP, predict the relative ordering of the verb and the object?

   Methods: Extract the object-related features automatically from the corpus, for each verb-object pair. Run a logistic regression model to predict the ordering of verbs and objects based on each of these features separately.

3. Background on Hungarian. In this section, we discuss a few aspects of Hungarian that have been investigated in previous work and relate to word order. These aspects include the discourse-configurationality of Hungarian (Section 3.1), object definiteness (Section 3.2), noun incorporation (Section 3.3), the information content of verbs (Section 3.4) and prosodic classification of verbs (Section 3.5).

3.1. DISCOURSE-CONFIGURATIONALITY. As we mentioned, it is widely accepted that in Hungarian, the syntactic structure of a sentence is determined by its information structure — in other words, Hungarian is considered by many to be a discourse-configurational language (É. Kiss 1995), even though there are disagreements as to what this term entails (Gecség & Kiefer 2009). Reviewing the extensive literature on Hungarian syntax is beyond the scope of this paper. Here we provide a brief overview of some of the key properties of Hungarian syntax that are relevant to the study of verb-object ordering.

At a high level, a Hungarian sentence is divided into a logical subject (which some call a topic) and a logical predicate (which some call a comment), in this order. The preverbal part of the predicate, following the logical subject, is also called the "focus field" (Brody 1990) and it is reserved for the focus constituent. A variety of different elements can occupy the focus position, including preverbal particles, negative quantifiers, bare NPs and focused definites and indefinites. Generally, these elements are considered to be in complementary distribution (Gecség & Kiefer 2009), but there is evidence suggesting that their patterning is more complex (Farkas 1986; É. Kiss 2002). Related to the positions of different grammatical elements in the sentence, there is also disagreement as to whether Hungarian has a neutral word order (SVO) (Kiefer 1967; Kálmán et al. 1986; Marácz 1989) or whether there is no neutral word order in Hungarian (É. Kiss 1981; É. Kiss 2002). Given the close and complex interaction between semantics, pragmatics and word order in Hungarian, many aspects of the theory of Hungarian syntax (e.g., the definition of a logical subject or what it means for a sentence to be have "neutral" word order) are much contested.

In this paper, we study the relative position of the verb and its object, given that the work of Komlósy (1989) suggests that the verb plays a role in the word order of the sentence. The most relevant aspect of Hungarian syntactic theory to us is that when the object is preverbal, it can either be in the topic or focus position. The examples[2] in ((2)) show the possible positions that the grammatical object *Sárit* 'Sarah-ACC' can occupy. The object is preverbal both when it is the topic of the sentence ((2)a) and when it is focused ((2)b) and it is postverbal when it is neither focused nor topicalized ((2)c).

(2) a. [$_{\text{Topic}}$ Sárit] [$_{\text{Pred}}$ elütötte egy autó]
[Sarah-ACC] [hit a car]
'A car hit Sarah. / Sarah was hit by a car.'

b. [$_{\text{Topic}}$ Józsi] [$_{\text{Pred}}$ [$_{\text{Focus}}$ Sárit] szereti]
[Joe][[Sarah-ACC] loves]
'It is Sarah whom Joe loves.'

c. [$_{\text{Topic}}$ Józsi] [$_{\text{Pred}}$ szereti Sárit]
[Joe] [loves Sarah-ACC]
'Joe loves Sarah.'

In our analyses, we do not distinguish between topicalized and focused objects, nor between their contrastive vs. non-contrastive subcategories (another distinction that is beyond the scope of this work). Identifying the discourse function would require manual labeling, since there is no reliable automatic method for doing so. Moreover, these discourse functions can be challenging to even label manually based on text, since intonation plays a key role in their in-

[2] The examples are based on the ones presented in different chapters of É. Kiss (2002).

terpretation (É. Kiss 2002). Therefore, we do not incorporate discourse function into our analyses, which is a limitation of this work. However, we believe that studying the ordering preference of verbs, even without knowing the information structure of the sentence, is valuable as we can learn whether and to what extent verb meaning has an influence on word order that persists across discourse contexts.[3]

3.2. OBJECT DEFINITENESS. In Hungarian, the definiteness of an object is closely linked to its discourse function, with definite objects being more likely to represent known information and indefinite objects being more likely to represent new information. Therefore, the definiteness of the object can be predictive of its syntactic position. Specifically, since the focus position is reserved for non-presupposed information (É. Kiss 2002), indefinite objects are more likely to occupy this slot than definite ones. In contrast, definite objects are more likely to follow the verb, or occasionally to be topicalized. Even though the close relationship between the definiteness and the syntactic position of objects is assumed and/or mentioned in the literature on Hungarian syntax, the extent to which object definiteness is predictive of verb-object ordering has not been studied. We incorporate object definiteness as a feature into our predictive model to compare its effect to that of verb semantics.

Hungarian is well known for marking the definiteness of the object on the verb. Specifically, Hungarian has two inflectional paradigms, the objective and subjective conjugations. The objective conjugation signals the presence of a definite object. These conjugations are relevant for us because they allows us to detect the definiteness of the object based on the morphological features of the verb.

3.3. NOUN INCORPORATION. Bare objects, which constitute a special subcategory of indefinite objects, are in fact required to be in the preverbal focus position unless they is another element in contrastive focus. Such preverbal bare NPs are considered to be instances of noun incorporation and their syntactic behavior is very similar to that of preverbs, forming a semantic and intonational unit with the verb (Kiefer 1990; Farkas & Swart 2003). Since the relative ordering of bare nouns and verbs is not flexible, unlike the relative ordering of non-bare objects, we remove examples with bare objects from our analyses.

3.4. THE INFORMATION CONTENT OF VERBS. Given the close relationship between information structure and word order in Hungarian, it is natural to wonder whether the information content of the verb (in the particular context) plays an important role in Hungarian word order. For example, do verbs with relatively low information content (so called "light" verbs), such as *tesz* 'make/put' and *ad* 'give', tend to follow their objects and verbs with relatively high information content tend to precede their objects? To our knowledge, this question has not been empirically investigated, and we also leave this question for future work to address, given the reasons outlined below.

First, the information content of the verb is heavily context dependent. Factors that may influence the informativeness of the verb include the discourse context and the nature of the arguments (e.g., their grammatical function and semantic properties). The influence of the arguments is why previous work studies the lightness of verbs in particular *constructions* rather

[3] We believe that incorporating discourse context would only strengthen our findings since doing so would allow us to remove examples where the object is in the contrastive topic or the contrastive focus position, given that contrastive constructions are considered less "neutral" due to the added emphasis on the contrastive element.

than the lightness of verbs in isolation (Vincze & Csirik 2010).[4] As we discussed in Section 1, we are interested in the influence of verb meaning on verb-object ordering across contexts, and we do not consider context-dependent features in this study.

Second, due to our experimental design, our dataset only has a limited number of verbs that occur in light verb constructions identified by Vincze & Csirik (2010). This is because canonical examples of light verb constructions, as reported by Vincze & Csirik (2010), involve bare nouns and verbs. We remove such constructions from our analyses for the reasons described in Section 3.3. Moreover, several light verbs that occur in light verb constructions tend to take multiple arguments (e.g., *tesz* 'make/put' frequently occurs with directional arguments). We remove verbs that frequently take multiple arguments from our data to control for confounds (e.g., the influence of the other argument on the ordering of the object and the verb). Due to the limited presence of light verbs, our data is not best suited for studying the influence of verb information content on verb-object ordering.

Even though studying the influence of verb informativeness on verb-object ordering is beyond the scope of this current work, we believe that it is a potentially relevant factor and we plan to address this question in future work.

3.5. VERBS SHARING AN ORDERING PREFERENCE. As discussed in Section 1, our work builds on that of Komlósy (1989), who identifies systematic ordering preferences for a group of Hungarian verbs. Specifically, Komlósy describes the verbs' prosodic behavior, which maps onto their ordering preferences, given that verbs that are on the left edge of the predicate (preceding their objects) tend to carry the main sentential stress.[5] He found that certain verbs tend to avoid carrying the main sentential stress (e.g., *talál* 'find', *tartalmaz* 'contain', *marad* 'remain') while others tend to seek it (e.g., *utál* 'hate', *tud* 'know', *emlékszik* 'remember'). He classifies such verbs into two classes, which he calls stress-avoiding and stress-seeking verbs, respectively, and classifies verbs with no strong preference as regular verbs. Kálmán et al. (1986) further break down Komlósy's stress-seeking verbs and stress avoiding verbs into obligatorily stressed verbs, potentially stressed verbs, obligatorily unstressed verbs and potentially unstressed verbs.

Our classification of verbs into semantic classes is inspired by Komlósy (1989), since even though his classification is not motivated by semantics, there are apparent semantic similarities shared among verbs he groups together. For example, many stress-seeking verbs are experiencer-subject psych verbs and many stress-avoiding verbs express spatial configuration.

4. Methods.

4.1. DATA. In this section, we describe how we build our dataset of verb-object pairs. Our data comes from the Hungarian Gigaword Corpus (Oravecz et al. 2014), the largest, carefully curated multi-genre corpus of Hungarian containing 1.5 billion tokens. The distribution of tokens across genres and the source of the texts is summarized in Table 1. The corpus is tok-

[4] Vincze & Csirik (2010) define light verb construction based on Sag et al. (2002) as a type of lexicalized phrase or flexible expression that is neither idiomatic nor productive, and its meaning is not completely compositional. Noun + verb combinations, such as *bejelentést tesz* 'make an announcement', are one category of light verb constructions they recognize.

[5] In Section 3.1 we mentioned that there is a close relationship between the intonation of a sentence and its information structure. The main sentential stress in a Hungarian sentence falls on the first major constituent in the predicate. Therefore, if there is a preverbal argument, it carries the main sentential stress as opposed to the verb.

enized and morphologically analyzed via HunMorph (Trón et al. 2005), an FST-based parser that produces quite reliable annotations.

Genre	% of Tokens	Source
Journalism	42.0%	Daily / weekly newspapers
Personal	22.2%	Social media
Literature	14.5%	Digital Literary Academy
Official	8.8%	Documents from public admin.
(Popular) science	7.2%	Wikipedia, Hungarian Electronic Library
(Transcribed) spoken	5.4%	Radio programs

Table 1. Distribution of genres in the Hungarian Gigaword Corpus.

4.1.1. DEPENDENCY PARSING. To obtain verb-object pairs, we perform dependency parsing on the data. We experiment with four different parsers: *magyarlanc* (Zsibrita et al. 2013), a widely used toolkit for lemmatization, dependency parsing and morphological analyses in Hungarian; Hungarian models for SpaCy[6]; the Stanford NLP parser (Qi et al. 2018) and our own, rule-based parser. With the rule-based parser, we identify objects in the morphologically analyzed corpus by extracting nouns with accusative case that are only separated from the verb by adverb(s) and/or a preverb, or determiners and adjectives (if the noun is postverbal).

We randomly sample a list of 20 sentences from each genre in the corpus and manually compare the results of the four parsers. Specifically, we look at whether each parser accurately identifies verbs and their nominal objects. We find that the StanfordNLP parser and the rule-based parser are by far the best in terms of precision. However, the Stanford NLP parser outperforms the rule-based parser in terms of recall, which is expected, as it is able to identify longer-range dependencies. Given its performance, we used the Stanford NLP parser to extract verb-object pairs. Henceforth, we refer to each co-occurrence of a verb-object pair in the corpus as a TOKEN.

4.1.2. MORPHOLOGICAL ANALYSIS. The StanfordNLP parser generates part-of-speech tags and morphological analyses for the words, based on tags that are supported by the Szeged Dependency Treebank (Vincze et al. 2010).[7] As mentioned in Section 3.2, we use these definiteness tags on the verbs (`def` or `ind`) to determine the definiteness of the object. We find that the StanfordNLP parser detects definiteness marking on verbs with high accuracy, but unfortunately its lemmatization capability is very poor. Therefore, we re-lemmatize the verbs and objects using HunMorph (Trón et al. 2005), which often provides multiple possible lemmas. We use all possible lemmas (separated by a forward slash) to represent each verb and object. Henceforth, we refer to these combinations of possible lemmas simply as LEMMAS.

4.1.3. FILTERING. We perform multiple filtering steps to minimize the number of confounds in our data. Below we explain the goal and the details of each step. Before filtering, we start out with 17 million TOKENS, representing 35,838 unique verb LEMMAS, which includes verb LEMMAS prefixed with preverbs.

[6] https://github.com/oroszgy/spacy-hungarian-models
[7] https://universaldependencies.org/treebanks/hu_szeged/index.html

Preverbs. Preverbs occupy the preverbal slot, except when the sentence has contrastive focus, in which case they occur postverbally. Therefore, verbs with preverbs are constrained in their ordering with respect to their objects compared to other verbs. To eliminate the confound of preverbs having an influence on the ordering of verbs and their objects, we remove all verbs with preverbs. We detect verbs with preverbs via 1) morphological analysis, since preverbs can be prefixed to the verb, and 2) via dependency parsing (checking if there is a `compound:preverb` relation), in case the preverb is separated from the verb. Given the large number of preverb + verb combinations in Hungarian, this step results in the removal of 24,779 verb LEMMAS, so 69% of the LEMMAS we started out with. At the end of this step, our dataset contains 11,059 unique verb LEMMAS.

Bare objects. Bare objects – i.e. common nouns without modifiers – also behave in a special way in Hungarian, as they are considered to be part of incorporating constructions (see Section 3.3). The behavior of bare nouns is similar to preverbs, in that they always precede the verb in the absence of contrastive focus. Hence, we remove these objects from our analyses as well. We detect if a noun is bare by checking if it has zero dependents — we only keep a noun with zero dependents if it is a proper noun.

Transitivity. To ensure the verbs are transitive, we keep those verbs that occur with objects in the corpus at least 30% of the time. We acknowledge that this filtering step generates false negatives, as they might be transitive verbs that occur with overt objects less than 30% of the time. However, it was more important for us to ensure high precision than recall — i.e. to make sure that the verbs we find are indeed transitive. This filtering step results in the removal of 4391 verb LEMMAS, with 6668 verb LEMMAS remaining.

Multiple arguments. We also wanted to make sure that our results are not skewed by verbs that frequently take more than one argument (e.g., ditransitive verbs and verbs with locative / directional arguments). The reason for this is that these verbs might still have ordering preferences, but not with respect to their objects but with respect to one of their other arguments. Such verbs would introduce a confound in our analyses, since by only looking at the ordering of the verb and the object, we would not be able to know if there is an additional argument influencing their relative ordering. Thus, we filter out those transitive verbs that occur with either of the following dependency relations more than 50% of the time: `iobj`, `obl`, `nmod:obl`, `advmod:obl`, `amod:obl` and `ccomp:obl`. As a result of this step, we remove 3034 LEMMAS and have 3634 LEMMAS remaining.

In addition, we also remove all TOKENS where there is an oblique argument beside the object, even if the LEMMA does not occur with obliques more than 50% of the time. This step is to ensure that for the TOKENS we study, there is no additional argument influencing the verb-object ordering.

Frequency. To ensure that we have a large enough number of TOKENS for each verb for statistically robust analyses, we keep those verbs that occur at least 200 times in our data after the filtering steps. This way, even when we split the data into halves (Section 4.1.4), both halves will have about 100 TOKENS for each verb. These filtering steps yield 380 verb LEMMAS, which can be found in the Supplementary Material along with their frequencies in our training data (Section 4.1.4). Even though these verbs are all relatively frequent, there is still a discrepancy among their frequencies, as they follow a Zipfian distribution with a long tail (see

Figure 1 in the Supplementary Material). For example, the most frequent verbs, *ismer* 'know (someone)', *támogat* 'support' and *okoz* 'cause' are more than 20 times as frequent as the least frequent verbs *pontoz* 'score (a test)', *aktivizál* 'get (someone) to take action', *körvonalaz* 'outline' and *mormol* 'murmur'.

4.1.4. SPLITTING THE DATA. Our final dataset includes approximately 1.3 million TOKENS of verb-object pairs representing 380 verbs. To avoid overfitting to our data, we split it in two equal halves: a training set and test set. Both halves of the data contain at least 100 TOKENS for each verb. Following standard practice in machine learning, we use the training set to develop our model and we apply our finalized model to the test set.

4.2. VERB CLASSIFICATION. In order to study the effect of lexical semantics on verb-object ordering, we classify verbs into coarse-grained semantic categories. Each semantic class is created because 1) it represents a highly salient lexical semantic category or 2) because it represents a semantic distinction that seems relevant to ordering preference. To meet condition 1), we consult previous literature on the semantic classification of verbs (Levin 1993). To meet condition 2), we look at semantic distinctions between verbs with different ordering preferences in the training data to identify semantic features that clearly distinguish among verbs with different ordering preferences. It is important that we consider verbs' ordering preference only when we define semantic criteria for the verb classes, not when we assign the verbs to classes, so that we do not bias our experimental design.[8]

We define ten semantic classes, as well as a small OTHER class for 18 verbs that are polysemous or cannot be assigned to any of the categories. The ten classes are:

1. ACTIVITY (e.g., *keres* 'search for', *firtat* 'dwell on', *foglalkoztat* 'employ, occupy')
2. AFFECT (e.g., *tisztít* 'clean', *vereget* 'hit at', *sürget* 'urge')
3. CHANGE (e.g. *aktivál* 'activate', *érlel* 'ripen', *mélyít* 'deepen, aggravate')
4. COVERING (e.g. *övez* 'surround', *óv* 'guard', *tartalmaz* 'contain)
5. CREATION/REPRESENTATION (e.g. *alkot* 'create', *szemléltet* 'illustrate', *szaporít* 'breed')
6. EVALUATION/EXPERIENCE (e.g., *gyűlöl* 'hate', *csodál* 'admire', *un* 'be bored by')
7. INGESTION (e.g., *fogyaszt* 'consume', *fal* 'devour', *kortyol* 'take sips of')
8. OWNERSHIP (e.g., *birtokol* 'own, possess', *érdemel* 'deserve', *illet* 'belong to')
9. PERCEPTION (e.g., *hall* 'hear', *vizsgál* 'examine', *érzekel* 'perceive')
10. PREFERENCE (e.g., *preferál* 'prefer', *választ* 'choose', *latolgat* 'ponder on (a decision/choice)')
11. ? (OTHER) (e.g. *szerkeszt* 'edit', *dédelget* 'fondle, pamper', *hallat* 'make heard')

[8] Classifying verbs based on their ordering preferences would mean that we are not actually classifying them along the semantic definitions of classes we set up. This would clearly introduce bias and prevent us from meeting our research goal of studying the effect of verb semantics on verb-object ordering.

For most of our verb classes, both conditions 1) and 2) are met. There are two exceptions where one of these conditions factored much more prominently into our decision to create the verb class than the other. The first one is the separation of change of state verbs (CHANGE) from verbs implying force exertion (AFFECT) — here, condition 1) holds more strongly than condition 2) (see Levin 1993; p. 240). The second exception is verbs implying preference (PREFERENCE), where condition 2) is met, but condition 1) less so.

Our flowchart[9] shows the classes and the decision procedure we used to assign verbs to classes. The series of questions that lead to each category in the flowchart constitutes the definition of each category. We color code each verb class in the flowchart for their ordering preference, which is based on patterns observed in the training data. We discuss the ordering preference of each verb class in Section 5.

4.3. OBJECT FEATURES. In Section 1, we motivated our use of two control features, object definiteness and object NP weight, which we compare with the effect of the verb's semantic class.

Object definiteness. In Section 3.2, we motivated our use of object definiteness as a feature and provided a brief overview of definiteness marking in Hungarian. We detect the definiteness of the object as part of the morphological analysis (Section 4.1.2), by looking at the definiteness marking on verbs.

Object NP weight. The complexity of constituents can be predictive of their relative ordering, given that "constituents in many languages tend to occur in increasing size and complexity"(Wasow 1997; p. 81). Complexity and size can be calculated in different ways. Bresnan et al. (2007) uses length of the constituents as a predictor of their relative ordering in the dative alternation and Benor & Levy (2006) uses the number of syllables in each noun as a predictor of their ordering in binomial expressions. We use the weight of the object NP (i.e. the number of elements in it) as an estimate of its complexity. We estimate object NP weight using the StanfordNLP parser (Qi et al. 2018), by calculating the number of dependents of the object's head noun.

4.4. LOGISTIC REGRESSION. We use a simple logistic regression model to see how well our features can predict the ordering of verbs and their objects. The binary response variable is the ordering of verbs and their objects (0 for OV, 1 for VO). Our set of predictors include categorical variables – the verb's lemma, the verb's semantic class, the definiteness of the object – and a continuous variable – the weight of the object NP. The model estimates the probability p of a verb preceding the object via

$$p = 11 + e^{(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}$$

where the β_i is the weight (or parameter) associated with variable x_i. As for categorical variables, each verb category is assigned a separate binary variable, except for the alphabetically first category (in our case, the OTHER category represented by ?), which is estimated by the intercept (β_0). Each continuous variable is assigned to a single continuous variable.

4.4.1. ESTIMATING ORDERING PREFERENCE. We estimate the ordering preference of each verb by observing if it is significantly more likely to precede or follow its objects in our data.

[9] https://ddemszky.github.io/hungarian_verbs/verb_classification_flowchart.pdf

To this end, we build a logistic regression model with a single predictor, the verb's lemma, and estimate the parameters for each verb. The model estimates the log odds of a verb preceding or following its object. If the log odds is negative, then the verb is more likely to precede its object, and if the log odds if positive, then it is more likely to follow its object.

We determine the significance of the estimate for each verb by z-scoring the log odds (dividing the log odds by the standard error). We consider the ordering preference of a verb to be significant if the absolute value of the z-score exceeds 2, which means that the estimate is greater than 2 standard deviations. We follow the same procedure for determining the ordering preference of each semantic class, by running a logistic regression model with verb class as a single predictor.

4.4.2. MODEL ACCURACY. We treat the accuracy of each model as an estimate for how well the features in that model can explain verb-object ordering in our data. Accuracy is defined as the proportion of TOKENS whose ordering the model predicts correctly. To see whether our models are overfitting to the training set, we also use our models (trained on the training set) to predict the ordering of TOKENS in the test set. If the difference between the training and test accuracies is not significant (as measured by a two-sample t-test on the training vs test predictions), that means that our parameter estimates are not biased towards the training set.

5. Results & Discussion. In this section, we address each of the research questions outlined in Section 2.

5.1. ORDERING PREFERENCES OF VERBS. We find that 280 out of 380 verbs (74%) have a significant ordering preference. Out of the verbs with a significant ordering preference, 106 (38%) have an OV preference and 174 (62%) have a VO preference. The fact that more verbs have a VO preference than an OV preference shows that focusing objects (resulting in an OV order) is less frequent across verbs than not focusing them. It is interesting that OV-preferring verbs still do make up a significant proportion of verbs with an ordering preference (38%). These OV-preferring verbs are transitive verbs that tend to occur with focused objects, and they form a larger group than has been identified in previous work by Komlósy (1989). The ordering preference of all verbs in our data is included in the Supplementary Material.

In our first research question, we also asked how well we can predict the ordering of verbs and objects in our data based on the verb exclusively. We sought to answer this question using a model that only includes the lemma of the verb as a feature — we call this model the VERB-ONLY model. We found that the accuracy of this model is 68%, as shown in Table 2, with no significant difference between the training and the test set ($p \approx 0.3$, obtained via a two sample t-test).[10] We compared the accuracy of the VERB-ONLY model with the majority baseline that randomly predicts VO ordering with the same frequency as observed in the training data (53%). The accuracy of the VERB-ONLY model is 1.3 times better than the majority baseline with very high significance ($p < 0.001$). This result suggests that the verb does explain a significant portion of the variance in verb-object order.

5.2. ORDERING PREFERENCES OF VERB CLASSES. In Section 4.2, we described our verb classification procedure; this procedure included identifying a small set of semantic classes for

[10] It is important to note that the upper bound on classification accuracy is lower than 100%, as there is also some amount of free variation in verb-object ordering in Hungarian (É. Kiss 1994). We leave the estimation of the amount of this free variation for future work.

Model	Train Acc.	Test Acc.
majority baseline	52.91%	53.04%
VERB-ONLY	67.95%	67.82%
CLASS-ONLY	65.04%	64.93%
OBJECT-WEIGHT-ONLY	54.90%	54.99%
DEFINITENESS-ONLY	57.29%	57.17%
COMPLEX (class+object-size+definiteness)	67.63%	67.52%

Table 2. Accuracy of logistic regression models, using different features, on the training and the test sets.

the verbs, which are salient based on previous literature and based on the semantic features that are relevant for the verbs' ordering preference. Below we summarize our high-level findings regarding the relationship between the semantic verb classes and their ordering preference.

- Stative verbs tend to be OV-preferring, especially if they denote location or spatial configuration. Both experiencer-subject and experiencer-object psych verbs are exceptions among stative verbs, as they are nearly all VO-preferring.
- Non-stative verbs, especially ones that entail the subject doing something to or with an existing object, tend to be VO-preferring. Verbs of creation / representation, which do not entail an existing target object,[11] are exceptions to this, as they tend to be OV-preferring.

The verb classes are listed in Table 3, along with their ordering preference and example class members showing a significant ordering preference. The largest OV-preferring class is CREATION/REPRESENTATION with 50 verbs, and the largest VO-preferring class is CHANGE, with 110 verbs. The smallest OV-preferring class is PERCEPTION, with 6 verbs, and the smallest VO-preferring class is INGESTION, with 11 verbs. The disparity between the class sizes can be partially explained by the disparity between the semantic granularity of the classes — for example, CHANGE is a much broader category semantically than INGESTION. In Section 4.2, we identified an OTHER class of 18 verbs (5%) that were either highly polysemous or they did not fit the definition of any class; this class, which included 18 verbs, was included in all our models.

In this research question, we also asked how well we can predict verb-object ordering based on the verb's semantic class. To this end, we construct a CLASS-ONLY model, which achieves an accuracy of 65%, with no significant difference between the training and test sets ($p < 0.001$). This result supports the way in which we categorized the verbs, as we find that the ordering preference of our verb classes largely account for the ordering preference of individual verbs. Specifically, the accuracy of the VERB-ONLY model (an upper bound on verb classification quality) is 68%, only 3% higher than the accuracy of CLASS-ONLY model. This

[11] As the footnote to the flowchart (Section 4.2) mentions, for verbs denoting representation, there's a distinction between source objects (the one being represented) and target objects (the representation). For other verbs, this distinction is not relevant.

Category	Pref.	Verb Counts			Example Verbs	
		OV	VO	No sig. pref.	OV	VO
ACTIVITY	OV	5	0	3	alkalmaz ‘employ’, kutat ‘research’, üzemeltet ‘operate’	
COVERING	OV	13	2	4	fed ‘cover’, díszít ‘decorate’, övez ‘surround’	óv ‘protect’, kerülget ‘go around’
CREATION/ REPRESENTATION	OV	35	2	13	alkot ‘create’, tükröz ‘mirror’, jelent ‘mean’	szaporít ‘breed’, vázol ‘sketch’
OWNERSHIP	OV	4	0	0	birtokol ‘possess’, érdemel ‘deserve’, illet ‘belong to’	
PERCEPTION	OV	3	0	3	hall ‘hear’, vizsgál ‘examine’	
PREFERENCE	OV	4	0	1	választ ‘choose’, preferál ‘prefer’, céloz ‘aim at’	
AFFECT	VO	14	47	30	csókol ‘kiss’, fenyeget ‘threaten’, sürget ‘urge’	hajszol ‘rush’, bántalmaz ‘hurt, abuse’, támogat ‘support’
CHANGE	VO	12	74	24	ihlet ‘inspire’, motivál ‘motivate’, tömörít ‘compactify’	aktivál ‘activate’, csökkent ‘reduce’, javít ‘repair’
EVALUATION/ EXPERIENCE	VO	2	39	15	gyászol ‘mourn’, hibáztat ‘blame’	bírál ‘judge’, csodál ‘admire’, gyűlöl ‘hate’
INGESTION	VO	2	9	0	fogyaszt ‘consume’, vedel ‘drink (a lot of)’	fal ‘devour’, iszik ‘drink’, olvas ‘read’
OTHER	OV	10	1	7	szerkeszt ‘edit’, szolgál ‘serve’	viszonoz ‘requite’

Table 3. Verb classes, their ordering preference (all of them are significant) and example verbs.

small difference in model accuracy is remarkable given that there is a large, 97% reduction in the size of the feature set from the VERB-ONLY model (380 verbs) to the CLASS-ONLY model (11 classes).

5.3. IMPORTANCE OF OBJECT-RELATED FEATURES. We construct a DEFINITENESS-ONLY and an OBJECT-WEIGHT-ONLY model to estimate the effect size of object definiteness and object NP weight, respectively. The DEFINITENESS-ONLY model achieves an accuracy of 57% and the OBJECT-WEIGHT-ONLY model achieves an accuracy of 55% (training and test performance are the same for both models, $p < 0.001$). These results indicate that these object-related features are not as important as features related to the verb (verb lemma and verb class).

Our COMPLEX model, which includes verb class, object definiteness and object NP weight as features, achieves 68% accuracy, with no significant difference between training and test performance ($p < 0.001$) Thus, using the COMPLEX model with only 14 features (11 for verb class + 2 for object definiteness + 1 continuous variable for object NP weight), we can predict verb-object ordering with the same level of accuracy as using the VERB-ONLY model with 380 features. The fact that there is no significant difference between training and test accuracy for any of the models shows that the parameter estimates are not overfitting to the training set.

6. Conclusion. In this paper, we investigated the role of three features, including the verb's semantic class, object definiteness and object NP weight, in determining verb-object order in Hungarian. We extract verb-object pairs and their associated object-related features from the Hungarian Gigaword Corpus, manually assign verbs to semantic classes, and build a logistic regression model to estimate the importance of different features. We discover patterns of semantic similarity among verbs with similar ordering preference, and we perform our verb classification based on these patterns.

We find that all of our features obtain significantly higher accuracy than the majority base-

line in predicting verb-object ordering. Even though using the verb lemma as a feature seems to be the most effective among our features, the complexity of the VERB-ONLY model (with 380 features) is significantly greater than that of our COMPLEX model, which contains all other features (14 features total). The VERB-ONLY model and COMPLEX models obtain the same accuracy, 68%, on both the training and the test sets. A CLASS-ONLY model achieves 65% accuracy, only 3% less than the VERB-ONLY model, suggesting that our semantic classification approximates the similarities among verbs in terms of ordering preference quite well.

The predictive power of our lexical semantic feature shows that it might be playing a non-negligible role in Hungarian word order. We hope that this investigation will lead to more studies in this domain. Promising avenues for future work include extending our analysis to a larger number of verbs (e.g., verbs with multiple arguments), adding additional features (e.g., the animacy or humanness of arguments) and better understanding the interaction among discourse context, lexical semantic factors and perhaps other factors, such as cognitive constraints, that may influence word order in Hungarian.

References

Benor, Sarah & Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* 82(2). 233–278. https://doi.org/10.1353/lan.2006.0077.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. *Cognitive Foundations of Interpretation*, 69–94. KNAW.

Brody, Michael. 1990. Some remarks on the focus field in Hungarian. Tech. rep. *UCL Working Papers in Linguistics* 2.

É. Kiss, Katalin. 1978. *A magyar szintaxis egy transzformációs generatív megközelítése*. Budapest: Hungarian Acaedemy of Sciences dissertation.

É. Kiss, Katalin. 1981. Structural relations in Hungarian, a "free" word order language. *Linguistic Inquiry* 12(2). 185–213. http://www.jstor.org/stable/4178216.

É. Kiss, Katalin. 1994. Scrambling as the base generation of random complement order. In Norbert Corver & Henk van Riemsdijk (eds.), *Studies on scrambling: Movement and non-movement approaches to freeword-order phenomena*, 221–256. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110857214.221.

É. Kiss, Katalin. 1995. *Discourse configurational languages*. Oxford, UK: Oxford University Press.

É. Kiss, Katalin. 2002. *The syntax of Hungarian*. Cambridge, UK: Cambridge University Press.

Farkas, Donka F. 1986. On the syntactic position of focus in Hungarian. *Natural Language & Linguistic Theory* 4(1). 77–96. https://doi.org/10.1007/bf00136265.

Farkas, Donka F. & Henriëtte de Swart. 2003. *The semantics of incorporation: From argument structure to discourse transparency*. Chicago: University of Chicago Press.

Gecség, Zsuzsanna & Ferenc Kiefer. 2009. A new look at information structure in Hungarian. *Natural Language & Linguistic Theory* 27(3). 583–622. https://doi.org/10.1007/s11049- 009-9071-7.

Kálmán, C. György, László Kálmán, Ádám Nádasdy & Gábor Prószéky. 1989. A magyar segédigék rendszere. *Általános Nyelvészeti Tanulmányok* 17. 49–103.

Kálmán, László, Gábor Prószéky, Ádám Nádasdy & C. György Kálmán. 1986. Hocus, focus, and verb types in Hungarian infinitive constructions. In Werner Abraham & Sjaak de Mey (eds.), *Topic, focus and configurationality,* 129–142. Amsterdam: John Benjamins.

Kiefer, Ferenc. 1967. *On emphasis and word order in Hungarian* (Indiana University Publications, Uralic and Altaic Series 76). The Hague: Mouton.
Kiefer, Ferenc. 1990. Noun incorporation in Hungarian. *Acta Linguistica Hungarica* 40(1-2). 149–177.
Komlósy, András. 1989. Fókuszban as igék [Verbs in focus]. *Általános Nyelvészeti Tanulmányok* 17. 171–182.
Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago press.
Marácz, László. 1989. Asymmetries in Hungarian (I). *Anuario del Seminario de Filología Vasca "Julio de Urquijo"* 24(2). 407–524.
Oravecz, Csaba, Tamás Váradi & Bálint Sass. 2014. The Hungarian Gigaword Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. 1719–1723.
Qi, Peng, Timothy Dozat, Yuhao Zhang & Christopher D. Manning. 2018. Universal dependency parsing from scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. 160–170.
Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multi-word expressions: A pain in the neck for NLP. *International Conference on Intelligent Text Processing and Computational Linguistics*. 1–15. Springer.
Trón, Viktor, András Kornai, György Gyepesi, László Németh, Péter Halácsy & Dániel Varga. 2005. HunMorph: open source word analysis. *Proceedings of the Workshop on Software*. 77–85. Association for Computational Linguistics. https://doi.org/10.3115/1626315.1626321.
Vincze, Veronika & János Csirik. 2010. Hungarian corpus of light verb constructions. *Proceedings of the 23rd International Conference on Computational Linguistics*. 1110–1118. Association for Computational Linguistics.
Vincze, Veronika, Dóra Szauter, Attila Almísi, György Móra, Zoltin Alexin & Jinos Csirik. 2010. Hungarian Dependency Treebank. *Proceedings of the Seventh Conference on International Language Resources and Evaluation LREC 10*.
Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9(1). 81–105. https://doi.org/10.1017/s0954394500001800.
Zsibrita, János, Veronika Vincze & Richárd Farkas. 2013. magyarlanc: A tool for morphological and dependency parsing of Hungarian. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 763–771.