Acta Linguistica Academica 71 (2024) 1–2, 202–218
DOI: 10.1556/2062.2023.00701

AKADÉMIAI KIADÓ

Hungarian auxiliaries revisited

ÁGNES KALIVODA[1*] and GÁBOR PRÓSZÉKY[1,2]

[1] HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary

[2] Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary

Received: April 21, 2023 • Revised manuscript received: October 5, 2023 • Accepted: October 5, 2023
Published online: March 11, 2024

ABSTRACT

This paper uses a corpus linguistic approach to investigate finite verbs co-occurring with infinitives. It aims to explore a range of similar verbs along a set of formal-distributional features based on Kálmán C. et al.'s (1989) study. We used hierarchical agglomerative clustering to analyze the data. The analysis identifies four clusters, two comprising verbs more auxiliary-like than the others. The results of this experiment are broadly similar to those of Kálmán C. et al. (1989); however, we also find remarkable differences. Most importantly, the so-called stress-avoiding verbs are likely to occur between the preverb and its associated infinitive, indicating that they are much closer to central auxiliaries than previously assumed.

KEYWORDS

auxiliaries, Hungarian grammar, data-driven approach, hierarchical agglomerative clustering, formal-distributional criteria

1. INTRODUCTION

The category of auxiliaries is controversial in the Hungarian linguistic tradition. It is the subject of ongoing debate whether there is such a group, and if so, what makes it different from the rest of the verbs. In the broadest sense, we can label all verbs with an infinitive argument with this term, but applying it only to defective verbs is more common. Defectivity can be phonological (striving for unaccentedness), morphological (empty paradigm cells), syntactic (a relatively bound word order), and semantic (incomplete argument frame). These properties are usually

* Corresponding author. E-mail: kalivoda.agnes@nytud.hun-ren.hu

interrelated. In what follows, we present a brief overview of the approach types that we consider the most fundamental.[1] We partly based our summary on the literature surveys written by Kenesei (2001) and Laczkó (2014), which provide further details on this subject.

The first type of approach is usually referred to as the traditional, descriptive one – see Tompa (1961), among others – characterized by presenting a list of what the author considers auxiliaries without setting explicit criteria. Instead, they refer to these words' "value of inflectional affixes". According to Kenesei (2001), this amounts to saying that they are words but have a function similar or equivalent to inflectional affixes (e.g., expressing tense, aspect, modality).

An outstanding representative of the second approach is Kálmán C. et al. (1989), a comprehensive descriptive study of Hungarian 'finite form + infinitive' constructions. This study aims to classify the finite verbs occurring in these constructions as auxiliaries or main verbs using strict formal-distributional criteria. This study has the most relevance for the present research, and we will discuss it at several points throughout the paper.

A third approach, significantly different from the previous one, appears in the early generative works of É. Kiss (1987, 1992), where the author assumes there are no auxiliaries in Hungarian. Her theory subsumes all verbal elements under category V, and it assigns the auxiliary-like distributional patterns and the semantic, argument structural properties of specific verbs to their lexical specifications.

Fourth, Kenesei's (2001) approach applies a wide range of criteria based on a list of auxiliary properties borrowed from Heine (1993). This list of criteria comprises 18 partially interrelated items, five of which are crucial for identifying Hungarian auxiliaries in Kenesei's framework. Laczkó (2014) summarizes these in the following way: (1) Hungarian auxiliaries have defective paradigms. (2) They cannot function as semantic predicates of sentences. (3) They cannot be complements of other predicates. (4) They cannot be nominalized. (5) In their presence, the main verb takes infinitival form.

All of the approaches above rely on intuition and native speaker judgment. Therefore, it seems justified to highlight the work of Modrián-Horváth (2009), the first study on this subject that extensively uses corpus data. This study sets three major criteria of auxiliary status: (1) the insertion of the finite form between the preverb and its associated infinitive, (2) its frequency of occurrence, and (3) the marked or unmarked status of the infinitive complement. Based on corpus data, two major types of auxiliaries seem to emerge. Still, the author emphasizes that there are no sharp boundaries between neither the two types nor these and other kinds of 'finite form + infinitive' constructions. This is an important note supported by the present study as well.

Since these approaches consider and prioritize different sets of properties and are motivated by different theoretical frameworks, their results are also quite diverse. In essence, *fog* 'will' is the only lexical item generally classified as an auxiliary verb (at least by approaches that assume the existence of Hungarian auxiliaries). The present study does not aim to compare or evaluate them. Our interest is in computational corpus linguistics and the applicability of its methods to the issue at hand. Thus, we opt for the approach best suited for a quantitative and largely automatic investigation. This is the approach of Kálmán C. et al. (1989), which defines

[1]The present overview focuses on papers concerned with the category of auxiliaries rather than how to analyze auxiliary constructions in different theoretical frameworks. Regarding the latter issue, we suggest the following studies: É. Kiss (2004) for a Chomskyan generative account, Laczkó (2014) for Lexical-Functional Grammar, Imrényi (2013) for projective dependency grammar, and Tolcsvai Nagy (2010) for a functional cognitive linguistic framework.

Hungarian auxiliaries based on syntactic and prosodic behavior. We aim to explore groups of similar verbs along a set of features based on the abovementioned study. We carry out this exploration using hierarchical agglomerative clustering.

The paper is organized as follows. Section 2 presents Kálmán C. et al.'s (1989) study focusing on its methods, which will be particularly important for our experiment. Section 3 describes our methodology in detail. Section 4 provides the cluster analysis results, followed by the interpretation of the results in Section 5. Finally, Section 6 summarizes the experiment and outlines prospects for further research.

2. A FORMAL-DISTRIBUTIONAL METHODOLOGY

The following summary presents the methodology of Kálmán C. et al. (1989) in detail since it is the starting point of the present approach. The group of finite forms under survey, the criteria of auxiliaryhood, the studied environments, and the procedure will be discussed in turn. Finally, we report the main results of the article.

In Kálmán C. et al. (1989), the set of studied lexical items comprises every verb and predicative nominal that might co-occur with an infinitive, except for the following: (1) verbs always negated, e.g., *(nem) átall* 'have the face to do st'; (2) verbs never occurring in affirmative or declarative clauses, e.g., *szíveskedik* 'kindly do st'; and (3) rare or archaic verbs, about which the authors had no intuition. Some types are represented only by a handful of items, as the authors assume these behave similarly. These are (1) adjectives, (2) nouns, (3) verbs of motion, (4) preverb-verb combinations, and (5) complex verb phrases without a copula, e.g., *jólesik* 'it feels good to do sth'.

Before we turn to the criteria of auxiliaryhood as defined in Kálmán C. et al. (1989), we need to outline some relevant facts about Hungarian sentence structure. First, two sentence types must be distinguished: neutral and non-neutral sentences. Neutral sentences are characterized by equal stresses on major constituents and relatively strict word order. On the contrary, non-neutral sentences usually have one prominent stress, and the word order is determined by the specific non-neutral clause types (e.g., progressive, imperative, wh-question). For more details, see Kenesei, Vago & Fenyvesi (1998).

In the case of neutral sentences, the finite form and its infinitive complement can appear in three word order patterns, each illustrated in (1). The finite form can precede the infinitive (1a), follow it unstressed – which is also called enclitic behavior – (1b), or appear interposed between the preverb and the infinitive that is lexically associated with the preverb (1c). The latter is a subtype of enclitic behavior, in which the finite form is enclitic to the preverb. The position before the finite form is called the verb modifier position in neutral sentences. The preferred word order is largely determined by the finite form co-occurring with the infinitive, and it can be mapped onto the finite form's prosodic behavior, see Komlósy (1989).

(1) a. *imád-om* ***el-mond-ani*** *a* *vélemény-em-et*
love-PRES.1SG/DEF away(PV)-speak-INF the opinion-1SG.POSS-ACC
'I love to speak my mind'

b. ***el-mond-ani*** *kényszerül-t-em* *a* *vélemény-em-et*
away(PV)-speak-INF be.obliged-PAST-1SG the opinion-1SG.POSS-ACC
'I was forced to speak my mind'

c. ***el*** *akar-t-am* ***mond-ani*** *a* *vélemény-em-et*
away(PV) want-PAST-1SG speak-INF the opinion-1SG.POSS-ACC
'I wanted to speak my mind'

According to Kálmán C. et al. (1989), the first criterion of auxiliaryhood is enclitic behavior: if the finite form is an auxiliary, it is stress-avoiding in neutral sentences, and the verb modifier position is taken by the infinitive, as in (1b). The second one, interposition, is a subtype of enclitic behavior attested when the infinitive has a preverb, as shown in (1c). The third criterion is based on the distinction between given and new information. If the finite form is an auxiliary, it can be unstressed even if it appears as new information in the clause. Finally, there is a criterion called striving for finiteness: the more auxiliary-like a lexical item is in Hungarian, the more it tends to appear in finite form. This is illustrated in (2a), where the verb *tud* 'can' appears as an infinitive complement of the verb *fog* 'will'. In contrast, the reverse case in (2b), showing *fog* 'will' as an infinitive governed by *tud* 'can', is infelicitous and barely attested in corpora. This indicates that *fog* 'will' is more auxiliary-like than *tud* 'can'.

(2) a. *meg* *fog-juk* *tud-ni* *valósít-ani* *az* *álma-i-nk-at*
PV take-PRES-1PL/DEF be.able-INF realize-INF the dream-PL-1PL.POSS-ACC
'we will be able to make our dreams come true'

b. ?**meg* *tud-juk* *fog-ni* *valósít-ani* *az* *álma-i-nk-at*
PV be.able-PRES-1PL/DEF take-INF realize-INF the dream-PL-1PL.POSS-ACC
'we will be able to make our dreams come true'

Regarding the environments, sentences considered must be declarative, positive and simple. The reason for that is the following: "Hungarian sentences have such powerful overriding rules for interrogation, negation, imperatives and complex arguments, that these would neutralize the differences stemming from the inherent property" of the auxiliary-like finite forms, namely, their tendency to enclisis (Prószéky et al. 1984, 169).

Having made the decisions presented above, the authors characterized each finite form along the following features: (1) What distributional pattern does it show depending on the prosody of the sentence? Does it appear in a different position depending on whether the sentence has flat or eradicating prosody? (2) What position does the finite form take within the construction? (3) Is the co-occurring infinitive a simple verb, a verb with a preverb in direct order, or a verb with a preverb in discontinuous order? The basis of this characterization was the authors' intuition and agreement. Finally, the authors determined groups of finite forms by manually evaluating and organizing the data. Table 1 summarizes their results.

3. DATA AND METHOD

Our research was inspired by Kálmán C. et al. (1989), and we tried to reproduce it as accurately as possible regarding the studied environments, lexical items, and feature set. We must note, however,

Table 1. Classification of finite forms co-occurring with infinitives according to Kálmán C. et al. (1989, 100). Number 3 following some of the verbs means that the verb is used only in 3rd person singular form. The sign □ indicates stress-avoiding verbs. The difference between II/A and II/B is assumed to be sociolinguistic. Finite forms of II/A may appear in intervening positions only in spoken/vernacular language. Those of II/B do so in vernacular use but show affix-like behavior in formal use

I. Auxiliaries		II. Emphatic verbs			III. Affix-like verbs
I/A Central auxiliaries	I/B Peripheral auxiliaries	Stress-preferring verbs		II/C Stress-requiring verbs	
		II/A	II/B		
akar *fog* *kell* *szokott* *tetszik* *tud* ('can')	*bír* *kezd* *kíván* *lehet 3* *mer* *óhajt* *próbál* *szándékozik* *szeretne* *talál* □ *tud* ('know') [Nominal + copula]	*illik 3* *sikerül* *szeret* Double agent verbs *(enged, hagy, segít)* [Nominal + copula]	*igyekszik* *iparkodik* *készül* *tartozik* *törekszik*	*bátorkodik* *imád* *siet* Negative verbs *(fél, ...)* [Nominal + copula] [Preverb-verb combination]	*kényszerül* □ *látszik* □ *tanul* *vágyik* □ *van* *vél* □ Double agent verbs *(hall, ...)* Verbs of motion *(megy, ...)*

that there are considerable differences between the settings of the two studies at some points, which we will discuss in detail. The present study takes a quite different direction in data analysis since it extensively uses unsupervised machine learning, which was not an option in the 1980s.

This section starts with presenting the corpus that serves as our data source. It is followed by the description of lexical items and features we considered to include in our dataset. A separate subsection is devoted to the presentation of the data collection and cleaning process since the outcome of the entire analysis is contingent on these preparatory steps. Finally, we introduce the applied data analysis. The codes and data created during the current study are available in the following GitHub repository: https://github.com/kagnes/hungarian_auxiliaries_revisited.

3.1. Text source

The Hungarian Gigaword Corpus (Oravecz, Váradi & Sass 2014) is a 1.04 billion-word, automatically annotated general corpus designed to represent a broad cross-section of Hungarian from the later part of the 20th century and the start of the 21st century. We use its newest – yet unpublished – version (Kalivoda et al. 2023), which differs from the original HGC in three aspects. First, it underwent several corpus cleaning steps, e.g., filtering extremely long sentences and duplicate or non-Hungarian paragraphs. Consequently, the overall text quality is higher, but the corpus size is smaller (776.9 million running words). Second, it contains the year of publication for 93.8% of the texts, in addition to the metadata already available in the original corpus: region and register. Finally, it is enriched with new annotation layers: detailed

morphological annotation created by emMorph (Novák, Siklósi & Oravecz 2016) and dependency relations following the Universal Dependencies standard (de Marneffe et al. 2021).

Dependency relations are of primary importance for this study since both the infinitive and the finite verb may have separable preverbs, which can be easily connected to their respective verb stems using this annotation layer. Furthermore, it helps to determine which finite verb – or predicative nominal – is the head of a given infinitive.

3.2. Entities and features

We decided to focus on simple verbs, disregarding nominals and preverb-verb combinations. The reason for this was to make our task easier about comparing the two studies since we would have expected many more lexical items if we did not impose this restriction on our sample. The feature set we developed comprises automatically extractable distributional properties of verbs that are likely to be relevant, based on Kálmán C. et al. (1989). These are (1) FIN INF = the given verb precedes the infinitive, (2) INF FIN = the verb is enclitic to the infinitive, (3) intervening = the verb appears between a preverb and its associated infinitive, and finally, (4) inf_form = the verb appears in the infinitive form. Regarding the last feature, we assume that auxiliaries strive for finiteness, meaning that the lower this value, the higher the likelihood of auxiliaryness. (3) illustrates the four features with the verb *akar* 'want'.

(3) a. *Akar-t-am* ***ven-ni*** *egy másik kemping-ágy-at.* → *FIN INF*
want-PAST-1SG buy-INF an other camping-bed-ACC
'I wanted to buy another camping bed.'

b. ***Beszél-ni*** *akar-ok* *vele-d.* → *INF FIN*
talk-INF want-PRES.1SG/INDEF with-2SG
'I want to talk to you.'

c. ***El*** *akar-ok* ***búcsúz-ni.*** → *intervening*
away(PV) want-PRES.1SG/INDEF say.goodbye-INF
'I want to say goodbye.'

d. *Akar-ni* ***kell*** *mind a két oldal-on.* → *inf_form*
want-INF have.to all the two side-SUP
'It has to be wanted by both sides.'

Kálmán C. et al. (1989) present an additional feature we cannot consider in the present study. It is the distribution of auxiliary-like items depending on whether they appear as given or new information in the sentence. This criterion can be applied if one has access to prosody annotation in the corpus; therefore, we had to disregard it.

3.3. Data collection and cleaning

As the first step of data collection, we extracted from the corpus all sentences containing an infinitive with a finite verb as its head. This initial set of data was then filtered considerably. We tried to examine only the environments used by Kálmán C. et al. (1989): declarative, positive and simple sentences. We preserved only sentences that met all the following conditions. First, the sentence ends with a full stop or a combination of a full stop and a quotation mark.

The prospective auxiliary bears neither a conditional mood suffix nor a subjunctive suffix, as these would indicate optative/desiderative and imperative/prohibitive sentences, respectively. Second, it does not contain a full stop, comma, semicolon, exclamation mark or question mark. This was needed to avoid complex sentences and structures such as the one illustrated in (4), which are frequent in novels and other narrative texts. In Hungarian corpora, it is generally analyzed as a single sentence. Regarding the exclusion of commas and semicolons, we know that we lose several good hits (e.g., simple sentences containing enumerations). However, precision is more critical for us in this task than recall.

(4) *– Engem senki sem szeret! – akar-t-am ordít-ani.*
– I.ACC nobody not.even love.PRES.3SG/INDEF – want-PAST-1SG scream-INF
'"Nobody loves me!" I wanted to scream.'

Third, the sentence does not contain words that are the most frequent indicators of negation or prohibition: *ne* 'don't', *nem* 'no/not', *se* 'not/neither', and *sem* 'not/neither'. Finally, it does not have a left periphery; it starts with a finite form, an infinitive, or a preverb associated with either of these (quotation mark or dash is allowed, though). This might seem an unusual criterion – and absent from Kálmán C. et al. (1989) – but it has proven to be an essential step in our experiment. Sentences with structural focus are virtually impossible to set apart from simple neutral sentences automatically if this difference is not annotated in the corpus. (5) illustrates the issue.

(5) *A feleség-e próbál-t rajta segít-eni.*
the wife-3SG.POSS try-PAST.3SG/INDEF he.SUP help.INF
Neutral reading: 'His wife tried to help him.'
Non-neutral reading (with focus): 'It was his wife who tried to help him.'

In speech, the intended meaning of (5) is obvious since the difference can be heard: in the first case, the verb *próbált* 'tried' is stressed, whereas in the second one, the noun phrase *a felesége* 'his wife' carries the primary stress. The sentence is ambiguous in written text unless a broader context is provided.

In summary, we had to impose strict requirements on the environment, resulting in losing a considerable number of good sentence candidates. However, we aimed at high precision – and accepted low recall – as too much noise in the data could have distorted the quality of clustering to a great extent. At the end of our filtering process, 12,287 sentences remained.

Manual normalization of the automatically extracted verb types was also unavoidable. This comprised two different processes. On the one hand, it was necessary to merge two or more lemmas in the case of slang forms (e.g., *kő, kék* → *kell* 'have to'); subtle, mostly dialectal differences (e.g., *köll* → *kell* 'have to' or *röstell* → *restell* 'be ashamed'); and finally, forms frequently deviating from standard orthography (e.g., *igyexik* → *igyekszik* 'strive', *teccik* → *tetszik* 'indicates politeness'). On the other hand, splitting one lemma into two distinct lemmas was justified in some cases when rather different distributions could be expected, based on previous literature as well as our intuition. The lemma *szeret* was split into *szeret* 'love' and *szeretne* 'would like', and *lesz* 'will be' into *van* 'be' and *lehet* 'could be'. Kálmán C. et al. (1989) suggest that the verb *tud* can be characterized by two different distributions depending on its distinct

senses 'can' and 'know how to'. Since we could not use any morphological clue for re-lemmatization, we had no choice but to leave this lemma intact. After normalization, we set a frequency threshold of 10 occurrences for the studied verb types since the clustering would have been less reliable if we included sparse data points. We must note that *talál* 'happen to', which is often held to be an auxiliary, did not reach this threshold. A closer look at the corpus data revealed that this verb appears mostly in non-neutral sentences, especially in the phrase *Azt találta mondani, hogy…* 'He/She happened to say that…'. This observation does not mean it must be ruled out as an auxiliary verb, but it would not have been possible to obtain reliable results for this verb with our method.

Each verb type was described using the four variables introduced in Section 3.2. The attested frequencies are for each feature summarized per verb, yielding a similarity matrix illustrated in Table 2.

In order to make the matrix suitable for cluster analysis, absolute frequencies had to be normalized. Normalization is the process of scaling individual samples to have a unit norm to avoid or at least lessen the negative impact of skewed frequency distribution on the clustering.

3.4. Analysis: hierarchical agglomerative clustering

Clustering is a machine learning technique that involves grouping a set of objects so that objects in the same group (called a cluster) are more similar than those in other clusters. One of the most common types of clustering is hierarchical agglomerative clustering. This is a bottom-up approach where each object is a singleton cluster at first, and pairs of clusters are successively merged while moving up the hierarchy. See Moisl (2015) for a detailed explanation intended for readers with a background in linguistics. Our analysis workflow is based on Hees's (2015) tutorial.

One of the most pressing questions in cluster analysis is deciding which set of parameters would be best for analyzing a given similarity matrix. One essential parameter is the distance metric: how the similarity has to be measured. The other is the method one selects to determine how the clusters should be merged. To make this decision as objective as possible, we performed clustering using every combination of metrics and methods available in SciPy (Virtanen et al. 2020). We calculated the cophenetic correlation coefficient for each resulting linkage matrix. This coefficient compares the pairwise distances of all studied objects – in our case, all verbs – to

Table 2. An excerpt from the similarity matrix

verb	FIN INF	INF FIN	intervening	inf_form
ad 'give'	19	15	0	214
ajánl 'recommend'	8	6	0	101
akar 'want'	159	883	1,382	16
bír 'be able'	6	4	10	23
enged 'let'	9	4	0	15
…	…	…	…	…

those implied by the hierarchical clustering. The closer the resulting value is to 1, the better the clustering preserves the original distances. We chose the parameters producing the highest cophenetic correlation coefficient (0.9035): average linkage method with Euclidean distance metric. We performed hierarchical agglomerative clustering with these parameter settings on the normalized matrix. Finally, we plotted the clustering results in a two-dimensional format by applying t-SNE for a more accessible presentation. t-SNE is an algorithm that calculates a similarity measure between pairs of objects in a high-dimensional space and in a low-dimensional space. It then tries to optimize these two similarity measures using a so-called cost function.

We implemented the analysis in Python 3.8 (Van Rossum & Drake 2009). We used the Pandas library (McKinney 2010) for dataframe handling, SciPy (Virtanen et al. 2020) for hierarchical agglomerative clustering, scikit-learn (Pedregosa 2011) for data normalization and t-SNE, and finally, Matplotlib (Hunter 2007) and adjustText (Flyamer 2018) for creating the scatterplot figures.

4. RESULTS

Applying the cluster analysis with the above-discussed parameters on the normalized similarity matrix yields the results in Figure 1. We must note that in the case of two-dimensional maps generated with t-SNE, the axes do not have a particular meaning, which is why they are abandoned in the figure entirely. The relative distances between low-dimensional data points convey the relevant information. Neighboring points in the input space, which was four-

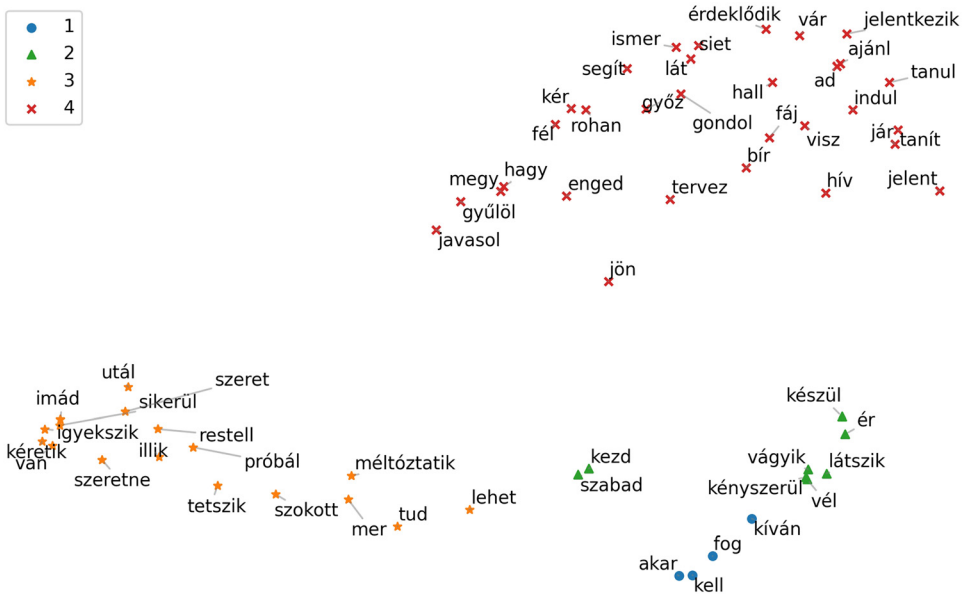

Figure 1. Clustering results shown as a two-dimensional map using the t-SNE algorithm

dimensional in our case, tend to be close to each other in the reduced, two-dimensional space. We must also add that t-SNE in itself is not an analytical tool. To provide a reliable interpretation of the results in Section 5, we considered a dendrogram (tree plot) representation where the whole process of cluster joining can be traced, as well as the original similarity matrix containing raw frequencies.

As a result of the analysis, four clusters could be discerned. The lexical items in each cluster are listed below:

- **Cluster 1:** *akar* 'want', *fog* 'will', *kell* 'have to', *kíván* 'wish'.
- **Cluster 2:** *ér* 'be allowed', *kényszerül* 'be obliged', *készül* 'prepare', *kezd* 'begin', *látszik* 'seem', *szabad* 'may, be allowed', *vágyik* 'desire', *vél* 'assume'.
- **Cluster 3:** *igyekszik* 'strive', *illik* 'be proper', *imád* 'adore', *kéretik* 'be asked for', *lehet* 'may', *méltóztatik* 'deign', *mer* 'dare', *próbál* 'try', *restell* 'be ashamed', *sikerül* 'succeed', *szeret* 'love, like', *szeretne* 'would like', *szokott* 'expressing a habit', *tetszik* 'indicates politeness', *tud* 'can, know how to', *utál* 'hate', *van* 'be'.
- **Cluster 4:** *ad* 'give', *ajánl* 'suggest', *bír* 'cope with', *enged* 'let', *érdeklődik* 'inquire', *fáj* 'hurt', *fél* 'be afraid', *gondol* 'think', *győz* 'cope with', *gyűlöl* 'detest', *hagy* 'leave, let', *hall* 'hear', *hív* 'call', *indul* 'set off', *ismer* 'be familiar with', *jár* 'go', *javasol* 'recommend', *jelent* 'mean, denote', *jelentkezik* 'apply for', *jön* 'come', *kér* 'ask for', *lát* 'see', megy 'go', *rohan* 'rush', *segít* 'help', *siet* 'hurry', *tanít* 'teach', *tanul* 'learn', *tervez* 'plan', *vár* 'wait', *visz* 'bring'.

In broad terms, there is one most defining common feature for each of these clusters. In the case of Cluster 1, this is the ubiquity of the intervening position. For Cluster 2, it is the preference for the enclitic position (INF FIN word order), and for Cluster 3, it is the contrary: a general preference for the starting position (FIN INF). The common feature of verbs in Cluster 4 is that they often appear as infinitives, which means they are low on the finiteness scale. In the next section, we take a closer look at these groups and attempt to interpret them while also comparing our results to those obtained by Kálmán C. et al. (1989).

5. INTERPRETATION OF CLUSTERING RESULTS

In order to facilitate the interpretation of our results, we display each cluster separately. A label is placed after each verb within the cluster, indicating to which group Kálmán C. and colleagues assigned the given verb. In addition, we use special characters as shorthands for information that Kálmán C. et al. (1989) also considered relevant, however, not group-forming. Table 3 summarizes our notations.

Looking back at the classification created by Kálmán C. and colleagues, the obtained large classes are auxiliaries (I), emphatic (II) and affix-like verbs (III). The set called **central auxiliaries (I/A)** comprises the following six lexical items: *akar* 'want to', *fog* 'will', *kell* 'have to', *szokott* 'expressing a habit', *tetszik* 'indicates politeness' and *tud* 'can'. These are described as having a uniform distribution in eradicating and level prosody. Turning to our results, three of these verbs are present in Cluster 1 (as shown in Figure 2), alongside *kíván* 'wish', which is labeled as a **peripheral auxiliary (I/B)** in Kálmán C. et al. (1989), based on its divergent behavior depending on prosody. Our experiment could consider prosody only to the extent it is discernible from word order. Based on the formal-distributional features we used here, *kíván* 'wish' is quite similar to the central auxiliaries.

Table 3. Notations used throughout the discussion

Label	Meaning
I/A	central auxiliary
I/B	peripheral auxiliary
II/A	stress-preferring verb, intervening in spoken/vernacular use
II/B	stress-preferring verb, intervening in vernacular and affix-like in formal use
II/C	stress-requiring verb
III	affix-like verb
□	stress-avoiding verb
*	double agent verb
°	verb of motion
?	not discussed in Kálmán C. et al. (1989)

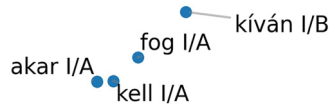

Figure 2. Verbs of Cluster 1

Cluster 1 of our experiment is characterized by the highest frequencies attested in the intervening position. If the given verb does not intervene, it is usually enclitic to the infinitive. These properties are interrelated, as was also discussed in Kálmán C. et al. (1989). These verbs rarely occur in the starting position. They are also at the top of the finiteness scale; they rarely occur as infinitives.

Four of the eight verbs in **Cluster 2** (see Figure 3) are classified as stress-avoiding verbs in Kálmán C. et al. (1989), i.e., they do not receive focus stress even if they are considered new information in a sentence. These are *kényszerül* 'be obliged', *vágyik* 'desire', *látszik* 'seem', and *vél* 'assume'. Their most prevalent common feature is that they are enclitic to the infinitive. They ended up being close to central auxiliaries because they are prone to intervene between the preverb and its associated infinitive, as shown in (6).

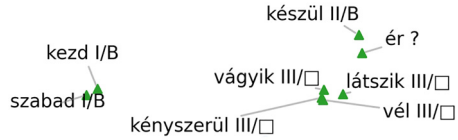

Figure 3. Verbs of Cluster 2

(6) a. ***vissza*** *vágy-nak* ***tér-ni*** *a* *piac-ai-nk-ra*
back(PV) wish-PRES.3PL/INDEF come-INF the market-PL-1PL.POSS-SUBL
'they wish to return to our markets'

b. ***el*** *látsz-ott* ***vesz-ni*** *a* *magas* *terem-ben*
away(PV) seem-PAST.3SG get.lost-INF the high hall-INE
'(he/she) seemed to be lost in the high hall'

c. ***fel*** *vél-t-em* ***fedez-ni*** *között-ük*
up(PV) think-PAST-1SG discover-INF among-3PL
'I thought that I discovered him/her among them'

d. ***el*** *kényszerül-t* ***hagy-ni*** *a* *tudomány-t*
away(PV) be.obliged-PAST.3SG leave-INF the academia-ACC
'he had no choice but to leave academia'

According to Kálmán C. et al. (1989), these verbs appear interposed only in vernacular use, but our corpus data do not clearly show this tendency. Modrián (2009) also reports that examples of interposition are attested in texts belonging to the 'belletristic' and 'press' registers of the Hungarian Gigaword Corpus.

Two further members of Cluster 2 are *kezd* 'begin' and *szabad* 'be allowed', labeled as peripheral auxiliaries in Kálmán C. et al. (1989). The verb *készül* 'prepare' can also be found here, the evaluation of which was quite different in Kálmán C. et al. (1989). In our data, it can be attested in an intervening position but more dominantly as being enclitic to the infinitive; see (7a). Finally, there is *ér* 'be allowed' that did not appear in Kálmán C. et al.'s classification. The explanation for this lies in the composition of our data source. Most of the corpus belongs to the 'personal' register, mainly social media. In these texts, phrases like the ones in (7b–d) are prevalent.

(7) a. ***indul-ni*** *készül-t-em*
leave-INF prepare-PAST-1SG
'I was about to leave'

b. ***meg-oszt-ani*** *ér*
PV-share-INF be.allowed-PRES.3SG/INDEF
'it is OK to share (the content)'

c. ***lájkol-ni*** *ér*
like-INF be.allowed-PRES.3SG/INDEF
'it is OK to give a "like"'

d. ***röhög-ni*** *persze* *ér*
laugh-INF of.course be.allowed-PRES.3SG/INDEF
'it is OK to laugh about it'

In **Cluster 3** (shown in Figure 4), the following pattern can be seen. The upper left segment is populated by lexical items that Kálmán C. et al. (1989) classify as emphatic verbs (e.g., *utál* 'hate', *imád* 'adore', *szeret* 'love, like'). Their points are located relatively densely, close to each other in the low-dimensional space. The lower left segment is taken by verbs that are central or peripheral auxiliaries, according to Kálmán C. et al. (1989). Here we see more scattered points, i.e., their behavior is less uniform (e.g., *tetszik* 'indicates politeness', *mer* 'dare', *lehet* 'may').

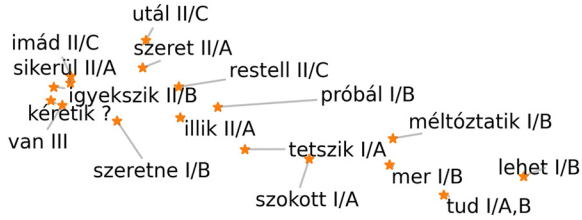

Figure 4. Verbs of Cluster 3

The main common feature of Cluster 3 is the high frequency of the finite verb's initial position. Indeed, this is what we expect in the case of verbs appearing in the upper left segment. However, it is surprising in the case of right-appearing verbs. We suspect that the lack of direct information about prosody led to this result, and many sentences in our sample could be pronounced with a focus stress, cf. (8). At the same time, these verbs often appear in an intervening or enclitic position (as we would expect). Therefore, they are adjacent to Clusters 1 and 2, mainly characterized by the latter positions.

(8) a. *Szok-t-ak* ***len-ni*** *nagy* *leértékelés-ek.*
used.to-PAST-3PL be-INF big discount-PL
'There used to be big sales.'
b. *Lehet* *rá* ***számít-ani.***
be.PRES.COND.3SG he/she.SUBL count-INF
'You can count on him/her.'
c. *Mer-em* ***remél-ni.***
dare-PRES.1SG/DEF hope-INF
'I hope so.' (lit. 'I dare to hope.')

Finally, two more members of Cluster 3 must be discussed. The first one is *kéretik* 'be asked for', which was not included in Kálmán C. et al. (1989). Our corpus data shows it is clearly a stress-requiring verb, see (9). The second is *van* 'be', which has a frequent, emphatic use in existential contexts.

(9) a. *Kér-et-ik* *máskor* *pontos-abb-an* ***fogalmaz-ni.***
ask-CAUS-PRES.3SG next.time exact-COMPR-ADVLZ word-INF
'You are asked to be more specific next time.'
b. *Van* *mi-ről* ***beszél-n-ünk.***
be.PRES.3SG what-DEL talk-INF-1PL
'There are things for us to talk about.'

Cluster 4 is quite mixed compared to Kálmán C. et al. (1989). The most prevalent common feature of these verbs is that they are frequently used as infinitives (see Figure 5). That is, they are the least auxiliary-like regarding their low position on the finiteness scale. Most of the verbs in this cluster are labeled affix-like verbs in Kálmán C. et al. (1989).

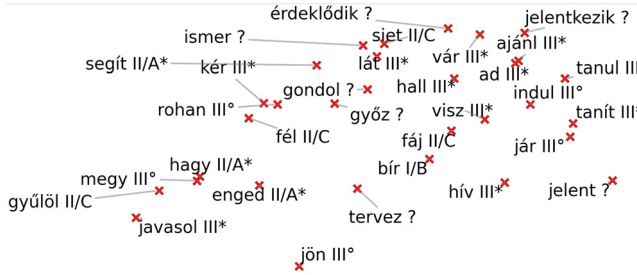

Figure 5. Verbs of Cluster 4

This cluster comprises every verb of motion in our sample (*indul* 'set off', *jár* 'used to go', *megy* 'go', *jön* 'come', *rohan* 'rush'). In addition, we find here every double agent verb. That is, *enged* 'let', *hagy* 'let, leave' and *segít* 'help' are no exceptions, contrary to the results of Kálmán C. et al. (1989). One unexpected result is the position of *bír* 'cope with', a verb considered a peripheral auxiliary in Kálmán C. et al. (1989). When this verb is used in finite form, it takes intervening or enclitic positions (as expected). However, it is more frequent as an infinitive complement of other, more auxiliary-like verbs, e.g., *bírni kell* 'it must be coped with', *bírni fogom* 'I will cope with it'.

6. CONCLUSION AND FUTURE WORK

This paper examined the distributional patterns of 60 finite verbs with an infinitive complement. The basis of our study was a sample of 12,287 sentences extracted from the corpus. By hierarchical agglomerative clustering, the verbs were grouped into four clusters, two of which comprise the most auxiliary-like verbs according to our criteria. We compared the results of our experiment to those of Kálmán C. et al. (1989). Our results are by and large similar. However, there are some rather unexpected differences as well. The most striking one is that the stress-avoiding verbs are prone to intervene between the preverb and its associated infinitive, which indicates that they are much closer to central auxiliaries than previously assumed. We would also like to point out that double agent verbs show uniform behavior based on our data, contrary to Kálmán C. et al.'s (1989) results.

An important lesson we learned is the following. No matter how carefully we preprocess the data, we cannot accurately filter the environments (so that only a specific set of neutral sentences remains) relying on the currently available corpus annotation. There is no one-to-one correspondence between prosody and word order, which causes difficulties if one has access to written text only.

The corpus-based approach presented here can be extended and continued in several directions. As highlighted above, prosody information is only partially discernible from word order. A possible line of further research could be manually annotating the prosodic patterns in a small text sample and training a language model on it to gain a large corpus annotated for prosody. If the resulting data are reasonably high quality, it becomes possible to use sentence prosody as a feature in a cluster analysis similar to the one presented here. Some of the potentially relevant

features of auxiliaryhood described in Kenesei (2001), e.g., defective paradigm and the impossibility/rarity of nominalization, can also be studied on corpus data.

In the current experiment, we decided to completely and deliberately ignore the lexical-semantic properties of verbs since semantic information is not directly available in the corpus. Nevertheless, annotating smaller samples of text with various semantic information could be instructive. These parameters, alongside the formal and distributional ones studied here, would make it possible to create behavioral profiles of these more or less auxiliary-like verbs; see Gries (2007) and Divjak & Gries (2006) for research in a similar vein. Related to this, we have to mention Bajzát (2022a). This study aims to explore infinitive constructions involving five auxiliary-like items (*tud* 'can, know', *akar* 'want', *szeret* 'love', *kíván* 'wish', *képes* 'able to') using cluster analysis on a nearly 2000-sentence sample, which is manually annotated for a wide range of semantic features. Combining this feature set with ours might yield further interesting results.

Finally, a promising direction of future research could be to conduct experiments of this sort on diachronic corpus data, which could outline the process of auxiliarization in Hungarian. There are qualitative studies concerning the grammaticalization of main verbs into auxiliaries – notably, Tolcsvai Nagy (2010) – but this topic has hardly been investigated quantitatively. As far as we know, the only study of this sort is Bajzát (2022b), focusing on the Middle Hungarian period. Quantitative research would be possible to a greater extent since a considerable amount of corpus data is available for every historical period of Hungarian, from the Old Hungarian period to the present day.

ACKNOWLEDGMENT

We thank two anonymous reviewers for their helpful comments on an earlier version of this paper. Ágnes Kalivoda's research has been supported by the OTKA PD project No. 142317, funded by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the PD 22 funding scheme.

REFERENCES

Bajzát, Tímea Borbála. 2022a. A premodális tartományokkal összekapcsolódó segédige/melléknév + főnévi igenév konstrukciók mondatszintű szemantikai vizsgálata [Sentence-level semantic analysis of auxiliary verb/adjective + infinitive constructions connected to premodal domains]. In S. Tátrai and G. Tolcsvai Nagy (eds.) A magyar mondat és kontextuális környezete [The Hungarian sentence and its context]. Budapest: Eötvös Loránd University (ELTE), Faculty of Humanities. 141–188.

Bajzát, Tímea Borbála. 2022b. A premodális tartományokkal összefüggő főnévi igeneves mintázatok klaszterezési lehetősége a történetiség perspektívájában [Clustering possibilities of infinitive constructions connected to premodal domains in a diachronic perspective]. In T. Forgács, M. Németh and B. Sinkovics (eds.) A nyelvtörténeti kutatások újabb eredményei [The most recent results of diachronic studies], Vol. 11. Szeged: SZTE BTK, Department of Hungarian Linguistics. 25–49.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal dependencies. Computational Linguistics 47(2). 255–308.

Divjak, Dagmar and Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. Corpus Linguistics and Linguistic Theory 2(1). 23–60.

É. Kiss, Katalin. 1987. Configurationality in Hungarian (Studies in Natural Language and Linguistic Theory 3). Boston, MA: D. Reidel Publishing Company.

É. Kiss, Katalin. 1992. Az egyszerű mondat szerkezete [The structure of the simple sentence]. In F. Kiefer (ed.) Strukturális magyar nyelvtan 1: Mondattan [Structural Hungarian grammar 1: Syntax], Vol. 1. Budapest: Akadémiai Kiadó. 79–177.

É. Kiss, Katalin and Henk Riemsdijk. 2004. Verb clusters: A study of Hungarian, German and Dutch. Amsterdam: John Benjamins Publishing Company.

Flyamer, Ilya. 2018. adjustText – Automatic label placement for matplotlib. Python. https://github.com/Phlya/adjustText.

Gries, Stefan Th. 2007. Corpus-based methods and cognitive semantics: The many senses of to run. In S. T. Gries and A. Stefanowitsch (eds.) Corpus-based approaches to syntax and lexis. Berlin & New York, NY: De Gruyter Mouton. 57–100.

Hees, Jörn. 2015. SciPy hierarchical clustering and dendrogram tutorial. Jörn's Blog. https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/. (March 20, 2023.)

Heine, Bernd. 1993. Auxiliaries: Cognitive forces and grammaticalization. Oxford: Oxford University Press.

Hunter, John D. 2007. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 9(3). 90–95.

Imrényi, András. 2013. The syntax of Hungarian auxiliaries: A dependency grammar account. In E. Hajičová, K. Gerdes and L. Wanner (eds.) Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013). Prague: Charles University in Prague, Matfyzpress. 118–127.

Kalivoda, Ágnes, Robert Malouf, Farrell Ackerman and Bálint Sass. 2023. Building a dependency treebank from the Hungarian Gigaword Corpus. Manuscript.

Kálmán C., György, László Kálmán, Ádám Nádasdy and Gábor Prószéky. 1989. A magyar segédigék rendszere [The system of auxiliaries in Hungarian]. In Z. Telegdi and F. Kiefer (eds.) Általános Nyelvészeti Tanulmányok [Studies in General Linguistics], Vol. XVII. Budapest: Akadémiai Kiadó. 49–103.

Kenesei, István. 2001. Criteria for auxiliaries in Hungarian. In I. Kenesei (ed.) Argument structure in Hungarian. Budapest: Akadémiai Kiadó. 73–106.

Kenesei, István, Robert M. Vago and Anna Fenyvesi. 1998. Hungarian (Descriptive Grammars). London: Routledge.

Komlósy, András. 1989. Fókuszban az igék [Verbs in focus]. In Z. Telegdi and F. Kiefer (eds.) Általános Nyelvészeti Tanulmányok [Studies in General Linguistics], Vol. XVII. Budapest: Akadémiai Kiadó. 171–182.

Laczkó, Tibor. 2014. On verbs, auxiliaries and Hungarian sentence structure in LFG. Argumentum 10. 421–438.

McKinney, Wes. 2010. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference. 56–61.

Modrián-Horváth, Bernadett. 2009. Gesichtspunkte zu einer funktionalen Typologie der ungarischen Infinitiv regierenden Hilfsverben. Acta Linguistica Hungarica 56(4). 405–439.

Moisl, Hermann. 2015. Cluster analysis for corpus linguistics. Berlin: De Gruyter Mouton.

Novák, Attila, Borbála Siklósi and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In N. Calzolari et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA). 1315–1322.

Oravecz, Csaba, Tamás Váradi and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In N. Calzolari et al. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik: European Language Resources Association (ELRA). 1719–1723.

Pedregosa, Fabian. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12. 2825–2830.

Prószéky, Gábor, György Kálmán C., László Kálmán and Ádám Nádasdy. 1984. Topic, focus, and auxiliaries in Hungarian. Groninger Arbeiten zur germanistischen Linguistik 24. 162–177.

Tolcsvai Nagy, Gábor. 2010. The auxiliary + infinitive construction in Hungarian. Acta Linguistica Hungarica 57(1). 143–164.

Tompa, József (ed.). 1961. A mai magyar nyelv rendszere [The system of contemporary Hungarian]. Budapest: Akadémiai Kiadó.

Van Rossum, Guido and Fred L. Drake. 2009. Python 3 reference manual. Scotts Valley, CA: CreateSpace.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau and Evgeni Burovski et al. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17. 261–272.