From verbal to adjectival: evaluating the lexicalization of participles in an Estonian corpus

Geda PAULSEN
Institute of the Estonian Language; University of Uppsala

Maria TUULIK
Institute of the Estonian Language

Ahti LOHK
Institute of the Estonian Language; Tallinn University of Technology

Ene VAINIK
Institute of the Estonian Language

This study addresses categorization issues related to adjective candidates in Estonian, focusing on the category of participles. The aim of the analysis was to assess the ranges of the prototypical adjective and to determine its degree of deviation on the prototypicality scale. The investigation was based on a group of validated adjectives – selected adjectives included in the *Basic Estonian Dictionary* – and two control groups of more and less lexicalized participles. We tested seven morphosyntactic corpus patterns characteristic of adjectives. The test patterns were based on the prototypical features of the adjective, as well as

Paulsen, G., Tuulik, M., Lohk, A., Vainik, E.: From verbal to adjectival: evaluating the lexicalization of participles in an Estonian corpus. Slovenščina 2.0, 10(1): 65–97.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.1.65-97

on observations made in the actual lexicographic analysis. To assess the sample words and determine the significance of the test patterns from the point of view of defining adjectivity, we used deviation analysis. The results of this study can be applied to establish a measure of adjectivity for lexicographic judgments when distinguishing, for instance, lexicalized participles from regular ones.

Keywords: corpus linguistics, lexicography, Estonian language, adjective, participle, deviation analysis

1 Introduction

A morphosyntactic analysis, from the corpus linguistics perspective, is a basic operation using inflectional paradigms and a base lexicon in part-of-speech disambiguation of the tokens in a text. For lexicographers, who first determine a word's lexical categorization, the morphosyntactic behavior of a lexeme in its natural contexts is essential information when judging its lexical classification. The data available in corpora yield potential new headwords, but automatic pre-processing is needed in order to make properly weighed decisions about the lexical affiliation of potential new lexemes, as the amount of material may be overwhelming.

In this study, we applied language technology and statistical analysis in order to aid lexicographers in structuring potential headwords. Our target lexical category is adjectives – we attempt to establish the ranges of similarity to the prototypical adjective based on a representative sample of predefined adjectives and to identify degrees for classifying a word (form) as an adjective. The background assumption driving the analysis is that an adjective is not a clearcut but rather a prototype-based category. To ascertain the adjectival core, we developed an evaluation methodology to assess the similarity of a word to an adjective, based on morphosyntactic corpus behavior. In other words, we seek to determine the tolerance ranges of the parameter values that distinguish adjectives from other words and can be used for comparison with the corresponding values of unclear cases.

To test the characteristic attributes of the Estonian adjective, we use a set of corpus patterns based on parameters that include morphological and syntactic features highlighted in the linguistic literature and

detectable in the corpus. The current study is also a re-evaluation of the test patterns used in our previous work (Tuulik et al., 2022). To capture a wider scope of adjectival corpus behavior, we introduce a new pattern: the predicative pattern.

We start our investigation with the rationale of the study and an overview of the Estonian adjective, as well as the participial categories, as described in Section 2. Here we identify the most relevant morphosyntactic properties of adjectives and participles as described in the literature. We proceed with the formation of a random sample of 100 adjectives from the headwords of the Estonian Basic Dictionary (Kallas et al., 2014; see also Kallas and Tuulik, 2011, Kallas et al., 2014), constituting the test group of "adjectives" as validated by lexicographers. To compare the prototypical adjectives with a close, but less clearly adjectival category, we contrast adjectives with participles displaying different degrees of lexicalization. To that end, we created two control groups of equal size: 1) participial independent headword candidates from the lexicographic database of Ekilex, and 2) a sample of regular verbal participles formed of common verbs. The hypothesis behind the composition of the two participial groups is that the participles group of the Ekilex entries are more lexicalized and resemble the reference group of validated adjectives to a larger extent than the regular participles.

To capture the adjectival corpus behavior, we elaborated test patterns detectable in the Estonian National Corpus 2019 and extracted frequency data on the sample words. The test samples, test patterns, methods applied to the data extraction and statistical processing are described in Section 3. Section 4 is devoted to the analysis of the extracted data. The absolute frequency data will be relativized, and the measurement of the adjectival corpus behavior and its limits are described in Section 4.1. After establishing the tolerance ranges of the validated adjectives, the respective values of the control groups will be related to these limits and the degrees of deviation will be calculated in Section 4.2. In Section 4.3, we evaluate the differentiation efficiency of each pattern. A concluding discussion of the results is given in Section 5.

2 Background

2.1 Adjectives and lexical decategorization in Estonian lexicography

This study was motivated by a challenge in the Estonian lexicography: the need to add PoS labels to a vast number of still under-specified keywords of the *Combined Dictionary of the Institute of the Estonian Language* (CombiDic). The current direction in Estonian lexicography is a unification of lexical resources (dictionaries and term bases) into a central superdictionary, the online public dictionary CombiDic. This process is supported by the dictionary writing system Ekilex. At the same time, lexicographic work is moving constantly towards a higher degree of automation and processing of corpora (Tavast et al., 2018; Koppel et al., 2019; Tavast et al., 2020).

A result of the automated processing of lexicographic data is that the lexical database of Ekilex includes automatically generated lists of dictionary entry candidates, requiring assessment of their degree of grammaticalization and/or lexicalization. The data are integrated from different sources,[1] containing words or word forms with different lexicographic statuses:

a) those not included in the CombiDic (*the CombiDic candidates*),
b) those included as headwords but without information about their lexical category (*under-specified headwords*),
c) those included in the CombiDic as headwords with PoS label(s) (*PoS-tagged headwords*).

Providing the underspecified Ekilex entries with PoS tags and assessing the CombiDic candidates for their potential status as lexical entries is an urgent lexicographic issue. Today, 72% (N = 255 691) of the total number of the public CombiDic keywords are missing PoS tags.[2] A survey of Estonian lexicographers (Paulsen, Vainik, Tuulik, Lohk 2019;

1 For instance, the participle forms included in one of the control groups (the Ekilex participles) of this study derive mainly from the databases of the *Estonian Collocations Dictionary* (2019) and the *Estonian-Russian Dictionary* (2018).

2 This value stems from an excerpt from all Ekilex databases (dictionaries, term bases and phrase collections) done by Kaur Männiko on 24. 1.2022.

Paulsen, Vainik, Tuulik 2020) revealed a need for automatic corpus-based solutions to determine the word class affiliation of a lexeme when there is more than one possible interpretation. Adjectives were pointed out as one of the most complicated categories, in particular the specification of participle forms as either verbal or adjectival (Paulsen et al., p. 188–189).

In a previous study (Tuulik et al., 2022), we tested six morphosyntactic corpus patterns that could differentiate adjectives from other words in 12 groups of words. Six groups of the selected words represented "neighboring" categories of adjectives (prototypical adjectives, less prototypical adjectives, adjectival participles, substantival adjectives, adverbial adjectives, and non-declinable adjectives; with regard to those categories closely related to adjectives, see Vainik et al., 2020, p. 122–123). In addition, we used six control groups representing clear cases of other word classes (verbal participles, substantives, adverbs, verbs, proadjectives and ordinals). All test groups contained 10 words each.

The tested parameters were, to different degrees, able to differentiate adjectival morphosyntactic behavior (Tuulik et al., 2022, p. 295–298). The next step was to measure the scalability of the adjectival behavior on a more representative set of adjectives and establish the tolerance ranges of prototypical adjectives. Since the sample of prototypical adjectives in the previous study was rather small (N = 10), the parameters had to be tested on a larger sample of adjectives that represent the best examples of their category. In this study, we focused on the overlapping area between adjectives and verbs. We chose participles as the contrasting test group to the prototypical adjectives for three reasons:

1) as a morphosyntactically close lexical group to adjectives, it is theoretically significant to examine where exactly participles differ from adjectives,
2) participles constitute one of the most problematic categorization issues for lexicographers,
3) participles are substantially represented among those words without clear lexicographic status in the Ekilex database (N = 1,542 in January 2022).

Theoretically, we rely on the prototype-based approach to linguistic categories. The latter was initially employed in the study of the internal structure of categories in experimental psychology by Eleanor Rosch (1973; 1975; 1978), and was also found useful in lexical semantics (see e.g., Berlin and Kay, 1969; Geeraerts, 1989). Hence, we assumed that the boundaries of prototype-based categories were not definite, and the members of a category might have different statuses: there might be more typical, "better" examples of a category. By a prototypical adjective we mean a lexeme displaying (to a certain extent) the morphological, syntactic, and semantic properties ascribed to this lexical category in the linguistic literature.

How can one then tell adjectives and participles apart? A prototype can be instantiated by the "best example" or described via a bundle of features, none of which is necessary nor sufficient to define the whole category. The present study is a test of the adjectival core features and the possibility of distinguishing more and less adjectival corpus behavior. As lexicographers need support in qualitative decision-making, we aimed to enhance the procedures used when setting boundaries. In our analysis, we combined the means of both prototypical and classic categorization, as the (gradual) deviation continuum we developed entailed basically binary and privative decisions.

The linguistic properties describing adjectives as a class will be discussed in the next section based on the example of Estonian adjectives. Since the adjective profile will be contrasted with the corresponding patterns of participles, we will also give an overview of Estonian participles.

2.2 The properties of adjectives and participles in Estonian

2.2.1 The Estonian adjective

There are no universal criteria for defining adjectives as a word class: adjectives may exhibit properties resembling nouns or verbs or neither of these two major categories. However, a distinguishable adjective class exists in every human language (Dixon, 2004, p. 1). Adjectives do not take major syntactic positions in sentences but occur in an attribu-

tive or predicative relation to the subject or object, modifying the noun. Semantically, adjectives describe nouns and portray their character.

In Estonian, a prototypical adjective is definable by a three-level bundle of features: morphological, syntactic, and semantic. The morphological processes characteristic of adjectives[3] involve inflection, forms of comparison, and derivation.[4] Like other word classes classified as nominals (adjectives, nouns, numerals and pronouns), Estonian adjectives are inflected for case[5] and number. The adjectival category of comparison involves the comparative suffix *-m* and the superlative suffix *-im*. There are no morphophonological restrictions on forming comparative forms. It is also possible to use the analytic superlative construction *kõige* "most" + comparative form, and some adjectives are only used in this construction (Viitso, 2001, p. 32–35, 42).

The adjective may constitute an adjective phrase by itself, or it may occur together with its modifiers as an attribute (1a), predicative (1b) or predicative adverbial (1c). The adjective is most clearly recognizable when used attributively, which is seen as the primary function of the adjective (Erelt 2017, p. 406). An Estonian adjective used as an attribute is typically prenominal and agrees with its head noun in case and number (as in (1a)), except for the terminative, essive, abessive, and comitative cases, which require the genitive of the adjective attribute (Pajusalu 2017, p. 382; Viitso 2001, p. 35), for instance *rõõmsate lasteta* [glad-GEN child-ABE] "without glad children". The adjectival predicative modifies the subject most often by using the copula verb *olema* "be". It usually appears in the nominative case (as in (1b)), but also other grammatical cases and elative occur. The predicative adverbial typically expresses a result state and occurs in the translative, es-

3 It should be mentioned that to simplify the morphological corpus analysis of Estonian, certain forms are treated differently in automatic morphoanalysis than in the traditional grammars: the comparative and superlative forms are analyzed as separate lemmas, and the present participles as adjectives (see e.g. Habicht et al., 2000).

4 Another characteristic of adjectives is the adjectival derivative suffixes (the most frequent are *-ne*, *-line*, *-lik*, *-kas*, *-jas*, *-tu*, *-us*; see Kasik, 2015, p. 348–367, and about adjectival derivation see Vare, 1984).

5 Estonian has 14 nominal cases: three grammatical (nominative, genitive and partitive) and 11 semantic or adverbial cases: illative (ILL), inessive (INE), elative (ELA), allative (ALL), adessive (ADE), ablative (ABL), translative (TRA), terminative (TERM), essive (ESS), abessive (ABE), and comitative (COM) (e.g. Viitso, 2003, p. 32).

sive or nominative cases in connection with a range of verbs of change[6] (Erelt, 2017b, p. 281–287; Erelt, 2017c, p. 289; Erelt, 2017a, p. 405). An adjective can also take a modifying adverb (1d).

(1a)	*Mahl*	*tehti*	*hea-de-st*	*õun-te-st*
	juice	was made	good-PL-ELA	apple-PL-ELA
	"The juice was made of good apples."			
(1b)	*Õun*	*on*	*väga*	*hea.*
	apple	is	very	good-NOM
	"The apple is very good."			
(1c)	*Jutt*	*muutus*	*igava-ks.*	
	chat	became	boring-TRA	
	"The chat became boring."			
(1d)	*üsna mugav*			
	"pretty cozy"			
(1e)	*palju mugava-m*;		*mugava-im*	
	much cozy-CMPR		cozy-SUP	
	"much cozier"		"coziest"	

The semantic properties of an adjective affect its ability to take comparative and superlative forms: the adjective allows for comparison if it encodes a scalar (degree) property,[7] e.g. the adjective *mugav* "cozy" in (1e). Comparison forms are thus not used with all adjectives, even when there are no morphophonological constraints (Viht and Habicht, 2019, p. 27). The distinction between a relative (scalar) and absolute (non-scalar) property also influences the structure of the adjective phrase: scalar adjectives can be modified by adverbs of intensity (*väga soe* "very warm" cf. ?*väga lingvistiline* "very linguistic") (see Erelt, 2017a, 406–408). However, the distinction is not absolute, as the ability of non-scalar adjectives to be modified by adverbs is not impossible in particular contexts, as example (2) shows:

6 These verbs include: *kujunema* "turn", *muutuma* "change", *minema* "go", *saama* "get", etc.

7 This dichotomy corresponds to the distinction between classifying adjectives and qualifying (attributive) ones: the former categorize the entity denoted by the noun as belonging to a certain type or class, while the latter describe the entity (e.g. Warren, 1984).

(2) *Lehed on südajad, alumised on peaaegu kolmnurksed*[8]
"The leaves are heart-shaped; the low ones are almost triangular."

The combination of semantic, morphological, and syntactic properties that define the Estonian adjective may lack some of the general features characterizing adjectives. There are words labeled as adjectives in Estonian that do not fulfill the agreement condition of the "best example of an adjective", because they are non-declinable (e.g., *kulla lapse-d* [dear child-PL] "dear children"). Moreover, other word classes may behave as adjectives in certain aspects. For instance, even though comparison is basically an adjectival property, nouns may adapt comparison forms in suitable contexts (*elu on lill* "life is (like) a flower": *elu on lillem* "life is more (like) a flower"). Nevertheless, the most distinctive example of a category carrying several semantic and morphosyntactic properties of adjectives is the basically verbal class of participles.

2.2.2 The Estonian participle

Participles are non-finite verbal forms situated on the border between verbs and adjectives. This implies that the participle suffixes function partly as grammatical, partly as lexical categories (e.g. Viht and Habicht 2019, p. 37), positioned between regular verbal endings and derivative suffixes that yield new lexemes. Estonian participles are related to verbs via inflection for voice and mood. Both present and past participles also show adjectival properties by functioning as attributes or predicatives in a sentence. Common to all participles is that it is possible to regularly form comparative and superlative forms of them (Kerge, 1998; Erelt, 2003, p. 63; Kasik, 2015, p. 369). An important distinction between present and past participles can be made by the verbal and nominal poles: while the non-declinable past participles[9] occur together with finite verb forms of the verb *olema* "be" (in compound

8 This example is taken from the corpus ENC 2019, subcorpus Web 2013.

9 On rare occasions, the past participle can inflect when used as a postposed attribute, agreeing with its head in case and number. Since this use is rather exceptional, we do not expect it to significantly influence the results. An example of a postposed participle is shown below:
inimese-l, tõrjutu-l ja allasurutu-l, on raske
person-ADE ostracize-PTCP-ADE and stifle-PTCP-ADE is difficult
"the person, ostracized and stifled, has difficulties"

tenses and negation), the present participles in Estonian show properties of nominal categories as they can be inflected for case and number. Modifiers characteristic of activity rather than the result of activity (or the possibility of those modifiers, in particular agentive, temporal and manner adverbials) may incline the interpretation towards the verbal (Erelt, 2017, p. 220). The participial endings, according to the tense and voice categories, for the verb *sööma* "eat" are presented in Table 1:

Table 1: *The Estonian participles*

	Personal	impersonal
present participles	***v***	***dav/tav***
	õun-a *söö-v* *laps* apple-PART eat-PTCP child "a child eating an apple"	*hommiku-l* *söö-dav* *õun* morning-ADE eat-PTCP apple "the apple being eaten in the morning"
past participles	***nud***	***dud/tud***
	laps *on* *söö-nud* *õun-a* child be-3SG eat-PTCP apple-PART "the child has eaten the apple"	*hommiku-l* *söö-dud* *õun* morning-ADE eat-PTCP apple "the apple that has been eaten in the morning"

The present impersonal participle form *söödav* is a good example of the decategorization patterns of participles: this form has the status of a headword in the CombiDic[10] as an adjective meaning "edible, satisfying, palatable", and even as a noun meaning "edibles". The adjectival reading enables this participle to obtain the interpretation of a predicative, not compound tense form. Semantically, the lexeme *söödav* shows the abstraction tendency of adjectivized participles when it comes to the concept of time: a characteristic of adjectivization is that the situation or property can be generalized to "at any time, always" (Kerge, 1998, p. 78; Erelt, 2017d, p. 823). The detachment from the verbal paradigm is complete when the participle receives an independent meaning with respect to its verb base (Kasik, 2015, p. 70).

The questions a lexicographer deals with when categorizing participle forms are: How can we distinguish verbal and nominal participles according to their morphosyntactic behavior? When can we say that a

10 https://sonaveeb.ee/search/unif/dlall/dsall/s%C3%B6%C3%B6dav/1

participle has distinctively become an adjective? We propose that, in practice, it is a matter of scaling the relative proportion of occurrences in a text corpus in respect to one or another pole. We do not expect the differentiation to be straightforward, but rather a question of tendencies.

3 Material and methods

The analysis of adjectives and participles was based on the morphosyntactic patterns identifiable in an annotated corpus. In the compilation of the test patterns, we aimed to capture the most salient attributive adjectival sequences, but also the most central non-attributive constructions. These patterns are presented in Section 3.1.

The patterns were tested on validated adjectives: the relative frequencies of corpus patterns of this group represented the reference point for further analysis and could be compared with the respective values of two control groups of participles. The principles behind the selection of the three test groups are discussed in Section 3.2. Section 3.3 presents details of the data extraction procedures. The method we used for assessing the distance of a participle from the prototypical adjectival behavior was deviation analysis, as explained in Section 3.4.

3.1 Catching adjectivity. The test patterns

The extraction of corpus sequences capturing the morphosyntactic behavior of the Estonian adjective is based on seven fixed patterns. The patterns are based on properties typical of the adjective, definable by two main parameters: the attributive and non-attributive adjectival functions. The test sequences must also be extractable by the corpus tagging system.

Most of the patterns reflect the properties assigned to adjectives in the linguistic literature. The third pattern is inspired by practical lexicographic work, and the fourth pattern has grown out of the analysis of corpus material. The seventh, the predicative pattern, is an addition to our previous investigation of adjectives (Tuulik et al., 2022, pp. 283–285). The term *test word* refers to any test word inserted into the search for the respective pattern. Six of the patterns are sequences; the comparative pattern counts, and thereby confirms, the existence

of the comparative form of a test word in the corpus. The test patterns used in this study are as follows:

1) The attribute pattern (ATTR) targets the sequence of the test word immediately preceding a noun. This pattern is based on the tendency of an adjective to modify the noun as an attribute. The collocational sequence ADJ_NOUN presumably reflects the most frequent use of adjectives (e.g. *väike laps* "little kid").
2) The agreement pattern (ATTR/AGR) is a sequence of the test word in the same case and number as the following noun. It tests the agreement of the test word and head noun, based on the ability of adjectives in the attributive function to agree in case and number with their head nouns (*väikes-te-l kivi-de-l* [*little*-PL-ADE *kid*-PL-ADE] "on the small rocks").
3) The sentence starter pattern (ATTR/ST) sets a syntactic restriction on the attribute phrase: the test word followed by a noun must be located at the beginning of a sentence. The purpose of this pattern is to differentiate verbal participles from adjectivized ones, for instance *Tuleval suvel…* "In the upcoming summer" is quite natural, but *Oleval suvel…* "In the being summer" is not.
4) The four-spot pattern (ATTR/4) measures the occurrence of the test word in a larger pattern, where it modifies a substantive and follows the sequence of an unspecified verb and an unspecified word (verb + X + test word + noun). According to our pilot study compiled to test the parameters (Tuulik et al., 2022), this pattern distinguished the main target of the present investigation – participles – from adjectives. With respect to other categories, this parameter was not as effective. This study will thus indicate whether this pattern should be kept in the test battery.
5) The adverb pattern (ADV) ascertains the sequence of an adverb preceding the test word. We expect the ability to take adverbial modifiers to be characteristic of the adjectives in the corpus, particularly with scalar adjectives.
6) The comparison pattern (COMP) estimates whether a word yields comparative forms. We restrict this test pattern to comparatives, assuming that the existence of a comparative form is a logical precondition for the possibility of a superlative. Moreover, since the

highest degree of comparison can, in parallel, be expressed by the analytic *most*-construction, involving the adverb *kõige* "most" and the comparative form of the adjective (*kõige väiksem* "most smaller"), the results are not quite representative.[11]

7) The predicative pattern (PRED) targets the sequence of the test word directly after the copula verb *olema* "be" or after *olema* and an adverb. Both patterns are characteristic of the Estonian predicative;[12] of course, these are also potential forms for the compound tense constructions involving participles, but in this case there may be additional sentential elements between the copula and participle. We do not specify the morphological form of the test word here as the Estonian predicative can be marked by several cases (the three grammatical cases and the elative case – see Section 2.2).

As the patterns described above indicate, there are two recurrent structural relations unifying the variables: patterns 1–4 include attributive phrases in a more or less fixed position in the sentence, and patterns 4 and 7 involve the pre-adverbial relationship of the test word (adverb + test word). Of the four attributive patterns, 1 and 2 can be classified as **general attributive patterns**, and 3 and 4 as **complex attributive patterns** with fixed positions in the sentence.

A summary of the division of test patterns according to attributive and non-attributive parameters is given in Table 2; the abbreviation TW stands for test word and the patterns are given in their logical form. Note that the elements in patterns containing TW are consecutive sequences.

11 There is an analytic way in Estonian to express even the comparative relation, using the (also comparative form) word *rohkem* (*rohkem väike* "more little"), but this is not as productive a pattern as the *kõige*-superlative and, although understandable, it is often not quite idiomatic.

12 Since the predicative can occur in four cases (nominative, partitive, genitive and elative), we do not restrict the morphological form of the test word. The predicate of this pattern is limited with *olema* "be" because we need to narrow down the corpus search pattern and focus only on the central case of predicative in Estonian, leaving out the cases of the predicative adverbial that enables the predicative to take the translative, essive and nominative cases in connection with verbs expressing change of state: *saama* "become", *muutuma* "turn", *kujunema* "develop into, turn", *minema* "go", etc.

Table 2: *Test patterns and parameters*

parameter	pattern name		abbreviation	search pattern
attributive	general attributive	attribute pattern	ATTR	(TW ∧ noun)
		agreement pattern	ATTR/AGR	$(TW \wedge noun)_{agreement}$
	complex attributive	sentence starter pattern	ATTR/ST	$(TW \wedge noun)_{sentence\ start}$
		four-spot pattern	ATTR/4	verb ∧ x ∧ (TW ∧ noun)
non-attributive		adverb pattern	ADV	(adverb ∧ TW)
		comparison pattern	COMP	∃ comparative
		predicative pattern	PRED	∀ olema "be" ∧ TW ∀ olema "be" ∧ (adverb ∧ TW)

3.2 The sample: validated adjectives and control groups of participles

The data set used in this study contained three sample groups. The reference group of our study, which we also call *the validated adjectives group*, consisted of a selection of lexicographically verified adjectives: 100 words extracted by random sampling from the 554 adjectives included in the *Basic Estonian Dictionary*.[13] We expected the adjectives included in this dictionary to be the most central and prototypical. To ensure the coherence of the sample, we excluded the lexemes that showed ambivalent behavior regarding their word class affiliation (e.g. *vabatahtlik*, interpretable both as the adjective "voluntary" and the substantive "volunteer"). We also excluded adjectives missing some central adjectival features, such as the non-declinable *eri* "separate, various".

We used two groups of participles as control groups of less prototypical cases to compare with the reference group. Control group 1 contained participles that by expectation incline towards adjectives, and control group 2 consisted of participles used predominantly in verbal contexts. All the participles were selected by random sampling and checked for their suitability. Both samples included all four participle

13 The dictionary includes the 5,000 most frequent Estonian words explained in simple language. In addition to frequency criterion, the words were selected according to their prominence for the learner of Estonian (Kallas et al., 2014: 1109–1110).

types (cf. Table 1 in Section 2.2) – personal present, personal past, impersonal present, and impersonal past participles – with an equal number of each participle type.

Control group 1, *the adjectivizing participles in Ekilex*, consisted of 100 participles that were CombiDic candidates or under-specified headwords of CombiDic. We expected most of these forms to behave as adjectives in the corpus texts and potentially to be tagged accordingly in the database. The random sample of Ekilex participles was extracted from the Ekilex database (N = 1,543).[14]

Control group 2, *the regular participles*, contained 100 participles for which we expected as little adjective behavior as possible. The verbs functioning as bases for the participles in this group were selected by random sampling from the approximately 1,000 verbs included the *Basic Estonian Dictionary*. The four types of participles were then formed and manually checked for their verbal use and sufficient frequency in the corpus.

The composition of the test groups was planned with the expectation that the morphosyntactic test patterns described in Section 3.1 would be able to distinguish the groups from each other. In other words, we hypothesize that the adjectivizing participles group of the Ekilex entries resembles the reference group of validated adjectives to a larger extent than the regular participles in the morphosyntactic corpus patterns we focus on.

3.3 The corpus extraction process

We extracted the data from the Estonian National Corpus 2019[15] (ENC 2019; see also Koppel and Kallas, 2022). The 1.5 billion token corpus ENC 2019 was pre-tagged, lemmatized, and disambiguated with the EstNLTKv.1.6 program, a natural language toolkit explicitly developed

14 The PoS-tagging status of CombiDic headwords changes along with the updating of the dictionary. This was extracted 24. 1. 2022.

15 The ENC-corpora are stored in the corpus query system Sketch Engine (Kilgarriff et al., 2004; Kilgarriff et al., 2014). We use the files of the ENC2019 uploaded from Sketch Engine to the home page of the Center of Estonian Language Resources. The frequency results of the Sketch Engine and CELR files may differ by up to 1%, as the last uses a slightly different approach by rejecting the data from broken sentences (Neeme Kahusk, personal communication). The ENC2019 subcorpora are available at https://entu.keeleressursid.ee/shared/7769/N66ZdfvwzQuXWIvIjnhVuX74oWmi1zrruZ1VpN8QE1Hj6jbfq5oMBxm8YQDrugyM

for the Estonian language and written in the Python programming language, executing basic NLP tasks (Orasmaa et al., 2016, p. 2460, Laur et al., 2020). In the tagging process, EstNLTK uses the tag set of the Vabamorf morphoanalyzer, which combines rule-based and probabilistic models, and its lemma disambiguation system based on the Vabamorf lexicon. According to Kaalep et al. (2012), EstNLTK's lemma disambiguation precision is around 0.94.

We applied a code[16] written in Python programming language for automatic data extraction. Table 3 presents the logical expressions used in data extraction.

Table 3: *The logical expressions used in test pattern extraction*

pattern name	search pattern	logical expression in the programming language of Python
ATTR attribute	TW + noun	i < sent_len – 1 and lemmas[i].lower() in test_words and postags[i+1] == "S"
ATTR/AGR agreement	$[TW + noun]_{agr}$	i < sent_len – 1 and lemmas[i].lower() in test_words and postags[i+1] == "S" and forms[i] == forms[i+1]
ATTR/ST sentence starter	$[TW + noun]_{st}$	i < sent_len – 1 and i == 0 and lemmas[i].lower() in test_words and postags[i+1] == "S"
ATTR/4 four-spot	verb + X + TW + noun	i < sent_len - 3 and postags[i] == "V" and lemmas[i+2] in test_words and postags[i+3] == "S"
ADV adverb	adverb + TW	i < sent_len – 1 and postags[i] == "D" and lemmas[i+1].lower() in test_words
COMP Comparison	∃ comparative	lemmas[i].lower() in comp_words
PRED predicative	*olema* "be" + TW *olema* "be" + (adverb) + TW	i < sent_len – 2 and lemmas[i].lower() == "olema" and postags[i+1] and lemmas[i+2].lower() in test_words

The frequency detection of test patterns was restricted by the limits of sentences, and those test pattern occurrences crossing the sentence boundaries were not considered. The test pattern identification counted lemmas of the test words (*lemmas[i].lower() in test_words*) with the exception of test words with the endings "dud", "nud" and

16 https://github.com/ahtilohk/PSG227/blob/main/Test-patterns_occurrences_in_ENC2019_without_estnltk_corpus_processing_module.py

"tud". These are the cases of the non-inflected past participles and for those only text words were considered.[17] In the extraction of the comparison pattern, we used a general code that searched for the occurrences of comparative forms on the basis of a manually composed list.[18] The test words were untagged throughout the extraction process, i.e., their tagging status was unspecified.

It should be borne in mind that the frequency results directly depend on the quality of the tagging system used, which in this study was based on the Vabamorf morphological analyzer, as incorporated in the EstNLTK program. We are aware of the possibility that tagging and disambiguation errors (e.g. ambiguities caused by inflectional homonymy[19]) may have affected our analysis. We did not manually correct the shortcomings of the automatic analysis because a lexicographer would receive a statistical analysis based on the very same corpus processing methods when using a potential application based on this model.

Since the absolute frequencies of test words adopting the corpus patterns were not comparable, we operated with relative frequencies: the absolute frequencies matching the test pattern requirements were divided by the general lemma frequencies.

3.4 Deviation analysis as a similarity measure

To identify the dissimilarity between the morphosyntactic behavior of the prototypical adjectives and the two control groups of participles, we employed a method that we call deviation analysis. It can be used for the systematic comparison of the measurements of a target phenomenon with the respective measurements of a standard. There is no predefined formula in this method, and the relevant parameters are measured and compared one-by-one. Based on the measurements, a

17 Instead of *lemmas[i].lower() in test_words* in the logical expression *len(word) > 3 and word[-3:] in ["nud", "dud", "tud"] and word in test_words* are used.

18 To implement the comparison test pattern, we manually formed comparative forms of all of the words in our samples, following the comparative formation rule: the singular genitive form of the test word + *m* (e.g., *väike* "small" : *väikse* [small.GEN] : *väiksem* "smaller").

19 As a morphologically rich language, form homonymy is not rare in Estonian and the forms of different lemmas may coincide, e.g. the form *koon* can be analyzed as the nominative case form of the noun *koon* "snout", and the third person present indicative form of the verb *kuduma* "drink" ("I knit").

range of tolerance can be specified to decide the acceptability of the rates of the target phenomenon as compared to the standard.

In this study, we took the relative frequencies of the corpus patterns as relevant measurements and defined a range of tolerance for every pattern based on the respective values of the reference group of the validated adjectives.[20] The results for the control group words could then be subjected to deviation analysis. Additionally, the counts of deviating criteria per word allowed us to establish a scale of dissimilarity to the corpus behavior of adjectives. By specifying the ability of each pattern to exclude regular participles from the adjectival tolerance ranges, we evaluated the differentiation efficiency of each pattern.

4 Deviation analysis of the sample words

In this section, we present the results of the corpus extraction data based on the seven test patterns and the three sample groups. The relativized frequency statistics of the 100 validated adjectives are provided in Subsection 4.1, and these data are the basis for defining the tolerance ranges of adjectival behavior (4.2). Next, we relate the results of the control groups to the tolerance ranges, which enables us to define the deviation ranges and establish the degree of deviation of the control groups in relation to the reference group (4.3), and to assess the efficiency of the test patterns (4.4).

4.1 The test results for the validated adjectives

The variation of the relative frequency results based on the 100 words of the validated adjectives group according to the outcomes in the seven test patterns is presented in a box plot in Figure 1, while the descriptive statistics behind the variation are presented in Table 4, below.

20 For an approach that treats the values of the respective test patterns as a joint measure of overall similarity vs difference, see Vainik et al. (in press).

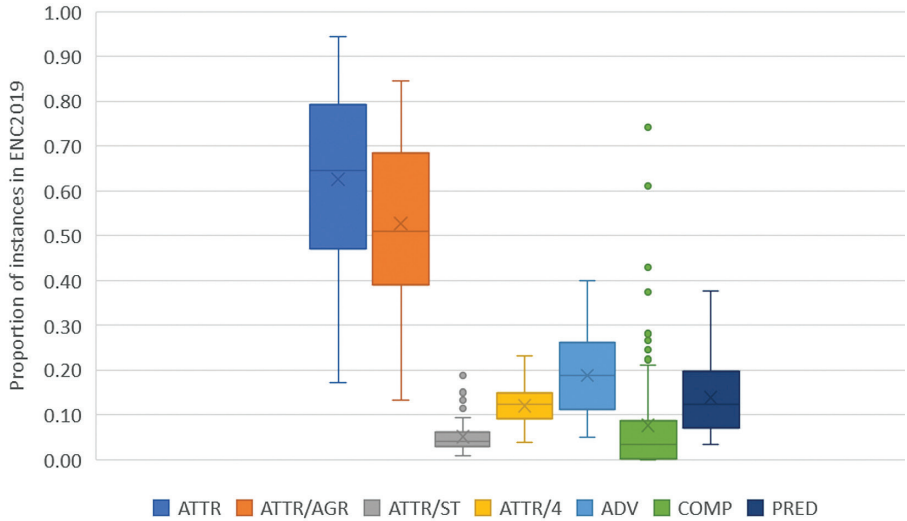

Figure 1: The division of the data of the test group of validated adjectives by the test patterns.

Table 4: *Descriptive statistics of the test group of validated adjectives by the test patterns*

	ATTR attribute	ATTR/AGR agreement	ATTR/ST sentence starter	ATTR/4 four-spot	ADV adverb	COMP comparison	PRED predicative
min	0.173	0.132	0.009	0.038	0.049	0.000	0.034
max	0.944	0.846	0.193	0.230	0.399	0.742	0.376
ave	0.626	0.527	0.051	0.120	0.188	0.076	0.138
stDev	0.192	0.185	0.034	0.037	0.084	0.120	0.074
median	0.644	0.510	0.042	0.123	0.187	0.034	0.124

Figure 1 and Table 4 show the average and median rates and the ranges of variation across the seven test patterns. The variation spans are notably wide, particularly in the two general attributive patterns: the attribute and the attribute agreement sequences.[21] The highest average and median values belong to the general attributive patterns, which is in accord with the assumption of the attributive function's prevailing status for adjectives (cf. Section 2.2; Erelt, 2017, p. 406). The high scores of the general attribute patterns (ATTR and ATTR/AGR) in-

21 For more about the correlation relations between the test patterns, see Vainik et al., (submitted).

dicate that the adjectives can quite freely function attributively in different positions of a sentence. The frequency of the complex attributive patterns (ATTR/ST and ATTR/4) is considerably more restricted, as these patterns combine multiple conditions besides the sequence of the test word and a noun (cf. Table 3 in Section 3.1).

The patterns with the most discrepant results are the four-spot-pattern and the comparison pattern, revealing several outliers. The non-attributive patterns – the adverb pattern (ADV), the comparison pattern (COMP), and the predicative pattern (PRED) – overall demonstrate relatively low levels of relative frequencies and variation ranges. The average rate of comparative forms is strikingly low, which is unexpected given the assumed prototypical nature of the validated adjectives. There are seven distinctly non-scalar adjectives without any occurrences in the comparison pattern, for instance *kahekordne* "double, two-floored" and *eelmine* "previous". The outliers deviating from the general tendency, i.e. the adjectives with exceptionally high results in the comparison pattern, are *kõrge* "high", *lihtne* "simple", *täpne* "precise" and *lahja* "lean".

4.2 Setting the adjectival limit ranges

The marginal rates of relative frequencies in the validated adjective group lay the foundation for postulating ranges of tolerance for the test patterns. The maximum and minimum values of the patterns (see Table 4), except for the comparison pattern, serve as the highest and lowest values of the corresponding tolerance ranges. The evaluation of the comparison pattern differs from other patterns: here we estimate the absolute frequency of a word's comparative form in the corpora. We consider an absolute frequency of higher than five occurrences to be a sign of non-occasional comparison formation, and hence not deviating from the adjective range.

To sharpen the contrast of adjectives from regular participles, we qualitatively adjusted the limits of the attribute pattern and the sentence starter pattern. In this process, we excluded a few highly deviating adjectives from the ranges of these patterns to better capture the essence of prototypical adjectives. The exclusion was done by comparing the test groups and considering a pattern's ability to differentiate

validated adjectives from regular participles individually. For example, setting the minimum value for the attribute pattern (ATTR) from 0.173 to 0.246 (by excluding the result of one validated adjective[22]), allowed us to differentiate 13 words from the regular participles group that would fit in the tolerance range if this one distinct adjective were not excluded. The ranges of tolerance are presented in Table 5.

Table 5: *The ranges of tolerance. Limits of the prototypical adjective for the seven test patterns*

test patterns	adjectival ranges (relative frequencies)
ATTR – attribute	0.246–0.920
ATTR/AGR – agreement	0.132–0.846
ATTR/ST – sentence starter	0.015–0.193
ATTR/4 – four-spot	0.038–0.230
ADV – adverb	0.049–0.399
COMP – comparison	*5 <
PRED – predicative	0.034–0.376

* The range of comparative forms is calculated in absolute frequencies

The analysis below, addressing the degrees of deviation from the adjective behavior and the differentiation efficiency of the test patterns, is based on the ranges defined in Table 5.

4.3 Assessing the control groups. Defining the deviation scales

Using the ranges of tolerance established in the previous section, we analyzed whether the test words fit into the limits set by the validated adjectives. To do that we counted both inclusive and non-inclusive results regarding the tolerance ranges for each pattern. To illustrate the analysis, we present the results for six control group participles across all seven test patterns in Table 6.

22 This word is *kade* "envious", which occurs extraordinarily rarely in the attributive function compared to other validated adjectives.

Table 6: *Deviation analysis of six test words from the control groups*

group**	test word	ATTR	ATTR/ AGR	ATTR/ ST	ATTR/ 4	ADV	COMP	PRED	total
Ekilex	*kleepuv* "sticking"	+	+	+	+	+	+	+	0
regular	*korratav* "being repeated, repeatable"	+	+	–	+	+	+	+	1
Ekilex	*maksustatud* "taxed"	+	–	–	+	+	–	+	3
regular	*kaevav* "digging"	+	+	–	+	+	–	–	3
Ekilex	*soostunud* "consented; swampy"	+	–	–	–	–	–	+	5
regular	*naerdud* "laughed"	–	–	–	–	–	–	+	6

** Ekilex = control group 1, the adjectivizing participles in Ekilex;
regular = control group 2, the regular participles

Let us now consider how much all of the words of the three test groups deviate from the tolerance ranges. Table 7 presents the test words measured according to the number of deviating patterns. As the table shows, 89 validated adjectives result in zero deviation and 10 adjectives deviate by one pattern, while one adjective deviates by two patterns. Based on the results for all three test groups, we define three degrees of deviation from the adjectival behavior: the no deviation (0–1 patterns deviating from the tolerance ranges), low deviation (2–3), and high deviation (4–7) scale.

Table 7: *The deviation scale*

no. of patterns deviating	0	1	2	3	4	5	6	7
adjectives (%)	89	10	1					
Ekilex participles (%)	6	27	39	16	9	3		
regular participles (%)	1	16	14	20	16	21	11	1
degrees of deviation	**no deviation**		**low**		**high**			

Although 99% of the validated adjectives fall within the highest prototypicality degree of no deviation, there are words that do not

score in all patterns even in this group, which is explainable by the adjustments of adjectival ranges described in Section 4.1. The test words with one deviating pattern are adjectives that do not form comparatives due to semantic restrictions (five adjectives, e.g. *ühetoaline* "one-roomed" and *vasak* "left") or reach the lower limit of the tolerance range in the attribute pattern (two adjectives, *kade* "envious" and *sõjaline* "military") or in the sentence starter pattern (five adjectives, e.g. *selge* "clear" and *vajalik* "necessary"). The validated adjective deviating in two test patterns is *kade* "envious", an adjective favoring non-attributive usage.

Comparing the results for all three test groups proves our hypothesis: the adjectivizing participles in Ekilex correspond to the adjectival behavior to a larger extent than the regular participles group. The deviation analysis shows that 33% of the Ekilex participles and only 17% of the regular participles match the no-deviation space with no or one pattern deviating from the tolerance ranges. Altogether 88% of the adjectivizing participles in Ekilex and 51% of the regular participles fall within the low or no deviation space. Only 12% of the Ekilex participles but 49% of the regular participles are situated at the high deviation level. Note also that all of the test words from the Ekilex group show at least two patterns within the ranges of tolerance.

4.4 Estimating the efficiency of test patterns

In this section, we evaluate the efficiency of the seven test patterns, i.e. the ability of each pattern to exclude regular participles from the tolerance ranges defined on the basis of the variation scope of the validated adjectives (see Table 5 in Section 4.1). The efficiency is assessed by the extent of the difference between the control group and reference ranges. Basically, the bigger the gap between the results of the two groups, the better the corresponding pattern's efficiency.

First, we calculate the differentiation efficiency of the patterns by comparing the results for the validated adjectives and regular participles group in terms of how many test words fit into the reference ranges of corresponding patterns. The results are presented in Table 8, with the patterns ordered from stronger to weaker efficiency, from left

to right. The values of the adjectives[23] represent 100%, and the corresponding ratio the result of the regular group – the gap between these two is the difference (the results for adjectives minus participles). For the collation of the data, the table also includes the results for the adjectivizing participles in the Ekilex group that fall, as we hypothesized, in between the validated adjectives and regular participles group.

Table 8: *The differentiating efficiency of test patterns; regular participles versus adjectives*

	COMP comparison	ATTR/ST sentence starter	ATTR/AGR agreement	ATTR attribute	ATTR/4 four-spot	PRED predicative	ADV adverb
adjectives	**94**	**95**	**100**	**99**	**100**	**100**	**100**
Ekilex participles	26	74	51	93	91	72	89
regular participles	**14**	**25**	**49**	**62**	**63**	**66**	**75**
difference	*85%*	*74%*	*51%*	*37%*	*37%*	*34%*	*25%*

According to the data presented in Table 8, the most efficient differentiator (the pattern that leaves the most regular participles out of the tolerance range) is the comparison pattern, also excluding a significant number of adjectivizing participles in Ekilex. The second strongest differentiator is the sentence starting pattern, which places 74% of Ekilex participles together with validated adjectives and leaves 74% of the words of the regular participles group outside of the tolerance range. The results for the attribute pattern and the four-spot pattern overlap, suggesting that the test battery would not suffer if one of them were excluded (at least for the analysis of participles). Overall, the results indicate that in each test pattern the adjectivizing participles in Ekilex fall between other test groups, exhibiting lower adjectival scores than the validated adjectives, but significantly higher scores than the regular participles.

The jitter plots below illustrate the distribution of the results of a strong differentiator (the sentence starter pattern, Figure 2) and a weak differentiator (the predicative pattern, Figure 3). The values – relative pattern frequencies for each test word – of the three test groups are

23 Note that the reason for the number of adjectives in three test patterns falling under 100 in Table 8 is that the adjectival ranges have been slightly adjusted (cf. Section 4.1 and Table 5).

presented on the x-axis; for the y-axis, the plots show randomly generated values, ensuring that the dots do not overlap.

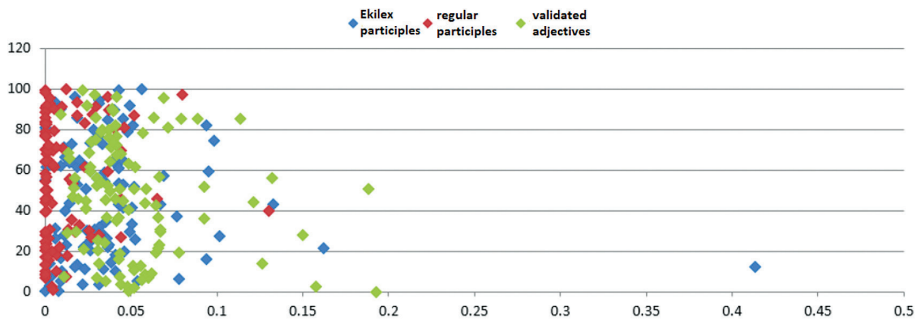

Figure 2: Distribution of results within the sentence starter pattern.

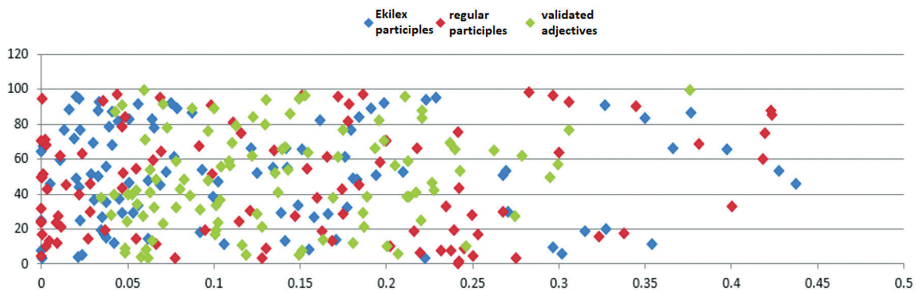

Figure 3: Distribution of results within the predicative pattern.

As the distribution of the results demonstrates, the two patterns differ considerably in their ability to differentiate the test groups. The distribution of the results based on the sentence starter pattern shows the results for the regular participles cumulating near a value of 0, while those for the adjectivizing participles in Ekilex and validated adjectives are split between 0 and 0.2. One of the weakest differentiators, the predicative pattern, spreads the results for the three test groups more evenly over a wider range, from 0 to 0.44.

5 Conclusions

The assessment of the limits of prototypical adjectivity carried out in this study confirmed that it is possible to capture the adjectival corpus behavior by morphosyntactic sequences typical of adjectives. We ap-

proached adjectivity via the most salient morphosyntactic properties of adjectives generalizable by the attributive vs. non-attributive opposition. Operationalizing these main parameters into seven sequential corpus patterns helped us to establish the ranges of variation within the defined limits of tolerance. We can conclude that the test patterns clearly distinguished the group of validated adjectives from the two control groups of participles.

The analysis showed that the validated adjectives in the test group were not homogeneous either: their results spread over three different ranges in terms of the results in the deviation analysis (zero, one or two patterns outside the ranges of the prototypical adjective), showing variance to a certain degree and proving the prototype-based nature of the adjective class. The deviation analysis resulted in a tripartite scale of similarity to adjectives in terms of deviation from the tolerance ranges set according to the variation of the group of adjectives. The overall scale of adjectivity was achieved by calculating the ratio of deviating and coinciding criteria (see the scale of deviation in Table 7). According to the deviation analysis, 12% of the adjectivizing participles in Ekilex and 49% of the regular participles were assessed as highly deviating from the validated adjectives, a result proving that the participles of these two groups differ in degrees of adjectivization. Moreover, and as we hypothesized, the adjectivizing participles in Ekilex (adjective candidates) fell closer to the validated adjectives than the regular participles.

The most accurate differentiation of the regular participles group from the validated adjectives was achieved by the comparison and sentence starter patterns. The results for the validated adjectives indicate that the occurrence of comparative forms is not necessarily frequent even for presumably prototypical adjectives, and thus may leave out words perfectly eligible for adding as dictionary entries. The infrequency of validated adjectives is striking, even in the sentence starter pattern. This indicates that the general adjectival properties (e.g. the simple attribute pattern or the predicative pattern) are not necessarily the clearest distinguishers between adjectival and non-adjectival behavior. The adverb pattern was the weakest differentiator according to the comparison of regular participles and validated adjectives. This result diverges from our pilot study with smaller test groups, in which

the adverb pattern was one of the three best differentiators (Tuulik et al., 2022, p. 296). We can conclude that the best differentiators are not necessarily the most typical adjectival properties (attributive and predicative), but more specific markers of potential morphosyntactic behavior. There are patterns that strongly exclude nonadjectival words (good excluders) and patterns that strongly include (good includers) adjectival words with the notion of prototypical adjectives.

The efficiency analysis revealed that the selection and constitution of patterns used in this study could be elaborated further to optimize the results. The comparison pattern extraction process would be facilitated by developing an automatic generator of comparative forms. Since the two attributive patterns, the attribute and four-spot patterns, show quite similar differentiating results, one of them can be left out of the test battery without weakening the results. The attribute pattern may be preferable as a necessary sequence since it shows slightly better results and is structurally simpler to use in the extraction process. As the results of the attribute agreement pattern were more or less the same in the control groups, due to the identical selection of the declinable present and non-declinable past participles in both, we can conclude that the agreement pattern could be more useful in connection with some other categories, e.g. in the assessment of the adjectival behavior of nouns.

The predicative and adverb patterns also need further adjustment: in their current forms they do not clearly differentiate regular participles from adjectivized ones, or from validated adjectives. One solution would be to include in the extraction code of the predicative pattern certain morphological restrictions by defining the predicative case forms. Moreover, setting the presence of the negation word *ei* "not" in the near context of the test word could also help to highlight verbal uses of participles. The adverb pattern may be elaborated by adding an inclusive search list of intensifying adverbs to the corpus extraction algorithm in order to avoid typical verb modifiers, e.g. adverbs of manner.

Ultimately, it is important to acknowledge the effects such patterns have in concurrence. But how many patterns are necessary to achieve the optimal results? In light of this study, we suggest that a proper test battery should include at least five patterns to capture the morphosyn-

tactic behavior of the versatile class of adjectives. The composition of patterns may be adjusted according to the lexical group targeted for assessment. It is also possible to use different methodological solutions and analyze the results for the test patterns in concurrence instead of a sum of separate values (for a Euclidean distance approach, see Vainik et al., in press).

The use of a quantitative approach can reveal unexpected aspects of a language, and the findings of this study have the potential to contribute to the knowledge of adjectives in Estonian, and also indicate the value of further investigations into this topic. When it comes to the contrasting focus of this study, the Estonian participles, the analysis revealed some similarities with adjectives as exemplified by the two control groups of participles. Another finding is connected to the subtypes of participles: the results of the deviation analysis show that the present participles congregate among the higher scores (in particular passive present participles, such as *hinnatav* "assessable") and the past participles fall within the lower scores. At least in part this is due to the fact that past participles cannot perform in the attribute agreement pattern, but there may be other factors affecting the similarity to the adjectives. Overall, the reasons for the general tendencies as well as outliers in the data deserve a closer, qualitative analysis.

In our opinion, the results of this study can be applied to develop a multi-parameter application for determining the relative adjectivity of a word or a word form, e.g. adjectivizing participles or nominals (for the border areas of adjectives with other lexical classes in Estonian, see Vainik, Paulsen and Lohk 2020). As the morphosyntactic patterns characteristic of a PoS are language-specific, so is the outcome of our examination. The results are, however, also adjustable for the analysis of other languages.

Acknowledgments

This research was supported by Estonian Research Council grant PSG227.

Abbreviations

ABE – abessive case
ADE – adessive case
ADV – adverb pattern
ATTR – attribute pattern
ATTR/AGR – attribute agreement pattern
ATTR/ST – sentence starter pattern
ATTR/4 – four-spot pattern
COMP – comparison pattern
CMPR – comparative form
ELA – elative case
GEN – genitive case
NOM – nominative case
PART – partitive case
PL – plural
PRED – predicative pattern
PTCP – participle
SUP – superlative
TRA – translative case

References

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press.

CombiDic = The EKI Combined Dictionary. (2020). Hein, I., Kallas, J., Kiisla, O., Koppel, K., Langemets, M., Leemets T., ..., & Voll, P. Institute of the Estonian Language. Retrieved from https://sonaveeb.ee (15. 11. 2022)

Dixon, R. M. W. (2004). Adjective classes in typological perspective. In R. M. W. Dixon & A. Aikhenvald (Eds.), *Adjective classes: a cross-linguistic typology*. Oxford: Oxford University Press.

Ekilex. Retrieved from https://ekilex.eki.ee/ (26. 11. 2021)

Erelt, M. (2017a). Omadussõnafraas [The adjective phrase]. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* [The Syntax of Estonian] (pp. 405–415). Eesti keele varamu III. Tartu Ülikooli Kirjastus.

Erelt, M. (2017b). Öeldistäide [The predicative]. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* [The Syntax of Estonian] (pp. 278–288). Eesti keele varamu III. Tartu Ülikooli Kirjastus.

Erelt, M. (2017c). Öeldistäitemäärus [The predicative adverbial]. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* [The Syntax of Estonian] (pp. 289–299). Eesti keele varamu III. Tartu Ülikooli Kirjastus.

Erelt, M. (2017d). Sekundaartarindiga laused [Sentences with secondary constructions]. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* [The Syntax of Estonian] (pp. 756–840). Eesti keele varamu III. Tartu Ülikooli Kirjastus.

Erelt, M. (2017e). Sissejuhatus süntaksisse [Introduction to syntax]. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* [The Syntax of Estonian] (pp. 53–89). Eesti keele varamu III. Tartu: Tartu Ülikooli Kirjastus.

The Estonian Collocations Dictionary = Eesti keele naabersõnad. (2019). Kallas, J., Koppel, K., Paulsen, G. & Tuulik, M., Institute of the Estonian Language. Retrieved from http://www.sonaveeb.ee (14. 11. 2022)

The Estonian-Russian Dictionary (2019). Laasi, H., Lagle, T., Leemets, H., Liiv, M., Pärn, H., Simm, L., ..., Tubin, V., (Comp.); Liiv, M., Melts, N., Romet, A., Kallas, J., Riikoja, E., Martoja, I., Smirnov, S., ..., & Veskimägi, E. (Eds.). doi: *10.15155/3-00-0000-0000-0000-0001BL*

Geeraerts, D. (1989). Prospects and problems of prototype theory. *Linguistics, 27*, 587–612.

Habicht, K., Kaalep, H.-J., Muischnek, K., Müürisep, K., & Rääbis, A. (2000). Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest. *Keel ja Kirjandus, 9*, 623–633.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers.

Kaalep, H.-J., Kirt, R., & Muischnek, K. (2012). A trivial method for choosing the right lemma. In *Baltic HLT* (pp. 82–89).

Kallas, J., Tiits, M., Tuulik, M., Koppel, K., & Jürviste, M. (2014). Eesti keele põhisõnavara sõnastik [The Basic Estonian Dictionary]. Tallinn: Eesti Keele Sihtasutus.

Kallas, J., & Tuulik, M. (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. *Estonian Papers in Applied Linguistics, 7*, 59–75.

Kallas, J., Tuulik, M., & Langemets, M. (2014). The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian. In A. Abel, C. Vettori & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress*: *The User in Focus, 15–19 July 2014, Bolzano, Bozen* (pp. 1109–1119). Bolzano/Bozen: Institute for Specialised Communication and Multilingualism. Retrieved from https://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_086_p_1109.pdf

Kasik, R. (2015). *Sõnamoodustus*. Tartu: Tartu Ülikooli Kirjastus.

Kerge, K. (1998). Vormimoodustus, sõnamoodustus ja leksikon: oleviku keksõna võrdluse all. Tallinn: TPÜ Kirjastus.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX*

International Congress, 6–10 July 2004, Lorient, France (pp. 105–116). Lorient: Université de Bretagne Sud. Retrieved from https://euralex.org/category/publications/euralex-2004/

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7–36.

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In I. Kosem, T. Zingano Kuhn., M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.). *Proceedings of the eLex 2019 conference: Smart lexicography, 1–3 October 2019, Sintra, Portugal* (pp. 434–452). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_24.pdf

Laur, S., Orasmaa, S., Särg, D., & Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference, May 2020, Marseille, France* (pp. 7152–7160). European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/2020.lrec-1.0.pdf

Orasmaa, S., Petmanson, T., Tkatšenko, A., Laur, S., & Kaalep, H-J. (2016). EstNLTK – NLP Toolkit for Estonian. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & P. Stelios (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia* (pp. 2460–2466). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

Pajusalu, R. (2017). Nimisõnafraas [The noun phrase]. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* [The Syntax of Estonian]. *Eesti keele varamu III. Tartu Ülikooli Kirjastus* (pp. 379–404).

Paulsen, G., Vainik, E., Tuulik, M., & Lohk, A. (2019). The Lexicographer's Voice: Word Classes in the Digital Era. In I. Kosem, T. Zingano Kuhn., M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.). *Proceedings of the eLex 2019 conference: Smart lexicography, 1–3 October 2019, Sintra, Portugal* (pp. 319–337). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_18.pdf

Paulsen, G., Vainik, E., & Tuulik, M. (2020). Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias [On word classes in contemporary lexicography: The lexicographers" view]. *Estonian Papers in Applied Linguistics, 16*, 177–202.

Paulsen, G., Vainik, E., Lohk, A., & Tuulik, P. (2021). Catching lexemes. The case of Estonian noun-based ambiforms. Electronic lexicography in the 21st century. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (Eds.), *Proceedings of the eLex 2021 conference: Post-editing lexicography, 5–7 July 2021, Brno, Czech Republic* (pp. 288–311). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_17_pp288-311.pdf

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

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology, General, 104*(3), 192–233.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, Lawrence Erlbaum, New York.

Tavast A., Koppel, K., Langemets, M., & Kallas, J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, *Alexandroupolis, 2021, online, Vol. 1* (pp. 215–223). Greece: Democritus University of Thrace.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts*, *17–21 July 2018, Ljubljana, Slovenia* (pp. 749–761). doi: 10.4312/9789610600961

Tuulik, M., Vainik, E., Paulsen, G., & Lohk, A. (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. doi: 10.5128/ERYa18.16.

Vainik, E., Paulsen, G., & Lohk, A. (2021). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, 7–11 September 2021, Alexandroupolis, Greece, Vol. 1* (pp. 119–130). Alexandroupolis, Greece: Democritus University of Thrace. Retrieved from https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p119-130.pdf

Vainik, E., Lohk, A., & Paulsen, G. (2021). The Distribution Index Calculator for Estonian. Electronic lexicography in the 21st century. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (Eds.), *Proceedings of*

the eLex 2021 conference: Post-editing lexicography, 5–7 July 2021, Brno, Czech Republic (pp. 121–138). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_07_pp121-138.pdf

Vainik, E., Paulsen, G., Tuulik, M., & Lohk, A. (in press). Towards the Morpho-syntactic Corpus Profile of Prototypical Adjectives in Estonian. *Estonian Papers in Applied Linguistics.*

Vare, S. (1984). *Omadussõnaliited tänapäeva eesti kirjakeeles.* [The adjectival suffixes in contemporary Estonian]. Tallinn: Valgus.

Viitso, T.-R. (2003). Structure of the Estonian language: Phonology, morphology, and word formation. In M. Erelt (Ed.), *Estonian language* (pp. 9–92). Tallinn: Estonian Academy Publishers.

Warren, B. (1984). *Classifying Adjectives*. Göteborg: Acta Universitatis Gothoburgensis.

Od glagolskega k pridevniškemu: vrednotenje leksikalizacije deležnikov v estonskem korpusu

Članek obravnava vprašanja kategorizacije, povezana s pridevniškimi kandidati v estonščini, s poudarkom na deležnikih. Cilj analize je bil oceniti razpon prototipskega pridevnika in določiti stopnjo njegovega odstopanja na lestvici prototipskosti. Raziskava je temeljila na skupini potrjenih pridevnikov – izbranih pridevnikov, vključenih v *Osnovni slovar estonščine* – in dveh kontrolnih skupinah bolj in manj leksikaliziranih deležnikov. Preverili smo sedem morfosintaktičnih korpusnih vzorcev, značilnih za pridevnike. Testni vzorci so temeljili na prototipskih značilnostih pridevnika in na opažanjih pri konkretni leksikografski analizi. Za oceno vzorčnih besed in določitev pomena testnih vzorcev z vidika opredelitve pridevnika smo uporabili analizo odklona. Rezultate študije je mogoče uporabiti za določitev merila pridevniške ustreznosti pri leksikografskih presojah, na primer za določanje leksikaliziranih deležnikov.

Ključne besede: korpusno jezikoslovje, leksikografija, estonski jezik, pridevnik, deležnik, analiza odklona