doi:10.5128/ERYa19.13

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 19, 225–244

TOWARDS THE MORPHOSYNTACTIC CORPUS PROFILE OF PROTOTYPICAL ADJECTIVES IN ESTONIAN

Ene Vainik, Geda Paulsen, Ahti Lohk, Maria Tuulik

Abstract. The transition zones between traditional word classes cause problems in lexicography. This research addresses the issue of estimating the level of adjectivization in Estonian by proposing a set of close-context indicators ("test patterns") based on the existing literature and detectable in annotated corpus text. The profile of prototypical adjectives (the "reference profile") is established by analyzing the normalized frequencies of the test patterns in a random sample of validated adjectives (N = 100). A scale of similarity to the reference profile is established by using the method of calculating Euclidean distances, which is considered a heuristic of the cumulative similarity vs. the difference. As a result, the scalar nature of the similarity to the reference profile is revealed, among both validated adjectives and the control group of yet underspecified lexicographic headword candidates (N = 100). The results are discussed in respect to improving the toolbox of the test patterns as well as in respect to future studies on some intriguing features of the actual corpus behavior of adjectives as compared to what would be expected by their morphosyntactic potential described in the literature.*

Keywords: lexicography, corpus linguistics, adjectives, lexical decategorization, Estonian

1. Introduction

Language is in constant flux, dynamic in its manifestations. This raises the issue of changes and multiplicity in word class affiliation in dictionary making. The Estonian lexicographers have described adjectives as one of the most problematic word classes due to its transition areas overlapping with nouns, verbs, and adverbs (Paulsen et al. 2019: 188–189, Vainik et al. 2020: 122–123). The present study focuses on the categorization issues around adjective candidates, in particular from the point of view of participles obtaining adjectival properties. It is common even in English, for example, to find an established adjective derived from a verbal participle, e.g., *interesting*.

* This research was supported by Estonian Research Council grant PRG 1978.

The large text corpora provide the lexicographers with a vast amount of raw data, organized as concordances and word sketches (see Kilgariff et al. 2004, Kilgariff et al. 2014). However, lexicographers have expressed a desire for specific tools that would provide statistical summaries of a word form's behavior as compared to the typical, or central, members of the word classes (Paulsen et al. 2019). This practical call has evoked a series of studies dedicated to the possibilities of defining and statistically measuring word-class-specific behaviors in the corpus in terms of scalable numerical heuristics (Paulsen et al. 2021, Vainik et al. 2021, Tuulik et al. 2022).

The aim of the present study is to delineate the defining features of adjectives detectable in an annotated corpus text and to establish a behavioral profile of prototypical adjectives in Estonian as a standard for comparison with so far unclassified word forms. We will also attempt to elaborate further on the statistical measure as a heuristic and the scale of similarity with adjectives.

The structure of the paper follows the steps taken in the study: after a few pieces of background information, we give an overview of the most relevant morphosyntactic properties of adjectives mentioned in the literature. A separate section is dedicated to the methods and materials. The reference profile of prototypical adjectives is established in the section of results and the similarities of the members of the reference group and the control group are measured there, too. The findings are discussed and the conclusions are presented in the final section of the paper.

2. Background

The current trend in Estonian lexicography is toward the unification of sparse dictionaries and term bases into a central database (called Ekilex, see Hein et al. 2020) handled via the dictionary writing system carrying the same name, while the aggregated super-dictionary is called EKI Combined Dictionary (CombiDic) and is accessible via the language portal Sõnaveeb[1] (Tavast et al. 2018, Koppel et al. 2019, Tavast et al. 2020). The aggregation has resulted in a set of somewhat unequally specified entries, as the material has been automatically derived from different kinds of sources with different foci of specification.

There is an urgent need to provide the as yet underspecified headword candidates in the Ekilex database with PoS tags, as the slot of PoS affiliation is included in the data model (Tavast et al. 2018). As much as 72% of the public CombiDic keywords (N = 255 691) lacked PoS tags at the time of starting this research.[2] The number of participle-like word forms among them was substantial (1542). We aim to contribute to the exploration of whether those words should be treated as regular members of the respective verb paradigms or have potentially acquired adjectival features, in which case they could be tagged as adjectives.

We do not assume all of the underspecified participle-like headwords to share similar status but expect a continuum of similarity with adjective-like behavior. This expectation arises from our theoretical understanding that follows the prototype-based model of categorization first described in psychology (Rosch 1973, 1975, 1978) and later also employed in linguistics (see e.g., Berlin, Kay 1969, Geeraerts 1989, in Estonian linguistics e.g., Erelt 1977).

[1] https://sonaveeb.ee

[2] The estimation is based on an extraction from all Ekilex databases (including dictionaries, term bases and phrase collections) on 24.1.2022. We thank Kaur Männiko for the excerpt.

The boundaries of prototype-based categories are not expressly defined, and the members of a category may have different statuses: there may be more typical, “better” examples of a category than others. A prototype can also be described via a bundle of features, none of which is necessary or sufficient for defining the whole category. In the present study, we attempt to use some of the characteristic features of adjectives described in the literature as constituting the profile of adjectives as a word class. One of the central notions in this study is, thus, the theoretical construct of a prototypical adjective, by which we mean an adjective that most clearly displays the morphological, syntactic and semantic properties ascribed to this word class in the linguistic literature.

2.1. The main features of adjectives in Estonian

The word class of adjectives can be found in every human language (Dixon 2004: 1). Adjectives do not take major syntactic positions in a sentence but occur in an attributive or predicative relation to the subject or object, modifying nouns. Semantically, adjectives describe the phenomena referred to by nouns. The syntactic and semantic features are accompanied by morphological features in Estonian, a morphologically rich language, which includes inflection, forms of gradation, and derivation.

Syntactically, an adjective constitutes a phrase by itself or together with its modifier(s), occurring as an attribute (1a), predicative (1b) or predicative adverbial (1c). The primary function of the adjective is attributive, where the components of the adjective phrase are most clearly recognizable (Erelt 2017a: 406). Estonian favors prenominal attributes (Pajusalu 2017: 382). Adjectives as attributes agree with their head nouns in case and number (as in 1a) except for the terminative, essive, abessive and comitative cases, which require the genitive of the adjective attribute (Viitso 2001: 35). The predicative modifies the subject via the copula verb *olema* ‘be’ and usually takes the nominative case but also the partitive, genitive[3] and elative cases are possible. The predicative adverbial typically expresses a result state and occurs in translative, essive or nominative case in connection with a range of verbs of change[4] (Erelt 2017c: 286–287, Erelt 2017d: 289, Erelt 2017a: 405). An adjective can be modified by an adverb (1d).

(1a) *Karud elavad paksu-de-s metsa-de-s*
bear-PL live-3PL thick-PL-INE forest-PL-INE
‘The bears live in thick forests.’

(1b) *Laps on väga rõõmus.*
child is very glad-NOM
‘The child is very glad.’

(1c) *Laps muutus rõõmsa-ks.*
child became glad-TRA
‘The child became glad.’

[3] The head noun of the Estonian predicative may occur in the comitative and abessive cases, requiring the genitive case of its attributive adjective, as in *Tüdruk on heledate juustega* [girl is fair-coloured-GEN hair-COM], lit. ‘the girl has fair-coloured hair’; and *Väide on kindla tõestuseta* [statement is firm-GEN proof-ABE], lit. ‘The statement is without firm proof’ (Erelt 2017c: 281–282).

[4] These are verbs: *kujunema* ‘turn’, *muutuma* ‘change’, *minema* ‘go’, *saama* ‘get’ etc.

(1d)	*äärmiselt*	*raske*
	extremely	difficult

The semantics of an adjective influences its gradability, i.e., the ability to derive comparative and superlative forms: an adjective generally allows for comparison if it encodes a scalar (degree) property, e.g., *nõrk* 'weak'. The comparative forms are marked by the suffix *-m* (*nõrgem* 'weaker' and the superlative with the suffix *-im* (*nõrgim* 'the weakest'); it is also possible to use the analytic superlative construction *kõige nõrgem* 'the weakest, lit. the most weaker'.

The semantic distinction between relative (scalar) vs. absolute (non-scalar) properties also affects the structure of the adjective phrase: scalar adjectives can be modified by intensifying adverbs (*väga nõrk* 'very weak' cf. ?*väga lingvistiline* 'very linguistic') (see Erelt 2017a: 406–408).

As the prototype model predicts, not all of the members of the adjective class share all of those features mentioned above (Viks 1977). For example, there is the atypical subclass of non-declinable adjectives that do not agree with their head nouns when used as attributes (e.g., *väärt sõbrale* [good friend-ALL] 'to the good friend'). There is also a subclass of gradation-defective adjectives, which do not derive comparative forms despite the lack of morphophonological or semantic constraints (e.g., *kaarjas* 'curve-like'; see Viht, Habicht 2019: 27). On the other hand, non-scalar adjectives can occasionally be modified by an adverb in a suitable context (e.g., *peaaegu kolmnurksed lehed* 'almost triangular leaves).

2.2. The adjectival features of participles

An example of a word class sharing several semantic and morphosyntactic features typical of adjectives are participles: the non-finite verbal forms that occupy the transition zone between verbs and adjectives. Participles have been defined as verb-derived adjectives within a verbal paradigm (Haspelmath 1994: 152). In Estonian, there are different endings for present (e.g., personal *tantsi-v* 'dancing'; impersonal *tantsi-tav* 'being danced') and past tenses (personal *tantsi-nud* '(has) danced'; impersonal *tantsi-tud* '(was) danced'). The suffixes function partly as verbal endings and partly as derivational suffixes yielding new lexemes (e.g., Viht, Habicht 2019: 37). Interestingly, both present and past participles can be used in the role of attribute or predicative in a sentence. Present participles can be inflected for case and number, agreeing with the head noun exactly as typical adjectives do. The past participles are non-declinable and resemble atypical (non-declinable) adjectives when used as attributes. In verbal use, participles occur together with the finite verb forms of the verb *olema* 'be' to form compound tenses, in which case it can be difficult to distinguish them from the role of the predicative, as they also use the copula verb *olema* 'be'. Common to all participles in Estonian is that it is possible to regularly form comparative and superlative (Kerge 1998, Kasik 2015: 369).

All of the features mentioned above make it difficult to categorize the PoS affiliation of participles, particularly because they can display their verbal and adjectival potentials depending only on the context. The idea of the current study is to try to develop a methodology for making distinctions based on statistics, i.e., the relative frequency of a word form's adjectival behavior as indicating its degree of adjectivization. We propose a series of close-context indicators of adjectival behavior (called “test patterns”).

In the following, we specify the behavioral profile of the prototypical adjectives and compare them with the corresponding values of a group of participles. In this regard, we target the underspecified (PoS-wise) word forms of the Ekilex database as a control, in order to see whether and to what degree the more adjectivized participles are distinguishable.

3. Materials and methods

The overall procedure consists of several steps. Data retrieval consists of 1) operationalizing the features of adjectival behavior into detectable patterns, 2) forming a reference group of adjectives and a control group of lexicographically underspecified participles, and 3) the retrieval of frequency data from the corpus. Then data analysis elicits: 1) normalizing the corpus frequencies, 2) establishing the profile of typical adjectival behavior, 3) a comparison of individual profiles of the words of both the reference and control groups, and 4) establishing a scale of adjectivity usable as a heuristic in lexicographic work. The methods of data retrieval are described in this section and those of data analysis are introduced in the Results section, where needed.

3.1. The test patterns of adjectival corpus behavior

The patterns capture the most characteristic features of adjectives: being used as a prenominal attribute, being used as a predicative in connection with the copula verb *olema* ‘be’, the possibility of being modified by a preceding adverb and the ability for (semantic) gradation. General attributive behavior is further defined by more specific conditions, such as the attribute’s agreement with its head noun (in case and number) and the condition of being placed right at the beginning of the sentence (see Tuulik et al. 2022 for justification).

The test patterns are composed mostly using the adjacency positions (bi-grams) occurring in the annotated corpus text. However, we leave the slot of a potential adjective unspecified in our coding as the patterns are targeted to discover instances of adjectival behavior possibly not accounted for as such in the existent tagging system.

The operationalization results in semi-specified test patterns (see Table 1) where the test word (TW) stands for the unit whose behavior is searched for. The pattern of searching for the comparative forms is exceptional: in that case, we manually derived the respective forms and extracted the lemma frequencies of comparative forms from the corpus. Also, the existence of comparative forms was taken as proof that there was gradation.

The toolbox of test patterns is a replication of the previous study (Tuulik et al. 2022), except for a complex attributive pattern (a condition of four sequential items) that was left out as too rare and not informative in many cases. The test pattern designed for catching predicative behavior is a new one that was missing from the first approximation.

Table 1. The features of adjectival behavior operationalized into the test patterns (TW = test word, the word which's behavior is studied)

Feature	Test pattern's name and abbreviations	Condition	Example	Test pattern
Attributive	attribute (general) – ATTR	A test word immediately precedes a noun	***roosa*** *jakk* 'pink jacket'	(TW ∧ noun)
	attribute (in agreement) – ATTR/AGR	A sequence of the test word in the same case and number as the following noun	***suures majas*** big-SG-INE house-SG-INE 'in a big house'	(TW ∧ noun)$_{\text{agreement}}$
	attribute (sentence starter) – ATTR/ST	A test word starts a sentence in the attributive position, followed by a noun.	***Roosa*** *jakk rippus…* 'The pink jacket was hanging…'	(TW ∧ noun)$_{\text{sentence start}}$
Predicative	predicative – PRED	A test word occurs directly after the Estonian copula verb *olema* 'be' or after the sequence *olema* and adverb	*Maja* ***on suur*** 'The house is big'	∀ *olema* 'be' ∧ TW* ∀ *olema* 'be' ∧ (adverb ∧ TW)
Being modified by an adverb	adverb – ADV	An adverb precedes the test word	***väga suur*** 'very big'	(adverb ∧ TW)
Gradability	comparative form – COMP	Presence of the comparative form	***suurem*** 'bigger'	∃ comparative

3.2. Forming the reference group and the control group

The reference group (N = 100) of adjectives was derived from the Basic Estonian Dictionary[5] (Kallas et al. 2014, see also Kallas, Tuulik 2011) by random sampling among headwords labeled as adjectives. We considered those as central and good examples of the class because it is a learner's dictionary presenting the most central and topical vocabulary; also, the PoS affiliation of the headwords has been validated by the lexicographers.

The control group (N = 100) of participle-like word forms was derived from the Ekilex database (from the list of as-yet underspecified headwords/candidates lacking the PoS tag[6]) and by random sampling. The sample was controlled, though, for the balance of the different types of participles (both present and past, personal and impersonal; see section 2.2).

3.3. Extracting test patterns from the corpus

The data was retrieved from the Estonian National Corpus 2019 (ENC 2019)[7], which is an annotated corpus of 1.5 billion tokens (Koppel, Kallas 2022) pre-tagged (i.e.,

[5] The dictionary contains the 5000 most frequent and central Estonian words explained in simple language; the number of adjectives is 554.

[6] The PoS-tagging status of CombiDic headwords constantly changes as the dictionary is updated. The date of extraction of our data is 24.1.2022.

[7] While all of the ENC-corpora are stored in the corpus query system Sketch Engine (see Kilgarriff et al. 2004, Kilgarriff et al. 2014), we use the files exported from the Sketch Engine cloud to the home page of the Center of Estonian Language Resources (https://entu.keeleressursid.ee). The frequency results may sporadically differ between the two data sources, up to1% (Neeme Kahusk, private conversation).

tokenized, lemmatized, morphologically analyzed and disambiguated) with the EstNLTKv.1.6 (Orasmaa et al. 2016: 2460, Laur et al. 2020), the precision of tagging being estimated at approximately 0.94 (Kaalep et al. 2012).

A special code (written in the Python programming language)[8] was run to extract the frequency data of the instances of matching the test patterns, as well as the lemma frequencies of all of the words in the reference group and in the control group. The extractor was set to find the patterns only within the sentence boundaries. The expressions are described in Table 2.

Table 2. The logical expressions for extracting the corpus patterns (TW = test word, the word which's behavior is studied)[9]

Patterns	**Test pattern**	**Logical expression in programming language of Python**
ATTR	(TW ∧ noun)	i < sent_len – 1 and lemmas[i].lower() in test_words and postags[i+1] == 'S'
ATTR/AGR	(TW ∧ noun)$_{\text{agreement}}$	i < sent_len – 1 and lemmas[i].lower() in test_words and postags[i+1] == 'S' and forms[i] == forms[i+1]
ATTR/ST	(TW ∧ noun)$_{\text{sentence start}}$	i < sent_len – 1 and i == 0 and lemmas[i].lower() in test_words and postags[i+1] == 'S'
ADV	∀ *olema* 'be' ∧ TW* ∀ *olema* 'be' ∧ (adverb ∧ TW)	i < sent_len – 1 and postags[i] == 'D' and lemmas[i+1].lower() in test_words
PRED	(adverb ∧ TW)	i < sent_len – 2 and lemmas[i].lower() == 'olema' and postags[i+1] and lemmas[i+2].lower() in test_words
COMP	∃ comparative	lemmas[i].lower() in comp_words

4. Results

The raw data of the test pattern frequencies were normalized, i.e., the number of test pattern matches was divided by the lemma frequencies of respective words[10]. In the following analysis, we operate only with the relativized values of test pattern instances.

4.1. The reference profile of validated adjectives

One of the aims of the study is to establish a behavioral profile of adjectives as a word class to be used as a standard while evaluating any other word form's behavior regarding its potential adjectivization. We expected the group of validated adjectives to be a reliable reference group for compiling such a profile. The results show, however, that there is considerable variance among the 100 random adjectives in respect to meeting the criteria of the test patterns (see Figure 1). Table 3 provides the details of descriptive statistics.

[8] https://github.com/ahtilohk/PSG227/blob/main/Test-patterns_occurrences_in_ENC2019_without_estnltk_corpus_processing_module.py

[9] Generally, lemmas are used to account for the test patterns (*lemmas[i].lower() in test_words*). Exceptionally, text words are used to account for the words with the endings 'dud', 'nud', 'tud', i.e., past participles that basically do not inflect (*[i].lower() in test_words* in logical expression *len(word) > 3 and word[-3:] in ['nud', 'dud', 'tud'] and word in test_words*). The code for extracting comparative forms is a co-product of the general code that searches for the lemma frequencies on the basis of a manually composed list of comparative forms.

[10] The data file of raw and relativized data and analysis is available at https://github.com/ahtilohk/PSG227/blob/main/Data_File_English_Descrption_Euclidean_distance_23.01.23.xlsx

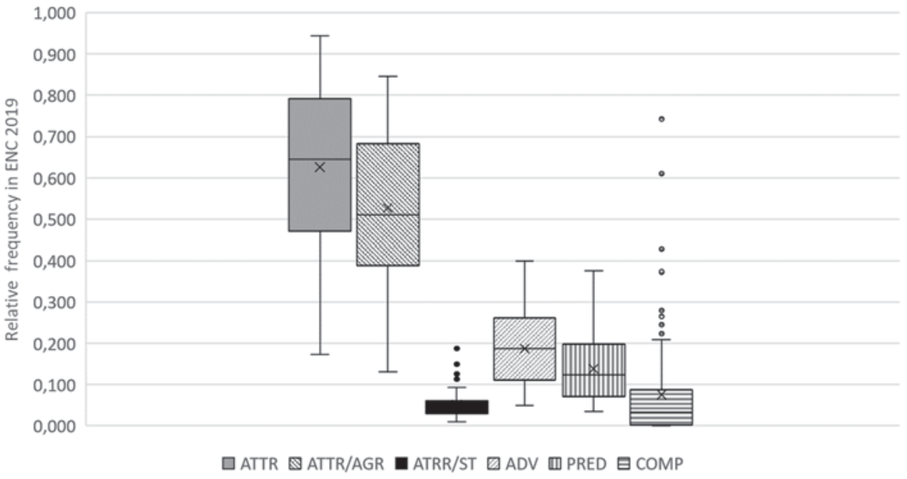

Figure 1. General distribution of the data of the 100 validated adjectives across the test patterns

Table 3. Descriptive statistics of the 100 validated adjectives across the test patterns (min = Minimum value, max = Maximum value, ave = Average, StDev = Standard deviation)

Statistics	ATTR	ATTR/AGR	ATTR/ST	ADV	PRED	COMP
min	0.173	0.132	0.009	0.049	0.034	0.000
max	0.944	0.846	0.193	0.399	0.376	0.742
ave	0.626	0.527	0.051	0.188	0.138	0.076
stDev	0.192	0.185	0.034	0.084	0.074	0.120
median	0.644	0.510	0.042	0.187	0.124	0.034

It is apparent that two of the patterns measuring attributive behavior (ATTR and ATTR/AGR) occur much more frequently, on average, than do the rest of the patterns. This finding confirms the theoretical assumption that the attributive role is dominant for adjectives (cf., section 2.1 and the reference to Erelt 2017a: 406). The general attributive pattern also demonstrates the biggest range of variance, thus revealing the existence of adjectives that score very high in the attributive role, as well as those that seldom do. The predicative role appears to be approximately four times less frequent, generally, than the attributive. The patterns targeted at measuring occurrence in predicative function (PRED) and the feature of being modifiable by an adverb (ADV) demonstrate moderate values on average and lesser ranges of variance. Two of the patterns – the one measuring attributive role right at the beginning of a sentence (ATTR/ST) and the one counting the comparative forms (COMP) – show controversial results. They fall at very low levels, generally, but also demonstrate that some of the adjectives score higher than the rest (see the outliers in Figure 1).

The lower level of the sentence-initial pattern is natural considering that the number of sentences in a corpus is lower than the number of bi-grams, which diminishes the probability of a bi-gram occurring in such a position. The adjectives scoring higher in the sentence-initial position have sparse semantic content with deictic properties and occasional pragmatic loads due to the layout of the information structure (e.g., *harilik* 'ordinary', *tänane* 'today's', *eelmine* 'previous', and *esialgne* 'initial').

The finding that the potential for gradation is hardly represented in our data might be explained by the predominance of non-scalar adjectives (e.g., *homne* 'of tomorrow', *vasak* 'left' and *järgmine* 'next'), which is not the case; there are only two words that produce no comparative forms: *kahekordne* 'double, two-floored' and *ühetoaline* 'one-room'. The low mean is due to the low relative frequencies of the comparative forms in general. The exceptional adjectives with the higher proportions of comparatives are *täpne* 'precise', *lahja* 'lean', *kõrge* 'high' and *lihtne* 'simple' (seen as outliers in Figure 1). The words represent qualities whose scale is more frequently discussed in the corpus.

The variance in our data is illustrated in Table 4, which presents a selection of adjectives with both attributive and non-attributive roles as their behavioral dominants. These are rather clear examples of distributional complementarity: it appears that among the validated adjectives there was a tendency to be biased to either more attributive or more predicative and modified usage.

Table 4. Examples of adjectives demonstrating the dominance of attributive vs. non-attributive patterns (the values higher than average are presented in bold in each column)

Adjective	ATTR	ATTR/AGR	ATRR/ST	ADV	PRED	COMP	**dominance**
sõjaline 'military'	0.94	**0.85**	0.03	0.08	0.03	0.00	attr
pidulik 'festive'	**0.89**	**0.76**	**0.09**	0.10	0.04	0.07	
tehniline 'technical'	**0.92**	**0.84**	**0.07**	0.10	0.05	0.01	
erialane 'specialized'	**0.86**	**0.76**	0.04	0.10	0.05	0.00	
kevadine 'vernal'	**0.82**	**0.71**	**0.11**	0.10	0.05	0.01	
füüsiline 'physical'	0.84	**0.77**	**0.05**	0.13	0.05	0.00	
alaline 'permanent'	**0.88**	**0.78**	0.04	0.06	0.06	0.00	
õudne 'awful'	0.37	0.33	0.04	**0.32**	**0.21**	0.08	non-ATTR
kasulik 'useful'	0.38	0.31	0.04	0.22	0.24	0.11	
lihtne 'simple'	0.36	0.28	0.03	**0.31**	**0.30**	**0.43**	
keeruline 'complicated'	0.34	0.27	0.02	**0.40**	**0.31**	**0.28**	
selge 'clear'	0.26	0.20	0.01	**0.20**	**0.27**	0.07	
kannatlik 'patient'	0.25	0.21	0.02	**0.21**	**0.38**	0.07	
kade 'envious'	0.17	0.13	0.01	**0.27**	**0.29**	0.00	

Table 4 presents only a fraction of the whole spectrum of variance. It is clear, however, that the higher-than-average values co-occur across the patterns measuring directly attributive behavior (ATTR and ATTR/AGR, less so for ATTR/ST), while the higher-than-average values accumulate alternatively on the patterns measuring predicativity (PRED), modifiability by an adverb (ADV) and gradability (COMP). The actual complementarity and/or co-variation among our test patterns is best revealed by a matrix of statistical correlations (see Table 5).

Table 5. Correlations between the patterns in the data set of validated adjectives. Only the statistically significant correlations (p = 0.000) are presented. Correlations stronger than 0.6 are in bold

Patterns	**ATTR**	**ATTR/AGR**	**ATTR/ST**	**ADV**	**PRED**	**COMP**
ATTR	1.00	0.97	0.60	–0.68	–0.71	
ATTR/AGR		**1.00**	**0.61**	**–0.66**	**–0.69**	–0.25
ATTR/ST			**1.00**	–0.51	–0.53	
ADV				**1.00**	**0.63**	0.40
PRED					**1.00**	0.23
COMP						1.00

The matrix shows rather strong positive and negative correlation coefficient values; it is only the pattern measuring gradation (COMP) that shows weak or no correlations and is, thus, a relatively independent characteristic. All of the attributive patterns are mutually positively correlated, which is a natural outcome as they were designed to measure attributive behavior in slightly different conditions. The moderate positive correlation between the adverb (ADV) and predicative (PRED) patterns can be explained by their partial overlap, as the sequence of adverb preceding a test word was also allowed in one of the search patterns for predicative behavior (see Table 1). The positive correlation between the patterns measuring gradation and modifiability by an adverb (COMP vs. ADV) is lower than expected, as the modifiability of the scalar adjectives by intensifying adverbs was a default assumption.

The negative correlation between the attributive and non-attributive patterns reflects the inclination of certain adjectives for either one or another type of usage, possibly due to semantic and/or pragmatic factors. The tendency noted in Table 4 thus proves to be a statistically significant result.

What comprises, then, the reference profile of the prototypical adjectives that could be used as a standard for deciding the level of adjectivization of yet underspecified word forms in the lexicographic work? In the present approximation[11], we decided to use the median values of each pattern as the benchmarks of adjectival behavior (See Table 3). We trusted the median values over the average, as the median is not influenced by occasional extremes in behavior and both attributive and non-attributive behavior do not fall too far from the median values. The reference profile of prototypical adjectives is an abstraction that consists of six data points. The adjectives in our dataset that fall the closest to the reference profile are *kahekordne* 'double; two-floored', *sõltumatu* 'independent', *positiivne* 'positive' and *ehtne* 'authentic'.

4.2. Comparing a word's behavioral profile to the reference profile

We used the method of Euclidean distances to assess words' cumulative similarity to a pre-set standard elsewhere (Tuulik et al. 2022) and we are going to adopt the same method here. The reference profile of prototypical adjectives (i.e., the series of median values in Table 3) is used as a standard against which the respective values of any word will be compared. We carry out the comparison among the adjectives

[11] For another approximation, see Paulsen et al. 2022.

in the reference group, as well as among the as yet underspecified participle-like word forms comprising the control group. The aim is to establish a measurable scale of similarity to the reference profile.

In general, Euclidean distance[12] is a method that measures proximities in Euclidean space between two data points that can be characterized by multiple parameters. The formula[13] is the following:

$$dist_{Euclidean} = \sqrt{\sum_{i=1}^{n=k} (V_i - O_i)^2}$$

The value of Euclidean distance is a positive number. The bigger the number, the bigger the measured difference between the phenomena and the less the similarity. The calculations thus provide us with a single value for each word that explicates its similarity to the reference profile.

We applied the formula to all words in both reference and control groups. The values of Euclidean distances range from 0.078 (the word *innustav* 'inspiring') to 0.88 (the word *ärrituv* 'irritable'), both from the control group of Ekilex words. The histogram in Figure 2 presents the general distribution of all data points independently of their group affiliation.

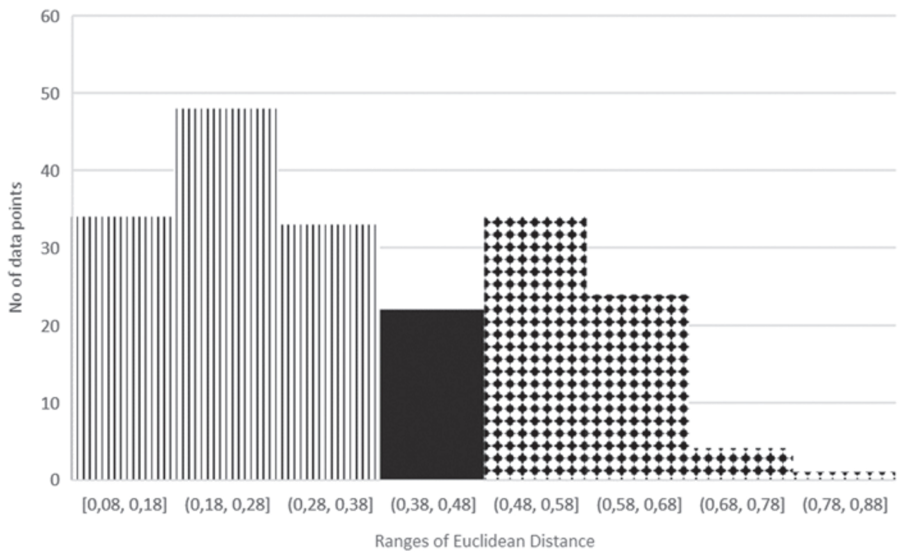

Figure 2. The distribution of the data according to a word's similarity to the reference profile

The histogram demonstrates two peaks of data accumulation. The cumulative area to the left corresponds to the low Euclidean distance values and can be interpreted as the range of maximum similarity to the reference profile. The left side covers the

[12] Within research related to language, the Euclidean distance is a popular distance metric in machine learning (Damien et al. 2011). This method is also applied in linguistic studies, for instance in sentence similarity measuring (Masanori 2021) and phonology research (Nycz, Hall-Lew 2013).

[13] V_i = the value of the i^{th} parameter of the phenomenon that is taken as a benchmark (i.e., in our case the i^{th} value of the reference profile); O_i = the value of the i^{th} parameter of the phenomenon whose proximity is measured (i.e., in our case the i^{th} value of a word's attested corpus behavior; k = the number of tested parameters (in our case k = 6).

range from 0.08 to 0.38, marked with a striped pattern; there are 104 such items in the data pool (52% of the total of 200). The other peak represents an accumulation of items moderately dissimilar to the reference profile (ranges from 0.48 to 0.68). The far right side, demonstrating maximum dissimilarity, is less populated. The range of dissimilarity covers 31,5 % of the total of 200 and is marked with a checked pattern in Figure 2. The intermediate range (0.38–0.48) is marked in solid color.

Figure 3 presents the results group-wise, revealing the inner-group variation of the Euclidean distance values and Table 6 presents the descriptive statistics. We have added the identified ranges of maximum (light grey) and minimum similarity (dark grey) to the original box plot, with three estimated degrees of similarity as a result: the maximum, intermediate and minimum similarity ranges.

Table 6. Descriptive statistics of the Euclidean distance measure (min = Minimum value, max = Maximum value, ave = Average, StDev = Standard deviation)

Statistics	Adjectives	Ekilex
min	0.082	0.078
max	0.717	0.830
ave	0.288	0.433
stDev	0.136	0.179
median	0.256	0.514

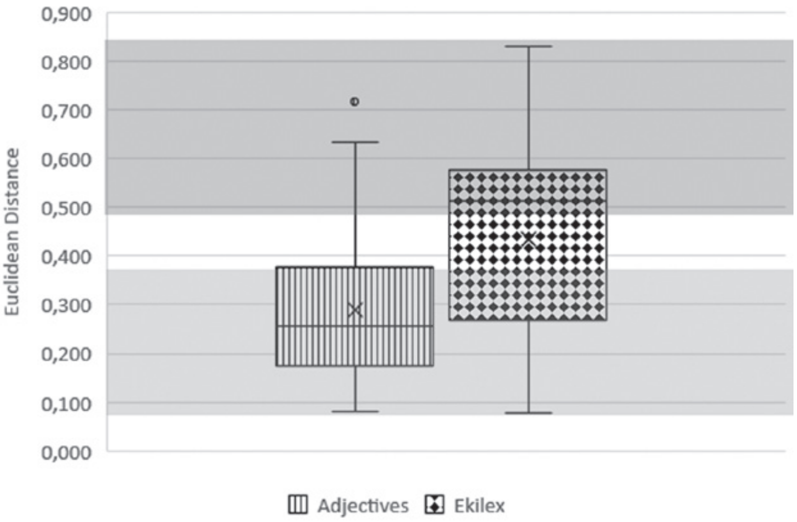

Figure 3. Variation of the Euclidean distance values of the test groups and the ranges of extreme similarity vs. dissimilarity from the reference profile

There is variation among the group of the validated adjectives in respect to similarity to the reference profile, as would be expected due to our prototype-based theoretical assumption of gradual membership and because of the observations made in the previous section about the bias of adjectives either towards attributive or non-attributive usage. The extreme outlier in Figure 3 is the word *kõrge* 'high', whose profile was most different from the reference profile, mostly due to the high relative frequency of the comparative form.

As the box plot graph in Figure 3 indicates, the distribution range of the control group (Ekilex) is much broader, which means that there is less similarity to the adjectival profile in this group, in general. This is an expected result because these words were the underspecified participle-like words of the Ekilex database, whose similarity to the adjectives needed to be estimated in the first place. Table 7 provides the proportions of the similarity ranges in each test group.

Table 7. The scale of similarity according to ranges of Euclidean distance

Test group	**Ranges of Euclidean distance (ED)**		
	The closest (ED ≤ 0.38)	**Intermediate (0.38 < ED < 0.48)**	**The most distant ED <= 0.48**
Adjectives	77%	14%	9%
Ekilex	27%	19%	54%

The range of closest distance (i.e., maximum similarity) includes, naturally, the majority of words from the reference group but also quite a number of words from the control group, which can be interpreted as their greatest level of adjectivisation (as measured by our test patterns). Some examples are *vohav* ‘proliferating’, *hõõguv* ‘glowing’, *innustav* ‘inspiring; enthusiastic’, *kergendav* ‘mitigating’, *ligimeelitav* ‘attracting’ and *uuritav* ‘explorable’). They are all present participles capable of and subject to declination when used as agreeable attributes.

The range of the most distant profiles (i.e., the maximum dissimilarity) contains some adjectives with the most deviant behavior, mostly in regard to missing the attributive usage and being used in a predicative role. These words include such personality traits as *kade* ‘envious’, *kannatlik* ‘patient’ and *ükskõikne* ‘indifferent’, and some general characteristics, e.g., *lihtne* ‘simple’, *keeruline* ‘difficult’ and *selge* ‘clear’ (see also Table 4 for the values on test patterns). The words *kõrge* ‘high’ and *täpne* ‘precise’ deviate differently by demonstrating high results in both attributive and gradation patterns.

Approximately one-half of the Ekilex test group appeared in the range of the greatest dissimilarity. A vast majority (90%) of them are past participles (e.g., *kirjeldatud* ‘described’, *kujunenud* ‘self-formed, evolved, turned into’, *küntud* ‘plowed’, *laekunud* ‘cashed in’, *laulatatud* ‘wedded’ and *lisandunud* ‘accrued’). Characteristic of the past participle forms are low results not only in the attribute agreement pattern (which is predictable, as these participles are basically non-declinable), but also in the simple attribute pattern. These forms also rarely have comparative forms, while they do have relatively higher scores in adverb and predicative patterns. Notice that the test pattern of collecting the predicatives may coincide with compound tense forms (*olema* ‘be’ + past participle), and the past participles may be preceded by manner adverbs and score relatively high in the adverb pattern. The participle forms of the phrasal verbs *etteantud* ‘given, lit. ahead+given’ *läbiviidud* ‘performed, lit. through+taken’, *luhtaläinud* ‘unsucceeded, lit. to+marsh+gone’, *väljakaevatud* ‘excavated, lit. out dug’, *ülestehtud* ‘done, lit up+done’ are exceptional in that they get very high scores in the attribute pattern.

The range of intermediate similarity contains words from both the reference group and the control group. These words in both groups tend to share the dominance of attributive usage over non-attributive usage.

The method of Euclidean distance allows us to have a compact measure instead of a cluster of distinct values (for a different solution, compare Paulsen et al. 2022). As a result, it is possible to say whether the word's general similarity fits the range of the most similar, the most dissimilar or the intermediate level of adjectivity. The same holds true for the behavior of validated adjectives, which form a natural continuum of the biggest to the smallest similarity with the reference profile.

5. Conclusion and discussion

The aim of the present study was to clarify the categorization issues around adjective candidates in Estonian lexicography, in particular from the point of view of participles obtaining adjectival properties. The idea was to create a statistical solution in order to decide the degree of adjectivization. We proposed a series of close-context indicators (called "test patterns") based on the features of adjectivity mentioned in the literature. The set of patterns was tested on a random sample of validated adjectives and on a control group of participle-like but yet underspecified lexicographic keyword candidates.

As a result, a reference profile of prototypical adjectives in Estonian was established and a methodology of comparing any word form's similarity to this profile was proposed. The Euclidean distance analysis – a complex measure considering jointly the values of all six patterns – enabled us to elaborate a tripartite scale of similarity, applicable to all words in the study, including the validated adjectives. The scalar nature of similarity to the reference profile was explicated in full accordance with our theoretical prototype-based understanding of categorization (Rosch 1973, 1975, 1978).

The Euclidean distance outcome can be used as a heuristic in lexicographic work. It is possible to calculate the individual profile of corpus behavior in terms of the similarity to the adjectival reference profile for any word without a PoS affiliation. The only precondition is that data about relative corpus frequencies of the patterns are needed. The results can be applied to develop a multi-parameter application to determine the relative adjectivity of a word or a word form, e.g., the adjectivizing participles or nominals (for the transition zones of adjectives with other lexical classes in Estonian, see Vainik et al. 2020). Since the morphosyntactic patterns characteristic of a PoS are language-specific, so is the outcome of our examination. The principles are, however, adjustable to the automatic analysis of other languages as well.

The study has its limitations. The data retrieved from the corpus directly depend on the quality of pre-existing morphological analysis. We are aware of the possibility of tagging and disambiguation errors (e.g., caused by ambiguities related to inflectional homonymy)[14] that may have an impact on the results; however, we did not correct the shortcomings of the automatic analysis manually because a potential application based on this model would apply the very same corpus processing

[14] Estonian is a morphologically rich language and the coincidence of forms of different lemmas is common, e.g., the form *joon* has two morphological interpretations: the nominative case form of the noun *joon* 'line' and the third person present indicative form of the verb *jooma* 'drink' ('I drink').

methods, and the statistics-based results of the analysis on the lexicographer's desktop would be the same.

Some interesting and challenging directions for further studies were revealed. The generally low level of comparative forms derived from adjectives with no semantic restrictions is an intriguing finding calling for examination. There is also the division of labor found among the group of validated adjectives. There appeared to be a subgroup of adjectives that did not follow the dominant use as attributes. Instead, the prevalence of the predicative role occurred. Such a division of labor is known in other languages (e.g., in English, see Bolinger 1967, Lassiter 2015: 145) but had not yet been described in Estonian data. This phenomenon needs further studies, especially for the potential of discovering a pattern for detecting the usage of words as predicative adverbials (see Example (1c) in section 2.1). Our current toolbox of patterns did not address the role of the predicative adverbial.

The strong positive correlations between measures of attributive behavior suggest that in future we could possibly give up the more specific ones, e.g., ATTR/AGR, to include the whole range of atypical non-declinable adjectives. The selection and constitution of patterns can be elaborated further. For instance, the extraction code of the predicative pattern can include certain morphological restrictions by defining the predicative case forms. The adverb pattern could include a search list of intensifying adverbs because it has been pointed out in the literature that certain kinds of modifiers (e.g., agentive, temporal and manner adverbials) are characteristic of actions rather than the result of actions (Erelt 2017b: 220). This may help in deciding whether the participle is meant as a regular verb form or as an entrenched unit interpretable as an adjective. An intriguing challenge would be to determine how to search for and distinguish participles marking habitual actions (rather than occasional events), which are supposed to be more adjectivized (see Kerge 1998: 78, Erelt 2017e: 823).

Abbreviations

ABE	abessive case
ADV	adverb pattern
ATTR	attribute pattern
ATTR/AGR	attribute agreement pattern
ATTR/ST	sentence starter pattern
COMP	gradation pattern
GEN	genitive case
INE	inessive case
NOM	nominative case
PART	partitive case
PL	plural
PRED	predicative pattern
TRA	translative case

References

Berlin, Brent; Kay, Paul 1969. Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press.

Bolinger, Dwight 1967. Adjectives in English: Attribution and predication. – Lingua, 18, 1–34. https://doi.org/10.1016/0024-3841(67)90018-6

Erelt, Mati 1977. Ebamäärasusest sõnade liigitamisel [‘About uncertainty in the classification of words’]. – Keel ja Kirjandus, 9, 525–528.

Erelt, Mati 2003. Structure of the Estonian language: Phonology, morphology, and word formation. – Mati Erelt (Ed.), Estonian Language. Tallinn: Estonian Academy Publishers, 9–92.

Erelt, Mati 2017a. Omadussõnafraas [‘The adjective phrase’]. – Mati Erelt, Helle Metslang (Toim.), Eesti keele süntaks. Eesti keele varamu, III. Tartu Ülikooli Kirjastus, 405–415.

Erelt, Mati 2017b. Öeldis [‘Predicate’]. – Mati Erelt, Helle Metslang (Toim.), Eesti keele süntaks. Eesti keele varamu, III. Tartu: Tartu Ülikooli Kirjastus, 93–239.

Erelt, Mati 2017c. Öeldistäide [‘The predicative’]. – Mati Erelt, Helle Metslang (Toim.), Eesti keele süntaks. Eesti keele varamu, III. Tartu Ülikooli Kirjastus, 278–288.

Erelt, Mati 2017d. Öeldistäitemäärus [‘The predicative adverbial’]. – Mati Erelt, Helle Metslang (Toim.). Eesti keele süntaks. Eesti keele varamu, III. Tartu Ülikooli Kirjastus, 289–299.

Erelt, Mati 2017e. Sekundaartarindiga laused [‘Sentences with secondary constructions’]. – Mati Erelt, Helle Metslang (Toim.), Eesti keele süntaks. Eesti keele varamu, III. Tartu Ülikooli Kirjastus, 756–840.

Damien, François; Wertz, Vincent; Verleysen, Michel 2011. Choosing the Metric: A Simple Model Approach. – Norbert Jankowski, Włodzisław Duch, Krzysztof Grąbczewski (Eds.), Meta-Learning in Computational Intelligence. Studies in Computational Intelligence, 358. Berlin–Heidelberg: Springer, 97–115. https://doi.org/10.1007/978-3-642-20980-2_3

Dixon, Robert M. W. 2004. Adjective classes in typological perspective. – Robert R. M. W. Dixon, Alexandra Aikhenvald (Eds.), Adjective Classes: A Cross-Linguistic Typology. Oxford: Oxford University Press, 1–49.

Geeraerts, Dirk 1989. Prospects and problems of prototype theory. – Linguistics, 27 (4), 587–612. https://doi.org/10.1515/ling.1989.27.4.587

Haspelmath, Martin 1994. Passive participles across languages. – Barbara Fox, Paul Hopper (Eds.), Voice: Form and Function. Typological Studies in Language, 27.Amsterdam: John Benjamins Publishing Company, 151–177. https://doi.org/10.1075/tsl.27.08has

Hein, Indrek; Männiko, Kaur; Kallas, Jelena; Koppel, Kristina; Langemets, Margit; Nurk, Tõnis; Plado, Merily; Vaus, Mari; Viks, Ülle; Tavast, Arvi; Laubre, Martin; Sharma, Yogesh; Niilo, Hardi 2020. Ekilex 2020. Eesti Keele Instituudi sõnastiku- ja terminibaas. Eesti Keele Instituut.

Kaalep, Heiki-Jaan; Kirt, Riin; Muischnek, Kadri 2012. A trivial method for choosing the right lemma. – Arvi Tavast, Kadri Muischnek, Mare Koit (Eds.), Human Language Technologies: The Baltic Perspective. Frontiers in Artificial Intelligence and Applications 247. IOS Press, 82–89. https://doi.org/10.3233/978-1-61499-133-5-82

Kaalep, Heiki-Jaan; Muischnek, Kadri; Müürisep, Kaili; Rääbis, Andriela; Habicht, Külli 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti keele testkorpuse morfosüntaktilise märgendamise kogemusest. [‘Do the available morphological descriptions of Estonian work on a real text?’] – Keel ja Kirjandus, 9, 623–633.

Kallas, Jelena; Tuulik, Maria 2011. Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 59–75. https://doi.org/10.5128/ERYa7.04

Kallas, Jelena; Tiits, Mai; Tuulik, Maria; Koppel, Kristina; Jürviste, Madis 2014. Eesti keele põhisõnavara sõnastik ['The Basic Estonian Dictionary']. Tallinn: Eesti Keele Sihtasutus.

Kasik, Reet 2015. Sõnamoodustus. Eesti keele varamu, IV. Tartu: Tartu Ülikooli Kirjastus.

Kerge, Krista 1998. Vormimoodustus, sõnamoodustus ja leksikon: oleviku kesksõna võrdluse all. Tallinn: TPÜ Kirjastus.

Kilgarriff, Adam; Rychlý, Pavel; Smrz, Pavel; Tugwell, David 2004. The Sketch Engine. – Geoffrey Williams, Sandra Vessier (Eds.), Proceedings of the Eleventh EURALEX International Congress. Lorient: Université de Bretagne Sud, 105–116.

Kilgarriff, Adam; Baisa, Vit; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vit 2014. The Sketch Engine: ten years on. – Lexicography, 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Koppel, Kristina; Tavast, Arvi; Langemets, Margit; Kallas, Jelena 2019. Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. – Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius (Eds.), ElectronicLexicography in the 21st century. Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 434–452.

Lassiter, Daniel 2015. Adjectival modification and gradation. – Shalom Lappin, Chris Fox (Eds.), Handbook of Contemporary Semantic Theory. Oxford: Wiley-Blackwell, 143–167. https://doi.org/10.1002/9781118882139.ch5

Laur, Sven; Orasmaa, Siim; Särg, Dage; Tammo, Paul 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20). European Language Resources Association (ELRA), 7152–7160.

Masanori, Oya 2021. Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees. – Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, 225–233.

Nycz, Jennifer; Hall-Lew, Lauren 2013. Best practices in measuring vowel merger. – The Journal of the Acoustical Society of America, 134 (5), 4198. https://doi.org/10.1121/1.4831400

Orasmaa, Siim; Petmanson, Timo; Tkachenko, Alexander; Laur, Sven; Kaalep, Heiki-Jaan 2016. EstNLTK – NLP Toolkit for Estonian. – Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: ELRA, 2460–2466. http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

Pajusalu, Renate 2017. Nimisõnafraas ['The noun phrase']. – M. Erelt, H. Metslang (Toim.). Eesti keele süntaks. Eesti keele varamu, III. Tartu Ülikooli Kirjastus, 379–404.

Paulsen, Geda; Vainik, Eene; Tuulik, Maria; Lohk, Ahti 2019. The Lexicographer's Voice: Word Classes in the Digital Era. – Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, Jose Pedro Ferreria, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius (Eds.), Proceedings of the eLex 2019 conference: Smart Lexicography. Brno: Lexical Computing CZ, s.r.o, 434–452.

Paulsen, Geda; Vainik, Ene; Lohk, Ahti; Tuulik, Maria 2021. Catching lexemes: The case of Estonian noun-based ambiforms. – Iztok Kosem, Michal Cukr, Miloš Jakubíček, Jelena Kallas, Simon Krek, Carole Tiberius (Eds.), Proceedings of the eLex 2021 conference: Post-Editing Lexicography. Brno: Lexical Computing CZ, s.r.o, 288–311.

Paulsen, Geda; Tuulik, Maria; Lohk, Ahti; Vainik, Ene 2022. From verbal to adjectival: Evaluating the lexicalization of participles in corpus. – Slovenščina 2.0, 10 (1), 65–97. https://doi.org/10.4312/slo2.0.2022.1.65-97

Rosch, Eleanor 1973. On the internal structure of perceptual and semantic categories. – T. E. Moore (Ed.), Cognitive Development and the Acquisition of Language. New York–San Francisco–London: Academic Press, 111–144.

Rosch, Eleanor 1975. Cognitive representations of semantic categories. – Journal of Experimental Psychology: General, 104 (3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, Eleanor 1978. Principles of categorization. – Eleanor Rosch, Barbara B. Lloyd (Eds.), Cognition and Categorization. Hillsdale–New York: Lawrence Erlbaum, 27–48.

Tavast Arvi; Koppel, Kristina; Langemets, Margit; Kallas, Jelena 2020. Towards the superdictionary: Layers, tools and unidirectional meaning relations. – Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras (Eds.), Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. 1. Greece: Democritus University of Thrace, 215–223.

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina 2018. Unified data modelling for presenting lexical data: The case of EKILEX. – J. Čibej, V. Gorjanc, I. Kosem, S. Krek (Eds.), Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts. Ljubljana, Slovenia, 749–761.

Tuulik, Maria; Vainik, Ene; Paulsen, Geda; Lohk, Ahti 2022. Kuidas ära tunda adjektiivi? Korpuskäitumise mustrite analüüs [‘How to recognize adjectives? An analysis of corpus patterns’]. – Estonian Papers in Applied Linguistics, 18, 279–302. https://doi.org/10.5128/ERYa18.16

Vainik, Ene; Paulsen, Geda; Lohk, Ahti 2020. A typology of lexical ambiforms in Estonian. – Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras (Eds.), Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. 1. Greece: Democritus University of Thrace, 119–130.

Vainik, Ene; Lohk, Ahti; Paulsen, Geda 2021. The distribution index calculator for Estonian. Electronic lexicography in the 21st century. – Iztok Kosem, Michal Cukr, Miloš Jakubíček, Jelena Kallas, Simon Krek, Carole Tiberius (Eds.), Proceedings of the eLex 2021 conference: Post-Editing Lexicography. Brno: Lexical Computing CZ, s.r.o, 121–138.

Viitso, Tiit-Rein 2003. Structure of the Estonian language: Phonology, morphology, and word formation. –Mati Erelt (Ed.), Estonian Language. Tallinn: Estonian Academy Publishers, 9–92.

Viks, Ülle 1977. Sõnaliik kui niisugune [‘Part of speech as such’]. – Keel ja Kirjandus, 9, 521–525.

Resources

CombiDic = The EKI Combined Dictionary 2022. Langemets, Margit; Hein, Indrek; Järviste, Madis; Kallas, Jelena; Kiisla, Olga; Koppel, Kristina; Leemets, Tiina; Mäearu, Sirje; Paet, Tiina; Päll, Peeter; Raadik, Maire; Risberg, Lydia; Tammik, Hanna; Tavast, Arvi; Tiits, Mai; Tsepelina, Katrin; Tuulik, Maria; Valdre, Tiia; Viks, Ülle; Sai, Edgar; Tubin, Valentina (Eds.). Institute of the Estonian Language. https://doi.org/10.15155/3-00-0000-0000-0000-08C0AL

Ekilex = The Estonian Dictionary and Termbase System Ekilex 2021. Institute of the Estonian Language. Arvi Tavast, Indrek Hein, Kaur Männiko (Developers). Martin Laubre, Yogesh Sharma, Hardi Niilo, Henri Kokk (Developers, LLC TripleDev). https://ekilex.eki.ee

ENC 2019 = Kallas, Jelena; Koppel, Kristina 2020. Estonian National Corpus 2019. Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-08565L

Sketch Engine. https://www.sketchengine.eu

EESTI KEELE PROTOTÜÜPSE ADJEKTIIVI MORFOSÜNTAKTILISE KORPUSPROFIILI JÄLIL

Ene Vainik[1], Geda Paulsen[1,2], Ahti Lohk[3], Maria Tuulik[1]

Eesti Keele Instituut[1], Uppsala Ülikool[2], Tallinna Tehnikaülikool[3]

Sõnavara kategoriseerimisel sõnaliikidesse valmistavad leksikograafias probleeme ennekõike üleminuekualad. Üks peamisi murekohti on raskus määratleda seejuures verbi ja adjektiivi vahelist piiri (Paulsen jt 2019, Paulsen jt 2020). Siinses uurimuses vaatleme partitsiipide adjektiviseerumisprotsesse korpusstatistika andmetele tuginedes. Lähenemine põhineb teoreetilisel eeldusel, et mistahes nähtusi kategoriseerivad inimesed alateadlikult liikmete sarnasust n-ö prototüüpsele esindajale ehk kategooria keskmele hinnates. See toob kaasa, et kategooria liikmed võivad olla selle prototüübiga kas rohkem või vähem sarnased; kategooria perifeerses osas võivad liikmed kuuluda juba ka mingisse naaberkategooriasse.

Adjektiividele omaste joonte väljaselgitamiseks korpuses kasutame testmustrite sarja, millest igaüks haarab potentsiaalse adjektiivi lähikonteksti. Kuus testmustrit põhinevad adjektiivide omadustel, mis on kirjanduses esile toodud ning ka eelmärgendatud korpusetekstides tuvastatavad. Kolm mustrit mõõdavad testsõna esinemist atribuudi rollis – eestäiendina üldiselt ning kahes kitsendatud positsioonis: ühilduvana põhisõnaga käändes ja arvus ning teiseks paiknevana lause alguses. Veel kätkesid mustrid esinemist keskvõrde vormis, laiendatavust vahetult eelneva adverbiga ning esinemist öeldistäitena st *olema* verbi jätkuna.

Prototüüpse adjektiivi korpuskäitumise profiil selgitati välja sajast sõnast koosneva juhuvalimi põhjal „Eesti keele põhisõnavara sõnastiku" adjektiividest. Kontrollrühm (N = 100) moodustati Eesti Keele Instituudi sõnastikubaasis Ekilex (Hein jt 2020) leiduvast sõnaliigimärgendita partitsiibist samuti juhuvalimina, silmas pidades erinevate partitsiibivormide võrdset esindatust.

Adjektiivi morfosüntaktilise käitumise prototüüpi valiti esindama katsetes kasutatud testmustrite suhteliste sageduste mediaanväärtused adjektiivide rühmas. Sarnasusmõõdikuna kasutasime eukleidilise kauguse meetodit, mis lubab analüüsida kõrvutatavate nähtuste mitmeid parameetreid korraga. Analüüsi tulemuseks on skaala, mis eristab määra, kuivõrd uuritav sõna sarnaneb oma korpuskäitumiselt tavalisele tüüpilisele adjektiivile. Analüüsi tulemusi lahkame testmustrite sarja tõhususe, aga ka testitud adjektiivide korpuskäitumise iseärasuste vaatenurgast.

Võtmesõnad: leksikograafia, korpuslingvistika, adjektiivid, leksikaalne dekategoriseerumine, eesti keel

Ene Vainik ((Institute of the Estonian Language) is a cognitive linguist. Her field of studies covers semantics, folk-psychology and, in particular, conceptualisations of emotions. Another related field has been affective computing (e.g., detecting emotions in written Estonian text and detecting emotional valence). Currently, she works for a project on part of speeches from the perspective of lexicography.
Roosikrantsi 6, 10119 Tallinn, Estonia
ene.vainik@eki.ee

Geda Paulsen's (Institute of the Estonian Language, University of Uppsala) research interests include lexical semantics, morphosyntax, word formation, lexicography, corpus linguistics, and contrastive linguistics.
Roosikrantsi 6, 10119 Tallinn, Estonia
Engelska parken, Thunbergsvägen 3 L, Box 636, 751 26 Uppsala
geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se

Ahti Lohk's (Tallinn University of Technology, Institute of the Estonian Language) research areas are validation of wordnet semantic hierarchies with graph-based methods and text mining algorithms for extraction of useful, new, and applicable information from unstructured texts.
Akadeemia tee 15A, 12618 Tallinn, Estonia
ahti.lohk@taltech.ee

Maria Tuulik's (Institute of the Estonian Language) research interests are lexical semantics of adjectives, lexicography, corpus linguistics, and boundaries of word classes.
Roosikrantsi 6, 10119 Tallinn, Estonia
Maria.Tuulik@eki.ee