From experiments to an application: the first prototype of an adjective detector for Estonian

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

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia
2 Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden
3 Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
E-mail: geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se, ahti.lohk@taltech.ee, maria.tuulik@eki.ee, ene.vainik@eki.ee

Abstract

In this study, we discuss the process of developing a multi-parameter application – the adjective similarity calculator (ASC) – that determines the relative adjectivity of a word or a word form. The tool relates the statistical summary of a word (form)'s corpus behaviour to the most typical and central aspects of the Estonian adjective: the adjectival corpus profile. To establish this profile, we use close-context patterns characterising adjectives and detectable in the corpus (see the experiments in Tuulik et al. 2022, Paulsen et al. 2022, and Vainik et al., 2023). The first prototype of the ASC will be evaluated based on clear cases of adjectives and PoS representatives overlapping with adjectival properties, but also based on words representing more distant classes. The main purpose of the application is to improve lexicographic work in categorisation procedures of the partly overlapping lexical categories to the adjective, particularly in such ambiguous cases as adjectivised participles, nouns and adverbs.

Keywords: language technology; lexicography; corpus linguistics; adjective; the Estonian language

1. Introduction

The identification of the boundaries between lexical categories is a common task in part-of-speech tagging and lexicographic procedures. In many languages, these boundaries can be rather blurred. One of the most problematic word classes for lexicographers working with Estonian is the adjective (Paulsen et al., 2019, 188–189), a category overlapping with the noun, verb, adverb, pronoun (see Vainik, Paulsen, Lohk, 2021: 122–123) and ordinal (e.g., Erelt, 2017: 63). Lexicographers need to make decisions about lexicalising participles, a phenomenon common for other languages as well (e.g. English, where participles tend to develop into full-blown adjectives, such as *blessed* and *hammered*). Another phenomenon yielding ambiguity between lexical categories is systematic polysemy (see Langemets 2010, 159–161), emerging as conversional transposition (see Vare, 2006:199), in which a word can be used in another category without changing its form, e.g. *vigur* 'trick' (noun); 'tricksy, prankish' (adjective).

The prototypical behaviour of a word class can be captured by using corpus data, in the form of a corpus profile gathering the central morphosyntactic patterns characteristic to the category (see Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023). Is this profile operational as a template for comparing particular words or word forms? Motivated by this question, we introduce the first working prototype of the Adjective Similarity Calculator (ASC). This multi-parameter application is designed as a tool for lexicographers working with contemporary Estonian. The ASC is based on a statistical summary of a word (form)'s corpus behaviour[1] in comparison to the most typical aspects of the Estonian adjective. To establish the adjectival corpus profile, we use a selection of the most central close-context patterns characterising adjectives and detectable in the corpus (see the experiments in Tuulik et al, 2022, Paulsen et al 2022, and Vainik et al., 2023). To measure the distance of a word from the adjectival profile, we have selected an approach we call conformity assessment, derived from the methods we have tested in our previous studies (see Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023).

The ASC elaboration process comprises two main optimisation issues: 1) the scope of the overlapping parts of speech targeted by the calculator, and 2) the optimisation of the thresholds of adjectivity on the basis of the results of a statistical analysis. The constituency of the set of automatically searchable test patterns should be applicable to all of the word classes overlapping with adjectives. The second issue involves adjustments to the method we use for calculating the distance of a word from the adjectival profile (see Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023).

We will begin with a short overview of the Estonian adjective and the theoretical foundations behind the development of the adjectival corpus profile in Section 2. Here we describe the idea behind the statistic and its calculation and explain the similarity estimation method we call conformity assessment. The details of its realisation as a script interacting with the corpus via Sketch Engine API are given in Section 3. Section 4 is devoted to the demonstration of the results illustrated by the examples from seven lexical classes. The results are compared with the decisions made by lexicographers in the EKI Combined Dictionary (CombiDic) and checked against the corpus data using the Sketch Engine tool Word Sketch and concordances. The problems and future directions of development are discussed in Conclusions.

[1] The mechanism of the tool developed in this study can roughly be compared to the Find X function of Sketch Engine, providing additional information about the usage of a word; the solution is described in Kilgarriff and Rychlý (2008). The Find X function uses frequencies of word forms to determine whether a word is predominantly used in plural or singular, whether a verb appears more in the present participle than in the passive form etc. The difference is, however, that our assessment battery is based on frequency data of a set of corpus patterns, not on frequencies of certain forms of a word.

2. Background

2.1 The Adjectival behaviour and its measurable patterns

In Estonian, there are five main word classes overlapping with adjectives: nouns, verbs, adverbs, pronouns and ordinals. Since the last two represent closed classes, we can say that the classes posing problems for lexicographers are mainly nouns, participles and adverbs. The noun-adjective type is the largest group showing ambiguity in word class[2], typically via transpositional derivation forming systematic polysemy networks (see Vare, 2006; Langemets, 2010). The second largest type is the adverb-adjective, consisting of words occurring in contexts typical of both classes, such as verbal or nominal modifiers. The transition zone between verbs and adjectives comprises the non-finite forms of verbs: participles[3], gerunds and supines. (For a typology of overlapping lexical categories in Estonian, see Vainik et al., 2021.) The determination of the lexicalisation degree of these forms is a challenge for lexicographers and also poses huge problems for automatic morphological analysis.

Hence, there are several lexical categories approaching the morphological, syntactic and semantic properties[4] of the adjective. Characteristically, the adjective occurs in a sentence together with a noun that it describes or modifies. The morphological characteristics of Estonian adjectives include inflection in case and number, forms of gradation and derivation. Syntactically, the adjective constitutes an adjective phrase by itself or together with its modifier(s). The constructions in which an adjective is most recognisable are those where it occurs as an attribute (1a) or as a predicative (1b). (About the Estonian adjective, see Viitso, 2001: 32–35, 42; Erelt, 2017: 405–406.) The adjective can be modified by an adverb in all of these configurations, exemplified below by the sentence (1b), where the intensifying adverb *täitsa* ’quite’ precedes the predicative adjective *põnev* ’exciting’.

(1a) *Matka-me* *lumis-te-s* *mäge-de-s.*
hike-1PL snowy-PL-INE mountain-PL-INE
‘We hike in the snowy mountains.’

[2] Based on an analysis of the database on words and forms that are ambiguous in terms of their PoS categorisation, compiled mainly from lexicographic sources (Vainik et al., 2021: 122).

[3] Participle endings in Estonian function partly as grammatical and partly as lexical suffixes (see Viht & Habicht 2019: 37); usually, participles are not regarded as independent PoS, except for corpus-tagging systems.

[4] The semantic properties an adjective typically describes centre around dimension, age, value and colour (Dixon 2006: 3–4); the adjective has no internal temporal structure (expressing states rather than activities and permanent rather than temporary characteristics, see e.g. Fábregas, Marín 2017); the adjective can have semantic valency (Helbig 1992; Haugen 2013).

(1b) *Film* *on* *täitsa* *põnev.*
film is quite exciting-NOM
'The film is quite exciting.'

A corpus-based application aimed at the identification of adjectival morphosyntactic patterns must focus on the structures that emerge as the most distinctive, as well as being detectable by the corpus tagging system. In our previous studies, we tested seven adjectival patterns (Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023). We screened out four patterns[5] that instantiate a central set of parameters of adjectival corpus behaviour. The selection is based on attributive and predicative constructions, but also the modifiability of an adjective candidate by an intensifying adverb (the abbreviation TW stands for the target word assessed for adjectival behaviour):

1) **the attribute pattern** (ATTR), targeting the sequence of the TW immediately preceding a noun. This pattern is based on the tendency of an adjective to modify the noun as an attribute (TW_NOUN), cf. *kollane pall* 'yellow ball'.

2) **the sentence starter pattern** (ATTR/ST) adds a syntactic restriction to the attribute phrase by restricting its location at the beginning of a sentence. This differentiates inter alia verbal participles from adjectivised ones (e.g. past participles in compound tenses require the preceding auxiliary verb *olema* 'be').

3) **the adverb pattern** (ADV) targets the sequence of ADVERB_TW, a characteristic pattern of the adjectives in the corpus, particularly with scalar adjectives.

4) **the predicative pattern** (PRED) combines two sequences: a) the copula verb *olema* 'be' followed by the TW, and b) a copula verb followed by an adverb and the TW.

To improve the distinction of adjectival behaviour, we added an inclusive list[6] of over 66 selected adverbs in queries of the adverb and predicative patterns (see Appendix 1). Hence, the patterns involving adverbs include only the adverbs typically modifying

[5] We have excluded e.g. the pattern ascertaining the agreement condition from the set of the attribute patterns because it excludes indeclinable adjectives and (also indeclinable) lexicalised past participles. Another pattern characteristic to adjectives left out of the final set is the gradation pattern, because the study of prototypical adjectives showed considerable variation in the occurrence of comparative forms (see Paulsen et al 2022: 89–92). Also, a precondition for the use of the gradation pattern is an automatic generator of comparative forms of any given word, which would considerably increase the "footprint“ of corpus data analysis.

[6] The list was compiled using the Sketch Engine word list tool, through which the 100 most frequent adjectives were extracted and the 30 most frequent adverbs for each of these adjectives were selected. The adverbs with frequencies of 10 or more were included in the list; some of the less frequent adverbs were included if they clearly expressed properties typical of adjective modifiers (e.g. intensifiers).

adjectives, leaving out, for instance, manner adverbs that predominantly modify verbs.

2.2 Conformity assessment and the estimated ranges of normal variation

The selection of the statistical method to calculate the similarity of a word with the prototypical adjective was based on previous experiments of three methods: conformity assessment[7], Euclidean distance and cosine similarity (Tuulik et al., 2022, Paulsen et al 2022a, Paulsen et al., 2022b, Vainik et al., 2023). Since the conformity assessment proved to be the most flexible (making possible the qualitative adjustment of the adjectival ranges of different lexical groups during the testing process) and, unlike the other tested methods, this enabled us to analyse the performance of a target word in different patterns separately[8], we chose this method as the similarity assessment measure for the ASC.

Conformity assessment allows for the systematic comparison of the relative frequency values of a target phenomenon with the respective measurements of a standard. There is no predefined formula in conformity assessment, and the relevant parameters are estimated and compared one by one. On the basis of the measurements, it is possible to identify the ranges of adjectival behaviour typical for each pattern. This approach relies on the prototype theory and the idea that a lexical class is not a clear-cut phenomenon but shows variance to a certain degree[9] (about the application of the prototype theory in lexical semantics, see e.g. Berlin & Kay 1969; Geeraerts 1989).

Using this approach, we operated with relative frequencies[10] of a target word's occurrences in the four selected corpus patterns (cf. Section 2.1). We defined a range of adjectival behaviour for every pattern based on the marginal rates of the 100 most central and prototypical adjectives in Estonian (for a detailed description of the setting of ranges and the selection of the sample adjectives, see Paulsen et al., 2022a[11]). These

[7] We have used the term *deviation analysis* for this method in our previous studies (Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023); the shift of perspective from deviation to conformity is for practical reasons: the application assessing a word's adjectivity counts matches of the behaviour of the prototypical adjective within the predetermined ranges of variation; thus, the process concerns compliance with the standard rather than deviation from it.

[8] This is important regarding the main user group – the lexicographers – who may need to acquire explicit information about the patterns that the target word performs, such as an adjective (or not).

[9] Our previous study (Vainik et al., 2023) indicated that even words marked as adjectives in dictionaries may differ in how high they score in different patterns. For instance, the actual usage of adjectives tends to incline towards either attributive (ATTR and ATTR/ST) or non-attributive (ADV, PRED) patterns. Hence, the patterns have a co-effect within a predetermined variation space.

[10] For frequency results to be comparable, the absolute frequencies of the corpus pattern occurrences are divided by the word's general lemma frequencies.

[11] The sample of prototypical adjectives was randomly selected from lexicographically verified adjectives in the Basic Estonian Dictionary (about the dictionary, see Kallas et al., 2014). Note also that the analysis in Paulsen et al. (2022a) was based on the state-of-the-art ENC

ranges represent the estimation of normal variation for adjectives and define the adjectival corpus profile. The ranges of adjectival behaviour of the four morphosyntactic patterns selected as the basis for the ASD are presented in Table 1:

Patterns	**adjectival ranges (relative frequencies)**
ATTR	0.246–1
ATTR/ST	0.015–0.193
ADV	0.01**–**1
PRED	0.036–0.344

Table 1: The ranges of adjectival behaviour, defining the adjectival corpus profile

Although the adjectival ranges primarily drew on the corpus behaviour of the sample of 100 prototypical adjectives[12], we adjusted the ranges qualitatively to improve their ability to differentiate other word classes, particularly participles from adjectives. For example, when setting the range for the adverb pattern (ADV), we excluded the results of highly deviating adjectives (the non-scalar adjectives, e.g. *ühetoaline* ‘one-room (flat)’, *vasak* ‘left’ and *homne* ‘tomorrow’s’) by raising the lower limit. Also, to avoid excluding perfectly clear adjectives (e.g. *haruldane* ‘rare’), we raised the upper limits of the attribute (ATTR) and adverb (ADV) patterns to the maximum (1). Table 2 provides examples where the relative frequency results of the example words are analysed as either a conforming result (1) or non-conforming result (0) to the ranges of adjectival behaviour of the four corpus patterns:

Word	**ATTR**	**ATTR/ST**	**ADV**	**PRED**	**Conforming patterns**
uhke ‘proud’	0.473 (1)	0.03 (1)	0.112 (1)	0.19 (1)	4
haihtuv ‘vanishing’	0.72 (1)	0.037 (1)	0.028 (1)	0.028 (0)	3
õnnitletud ‘congratulated’	0.116 (0)	0 (0)	0.041 (1)	0.136 (1)	2

corpus available at that time, the ENC 2019. All calculations done in the present study are based on the ENC 2021 corpus; also, the adjectival ranges have been checked on ENC 2021.

[12] In the testing process of this study, we used the representative sample (N = 100) of prototypical adjectives and two control groups of participles tested in our previous studies (Paulsen et al., 2022; Vainik et al., 2023); as control groups also functioned six samples of word groups representing lexical categories overlapping with the adjective, used in Tuulik et al. (2022).

Word	ATTR	ATTR/ST	ADV	PRED	Conforming patterns
hiir ‘mouse’	0.237 (0)	0.015 (1)	0.007 (0)	0.03 (0)	1
oskama ‘can, know’	0.11 (0)	0.003 (0)	0.005 (0)	0.017 (0)	0

Table 2: Examples of conformity assessment analysis

3. Creating the calculator

3.1 The prerequisites of the ASC

There are basically four main requirements for creating an ASC application:

1) knowledge of the normal variation within the patterns of adjectival behaviour;

2) an established scale of adjectivity;

3) the availability of a morphologically annotated corpus for retrieving the frequency data of patterns and lemmas;

4) a script communicating with the corpus and retrieving statistics on the occurrences of the input word in the selected patterns, as well as calculating conformity assessment results.

The first requirement, the ranges of normal adjectival variation for each selected corpus pattern, were presented in Section 2.2 (Table 1). Conformity assessment results in each corpus pattern are the basis for evaluating a word’s closeness to adjectival behaviour. The counts corresponding to the criteria allow us to establish a scale of similarity to the adjectival corpus profile, which brings us to the second requirement of our calculator. The values matching the ranges of adjectivity vary over five degrees, presented in Table 3 (the function of the colours is to facilitate the perception of the values; these colours are also used on the display of the ASC):

Values	Scale
4	very likely
3	likely
2	ambiguous
1	unlikely
0	very unlikely

Table 3: The scale of adjectivity

The third requirement of the ASC is its data source: the ENC 2021 corpus, currently the newest and largest corpus of the Estonian language, with 2.4 billion words (Koppel & Kallas, 2022b). The ENC corpora (Koppel & Kallas, 2022a) are stored in the corpus query system Sketch Engine (Kilgarriff et al., 2004; Kilgarriff et al., 2014). ENC 2021 is pre-tagged, lemmatised, and disambiguated with the estNLTK 1.6.9 program (Laur et al, 2020). This corpus contains eleven sub-corpora[13].

The fourth requirement of the ASC, a script retrieving the frequency data from ENC 2021 and linking the Sketch Engine system to the application, is described in the next subsection.

3.2 The algorithm

The algorithm[14] we used for evaluating the adjectivity of a given word utilises statistics queried via the corpus query system Sketch Engine's API[15]. First, we will provide an overview of the statistics queried and their query patterns.

To retrieve the necessary frequencies, we queried the Sketch Engine API using a specific set of query patterns. These patterns correspond to various occurrences of the input word in a given text corpus (in our case, ENC 2021). Table 4 displays the query patterns and the corresponding frequencies obtained through the Sketch Engine API.

Identification	**Definition**	**Query**
lemma_freq	overall frequency of the input word (lemma)	`[lemma = "lemma"]`
lemma_S_freq	the frequency of an input word followed by a noun	`[lemma = "lemma"] [tag = "S.* "]`
s_lemma_S_freq	the frequency of an input word when it is at the beginning of a sentence and followed by a noun	`<s>[lemma = "lemma"] [tag = "S.* "]`
Dlist_lemma_freq	the frequency of the input word if it is preceded by one of the predefined adverbs	`([lemma = "adv1"]|[lemma="adv2"] | …)[lemma="lemma"]`

[13] Web 2013, Web 2017, Web 2019, Web 2021, Feeds 2014–2021, Wikipedia 2021, Wikipedia Talk 2017, the Open Access Journals (DOAJ), Literature, the Balanced Corpus and the Reference Corpus.

[14] The code is available at https://github.com/PRG1978/A-multi-purpose-lexicographic-resource.

[15] About the communication with the Sketch Engine via automated HTTP requests, see more at https://www.sketchengine.eu/documentation/api-documentation/.

Identification	Definition	Query

be_DlistQ_lemma_freq	the frequency of the input word which may be preceded by one of the predefined adverbs preceded by "be" given as the base form	`[lemma = "be"]`
		`([lemma="adv1"]|[lemma="adv2`
		`"] | …)?[lemma="lemma"]`

Table 4: Frequency identification and query patterns

The queried statistics include the total frequency of the input word (lemma) and the frequency of the input word as part of a sequence (column "query" in Table 4). The first column shows five identifiers corresponding to the five frequencies obtained from the query pattern in the third column. The "lemma" in quotation marks represents the input word, while the query fragment "`([lemma="adv1"]|[lemma="adv2"] | …)`" represents the inclusive list of over 66 adverbs (see Appendix 1). Exceptions in the ASC queries are non-inflected past participles with the endings *-dud*, *-tud* and *-nud*; for those forms, only text words are considered, not lemmas. It is important to note that all data processing is based on the frequencies of the actual occurrences of different PoS-interpretations in the corpus; the PoS of the target words is not predefined.

To estimate the adjectivity of a given testing word, we normalise the frequencies obtained from Table 5 using formulas (1) to (4):

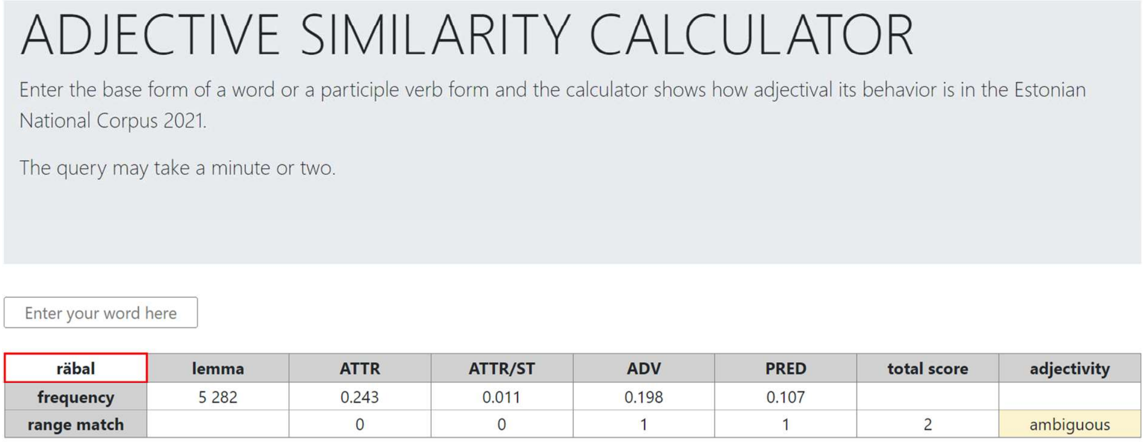
lemma_S_norm_freq = lemma_S_freq / lemma_freq (1)
s_lemma_S_norm_freq = s_lemma_S_freq / lemma_freq (2)
Dlist_lemma_norm_freq = Dlist_lemma_freq / lemma_freq (3)
be_DlistQ_lemma_norm_freq = be_DlistQ_lemma_freq / lemma_freq (4)

These formulas involve dividing the second to fourth frequencies by the overall frequency of the test word (first row of Table 4). The resulting normalised frequencies are then checked against a set of predefined ranges of adjectival behaviour (see Table 1 in Section 2.2), following the steps of conformity analysis. After that, the corresponding adjectivity rate from the scale of values (as established in Table 3, Section 3.1) is found and displayed on the screen together with frequency data and numeral values in each pattern.

4. The calculator at work

The ASC works on the web address https://adjcalculator.pythonanywhere.com/. It can be opened in a separate window of a web browser while working in a dictionary writing system or checking corpus data via Sketch Engine platform. The application is supported by the most common browsers (Microsoft Edge, Mozilla Firefox, Chrome, Safari and Brave).

Figure 1 presents the user interface of the ASC. There is a search box below the title and further below are tabular fields for the results of a query. The user needs to press the “enter” button on the keyboard to start the query.



ADJECTIVE SIMILARITY CALCULATOR

Enter the base form of a word or a participle verb form and the calculator shows how adjectival its behavior is in the Estonian National Corpus 2021.

The query may take a minute or two.

räbal	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	5 282	0.243	0.011	0.198	0.107		
range match		0	0	1	1	2	ambiguous

Figure 1: The user interface of the ASC

The word entered in the application – *räbal* ‘rag; miserable, shabby’ – is an existing entry in CombiDic, marked with two PoS tags: as a noun, mostly used in the plural (*räbalad* ‘rags’), and as an adjective (*räbal meeleolu* ‘shabby mood’). The result of the calculator, the value “ambiguous”, reflects its twofold PoS affiliation. The outcome also shows that its use as an attribute is below the level of prototypical adjectives while the score in the adverb pattern and the role of the predicative match the criteria of typical adjectives. For closer examination of the actual corpus behaviour of this word, one can look at the concordances and/or Word Sketch tool in Sketch Engine.

4.1 Quantitative parameters

A single query by ASC took 3.6–88.4 seconds during the test period of the prototype. Because the ASC retrieves the frequency data via the Sketch Engine API (see Section 3.3), the speed of the ASC is dependent on the smoothness of queries by Sketch Engine. The query time may be shorter if a request has previously been processed.

4.2 Evaluation of the ASC and its results

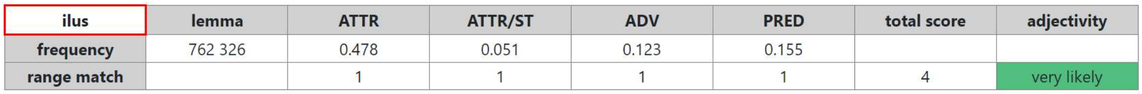
In this section, we test words from seven different lexical groups with different lexicographic status to demonstrate how the ASC works and to evaluate the results. The examples selected for analysis represent different subtypes of the main word classes and exemplify how the ASC works with both non-ambiguous and ambiguous cases regarding PoS categorisation. The categories examined are adjectives, nouns, verbs, adverbs, pronouns and numerals (both cardinals and ordinals). The participles, one of the most problematic areas in lexical categorisation, are not analysed in connection with verbs, but receive their own analysis in Section (4.2.4).

The words are checked for their status as a lexical entry in the CombiDic dictionary; the collocational analysis of the results is based on the Sketch Engine tool Word Sketch searching ENC 2021, the corpus the ASC also relies on. The usage examples come from ENC 2021, sometimes shortened to show the most relevant information.

4.2.1 Adjectives

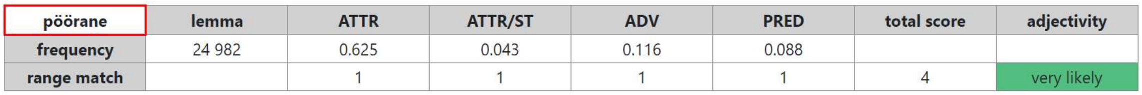
First, we test three adjectives that are headwords in the CombiDic, to see if they match the adjectival profile measured by the ASC. These are the root adjective *ilus* ’beautiful, pretty’, the derivative *pöörane* ’frantic, wild’, and the indeclinable adjective *eri* ’separate; different’. As the ASC results depicted in (2a–2c) show, all three adjectives achieve the highest results, scoring in all four patterns.

(2a) *ilus* ’beautiful, pretty’



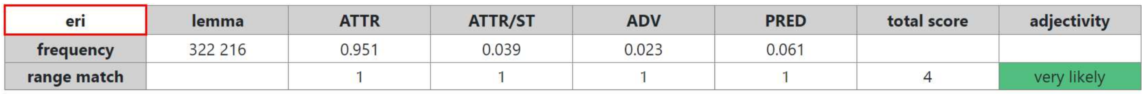
ilus	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	762 326	0.478	0.051	0.123	0.155		
range match		1	1	1	1	4	very likely

(2b) *pöörane* ’frantic, wild’



pöörane	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	24 982	0.625	0.043	0.116	0.088		
range match		1	1	1	1	4	very likely

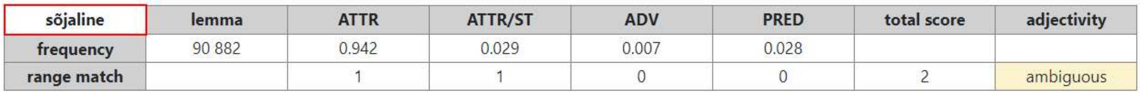
(2c) *eri* ’separate, different’



eri	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	322 216	0.951	0.039	0.023	0.061		
range match		1	1	1	1	4	very likely

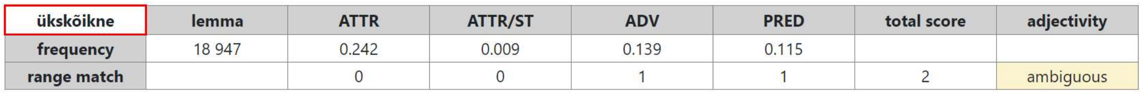
Let us now take a look at two adjectives – perfectly common and validated as adjectives in the CombiDic – categorised as ambiguous by the ASC. These adjectives are *ükskõikne* ‘indifferent’ and *sõjaline* ‘military’. The screenshots of the ASC analyses show the scores concentrating either to the left (2d) or the right side (2e) of the table. These results reflect a division of labour in behavioural profiles among adjectives: there are adjectives that are predominantly used as attributes and those prevalent in the predicative role (see Vainik et al., 2023). Such a differentiation is identified in other languages (for English, see Bolinger 1967, Lassiter 2015: 145) but, to our knowledge, has not yet been investigated in Estonian.

(2d) *sõjaline* ‘military’



sõjaline	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	90 882	0.942	0.029	0.007	0.028		
range match		1	1	0	0	2	ambiguous

(2e) *ükskõikne* ‘indifferent’

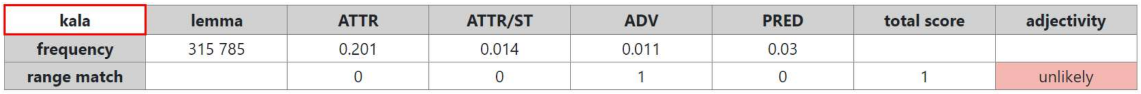


ükskõikne	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	18 947	0.242	0.009	0.139	0.115		
range match		0	0	1	1	2	ambiguous

4.2.2 Nouns

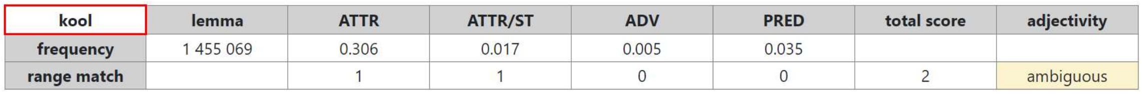
The examples of nouns tested for adjectival behaviour are the concrete noun *kala* ’fish’ (3a), the noun *kool* ’school’ (3b), with twofold semantic content denoting both a building and an institution, and the abstract noun *armastus* ’love’ (3c):

(3a) *kala* ’fish’



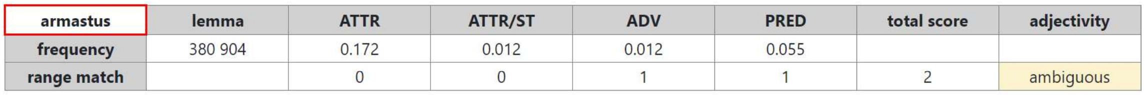
kala	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	315 785	0.201	0.014	0.011	0.03		
range match		0	0	1	0	1	unlikely

(3b) *kool* ’school’



kool	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	1 455 069	0.306	0.017	0.005	0.035		
range match		1	1	0	0	2	ambiguous

(3c) *armastus* ’love’



armastus	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	380 904	0.172	0.012	0.012	0.055		
range match		0	0	1	1	2	ambiguous

As expected, all three nouns show low results in the ASC; they also vary in the actual realisation of tested corpus patterns. The first of them, *kala*, receives the label “unlikely adjective”, with one matching pattern, the ADV. The most frequent adverbs preceding *kala* are the degree adverbs (*palju* ’a lot of’, *rohkem* ’more’, *peamiselt* ’mostly’), but also *lihtsalt* ’simply’, and *hoopis* ’instead; completely’. It is important to note that *rohkem* and *lihtsalt* are not included in the adverb list (cf. Appendix 1) because they are predominantly used as verb modifiers. The predicative pattern is possible but rather infrequent for *kala* (e.g. *hai on kala* ’a shark is a fish’).

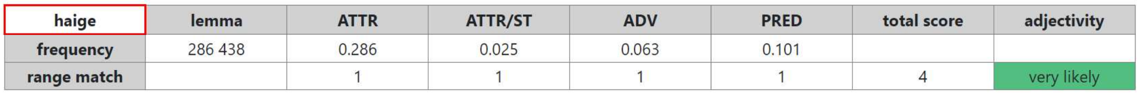
Why does the noun *kool* ‘school’ match the adjectives in the attributive patterns (ATTR and ATTR/ST)? The reason is the fact that, in Estonian, nouns can be used as genitive attributes, which is a frequent pattern for this word, as in the following collocations:

(3d) *kooli* *söökla* / *õpetaja* / *õpilane*
school.GEN canteen / teacher / pupil
'the canteen / teacher / pupil of the school'

The abstract word *armastus* 'love' shows relatively high results in adverb and predicative patterns. The ranges of the adverb pattern are relatively large, for instance, for manner adverbs (*lihtsalt* 'simply'); this noun is also modified by the adverbs included on our list of adjectival modifiers. Abstract nouns can be used predicatively as in *Jumal on armastus* 'God is love', and *elu on armastus* 'life is love'.

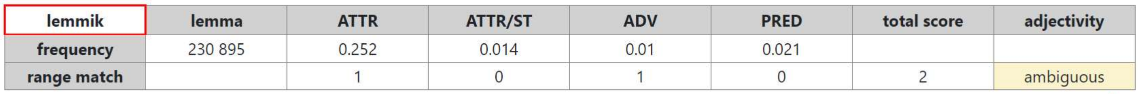
To test the ASC for more ambiguous cases of PoS manifestation, we examine the words *haige* 'sick; sick person' and *lemmik* 'favourite thing; favourite, dearest', both tagged as noun and adjective in CombiDic. These words represent productive patterns of nominalisation and adjectivisation: as a result of ellipsis, basically every adjective can employ the syntactic functions typical to nouns (i.e. occur as a subject, object or predicative), and some nouns can be used as modifiers (Vainik et al., 2021: 123). The example of nominalisation, *haige* (3e), is labelled "very likely adjective", corresponding to the adjective profile in every respect. The adjectivised noun *lemmik* (3f) matches only half of the patterns: apparently, this word still does not behave fully as an adjective.

(3e) *haige* 'sick, ill; sick person'



haige	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	286 438	0.286	0.025	0.063	0.101		
range match		1	1	1	1	4	very likely

(3f) *lemmik* 'favourite thing; favourite, dearest'

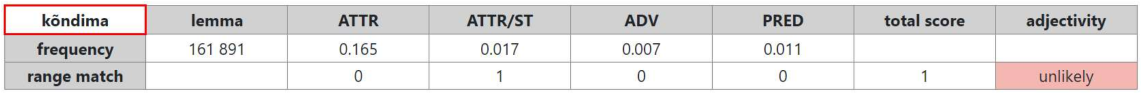


lemmik	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	230 895	0.252	0.014	0.01	0.021		
range match		1	0	1	0	2	ambiguous

4.2.3 Verbs

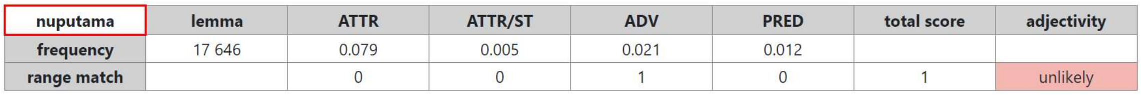
The verbs selected for illustration represent semantically different areas: the concrete motion verb *kõndima* 'walk' (4a) and two cognitive verbs, *nuputama* 'figure, contrive' (4b) and *mõtlema* 'think' (4c). The results show variation in corpus behaviour, even for the two cognitive verbs; the overall adjectivity assessments are very low ("unlikely adjective").

(4a) *kõndima* 'walk'



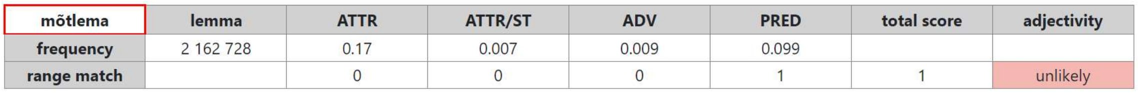
kõndima	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	161 891	0.165	0.017	0.007	0.011		
range match		0	1	0	0	1	unlikely

(4b) *nuputama* ’figure, contrive’



nuputama	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	17 646	0.079	0.005	0.021	0.012		
range match		0	0	1	0	1	unlikely

(4c) *mõtlema* ’think’



mõtlema	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	2 162 728	0.17	0.007	0.009	0.099		
range match		0	0	0	1	1	unlikely

The motion verb *kõndima* ‘walk’ shows one match with the adjectival profile (4a), in the pattern measuring precedence of a noun at the beginning of a sentence (ATTR/ST). Estonian is a pro-drop language; hence, the pronoun before a verb can be omitted and the sentence can start with the verb followed by an adverbial consisting of a noun in a semantic case form:

(4d) *Kõnnin* *auto-ni* / *tänava-le* / *sõbra-ga*
walk-3SG car-TERM / street-ADE / friend-COM
‘I walk to the car / to the street / with a friend’

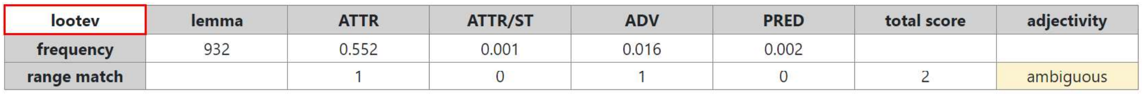
The two cognition verbs receive matches with the adjective profile, too, but in different patterns. The verb *nuputama* ‘figure, contrive’ (4b) often occurs after an adverb, which may coincide with adverbs typically modifying adjectives (e.g. the degree adverbs *natuke* ‘a little’, *palju* ‘a lot’ and *veidi* ‘a bit’). The verb *mõtlema* ‘think’ scores in the predicative pattern (4c) for the reason typical of verbs: the main aspect contravening the quality of the PRED-pattern is that the copula verb *olema* ‘be’ is also used as the auxiliary verb in present or past tense forms in connection with compound tempus. An example of *mõtlema* in a perfect tense is given in (4e).

(4e) *Ta* *on* *mõelnud* *töökoha* *vahetuse-le.*
He/she be-3SG think-PAST-PART job.GEN shift-ALL
‘He/she has been thinking about a job change.’

4.2.4 Participles

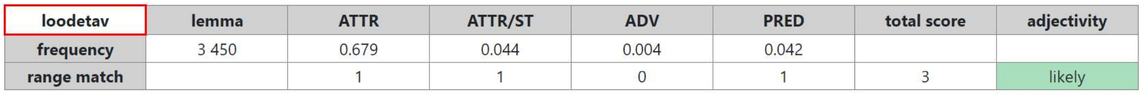
One of the target categories for the ASC analysis is participles, constituting a fuzzy area between verbs and adjectives. Here we analyse the present and past personal and impersonal forms of the verb *lootma* ‘hope, expect’ (see 5a–5d). None of these forms are headwords in CombiDic; however, two of them (*loodetav* (5b) and *loodetud* (5d)) receive quite high adjectivity assessments (“likely adjectives”). These results are to be expected, as the forms with higher scores in fact demonstrate both verbal and adjectival usage patterns in corpus data and the forms with lower results are exclusively used in verbal functions. Compared to the verbs analysed in the previous section, the ASC results show considerable variation.

(5a) *lootev* 'hoping'



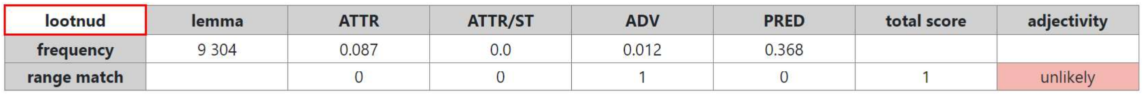
lootev	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	932	0.552	0.001	0.016	0.002		
range match		1	0	1	0	2	ambiguous

(5b) *loodetav* '(being) hoped, expected'



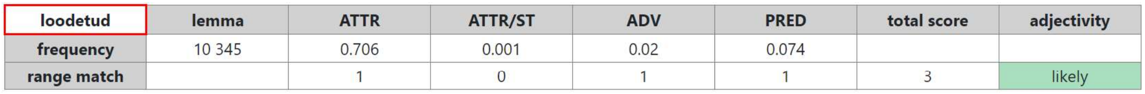
loodetav	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	3 450	0.679	0.044	0.004	0.042		
range match		1	1	0	1	3	likely

(5c) *lootnud* '(has) hoped, expected'



lootnud	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	9 304	0.087	0.0	0.012	0.368		
range match		0	0	1	0	1	unlikely

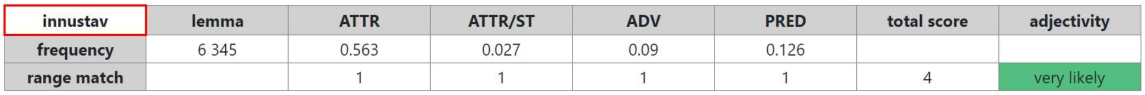
(5d) *loodetud* '(has been) hoped, expected'



loodetud	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	10 345	0.706	0.001	0.02	0.074		
range match		1	0	1	1	3	likely

Let us now analyse two examples of participles showing results from both extremes of the scale established in Table 3. A participle that might be considered a strong candidate for the status of the headword in the CombiDic is the present participle form *innustav* 'encouraging, inspiring'. This form does not yet have the status of a dictionary entry, but receives the highest value of adjectivity with the score "very likely adjective" (see 5e). The adjectival usage is also confirmed by the examples in the ENC 2021 corpus.

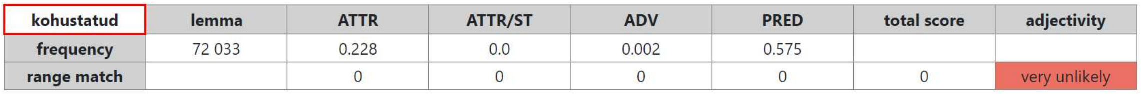
(5e) *innustav* 'encouraging, inspiring'



innustav	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	6 345	0.563	0.027	0.09	0.126		
range match		1	1	1	1	4	very likely

There are words or word forms with highly restricted usage, such as the participle *kohustatud* 'be obliged to', receiving 0 points in the ASC analysis. This past participle of an impersonal voice form is mainly used in the construction [X *on kohustatud* V$_{\text{inf}}$] 'X is obliged to V'. Therefore, we can see under-representation in all patterns except the predicative pattern *olema*_TW ('be'_TW), where this participle demonstrates clear overuse: the result of this pattern exceeds the adjectival ranges of 0.036–0.344, with a result of 0.575. This indicates that the upper limit of this range also functions well. The ASC analysis of *kohustatud* is presented in (5f):

(5f) *kohustatud* ‘obliged to’

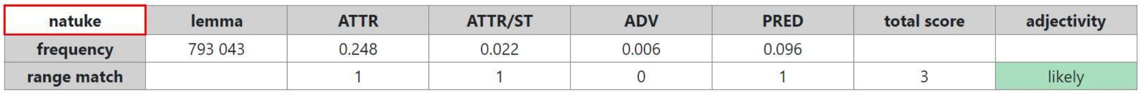


kohustatud	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	72 033	0.228	0.0	0.002	0.575		
range match		0	0	0	0	0	very unlikely

4.2.5 Adverbs

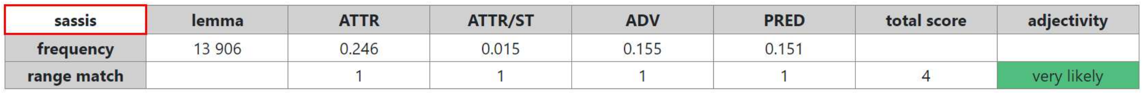
We have selected three words representing different types of adverbs: the degree adverb *natuke* ’a little, slightly’ (6a), the state adverb *sassis* ‘messy; confused’, indicating the physical or mental condition of the participant in an event (6b), and the sentence adverb *kindlasti* ‘certainly’ (6c).

(6a) *natuke* ’a little, slightly’



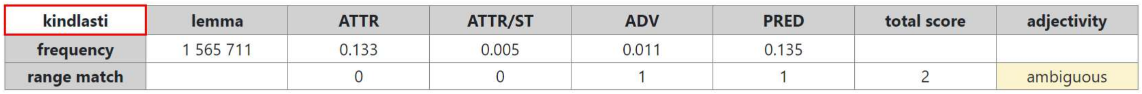
natuke	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	793 043	0.248	0.022	0.006	0.096		
range match		1	1	0	1	3	likely

(6b) *sassis* ’tangled, messy; confused’



sassis	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	13 906	0.246	0.015	0.155	0.151		
range match		1	1	1	1	4	very likely

(6c) *kindlasti* ‘for sure, certainly’



kindlasti	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	1 565 711	0.133	0.005	0.011	0.135		
range match		0	0	1	1	2	ambiguous

The first two adverbs score quite high in the ASC, labelled as “likely adjective”, whereas *kindlasti* ‘certainly’ is rated as “ambiguous”. The degree adverb *natuke* ‘a little’, as expected, conforms to the adjectival behaviour in both attribute patterns (in such collocations as *natuke aega/nalja* ‘a little bit of time/fun’) but is not modified by an adverb itself. As an intensifier, it precedes predicatives and thus occurs after the verb *olema* ‘be’ (*Uudis on natuke enneaegne* ‘The news is slightly premature’).

The fact that the state adverb *sassis* ‘messy; confused’ receives the highest rating, “very likely adjective” is quite predictable, as it belongs to a type of adverbs functionally overlapping with adjectives[16]. It is also frequently modified by the intensifying adverbs

[16] The adverbs belonging to this type can also be analysed as (locative) case forms of nouns, e.g. *lokki-s* ’curly’ [curl-INE] and, as in this example, a base noun (*lokk* ’curl’) may be detectable. The static locative semantics (inessive and adessive cases) lead to the adjective interpretation; the directional (illative/elative; alla-tive/ablative) forms of the same words (e.g. *lokk-i* ’into a curly state’ [curl-ILL]) are read as either an adverb or as the respective case forms of nouns but not an adjective (See more in Vainik et al., 2021: 124).

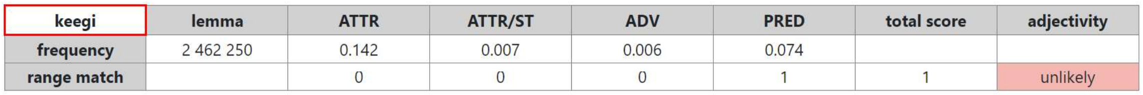
included on the list of adverbs modifying adjectives (*täiesti* / *lootusetult* / *veidi sassis* ‘completely / hopelessly / a bit messy’).

The fact that the sentence adverb *kindlasti* ‘certainly’ only receives two points is not a surprise, as this word, particularly at the beginning of a sentence, affects word order by subject-predicate inversion and is typically followed by the predicate of the sentence (see Lindström 14–15).

1.1.1 Pronouns

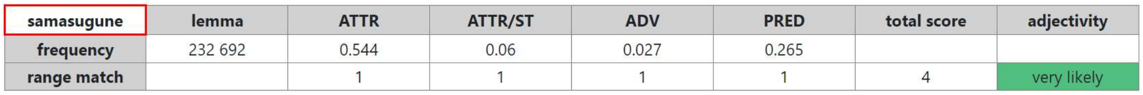
Estonian pronouns function similarly to nouns, adjectives or numerals (Erelt 2017: 59). Let us test the indefinite pronoun *keegi* ‘someone’ (7a), the compound demonstrative pro-adjective *samasugune* ‘(the) same’ (7b), and the pro-numeral *tosin* ‘dozen’ (7c).

(7a) *keegi* ‘someone’



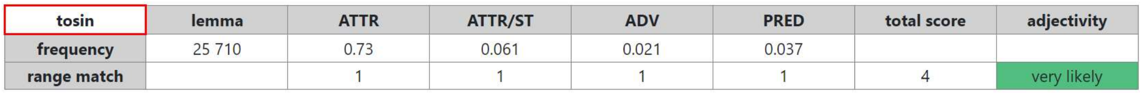
keegi	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	2 462 250	0.142	0.007	0.006	0.074		
range match		0	0	0	1	1	unlikely

(7b) *samasugune* ‘same’



samasugune	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	232 692	0.544	0.06	0.027	0.265		
range match		1	1	1	1	4	very likely

(7c) *tosin* ‘dozen’



tosin	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	25 710	0.73	0.061	0.021	0.037		
range match		1	1	1	1	4	very likely

The results correspond quite well with the word class the respective pronoun replaces. The pronoun *keegi* receives only one point in ASC and the label “unlikely adjective”, matching only the predicative pattern (see 7a). The proadjective *samasugune*[17] (7b) behaves as a true adjective and scores on the highest level (“very likely adjective”). Surprisingly, at least at first sight, the pronumeral *tosin* also receives the maximum score in ASC (see 7c). The usage patterns typical of an Estonian quantifier phrase explain the phenomenon: in the nominative case the quantifier governs its nominal complements by assigning to them the partitive case (*kaks õun-a* [two apple-PART]; see e.g., Erelt 2009: 19). This pattern explains the high score in the attribute pattern of *tosin*; this quantifier is often followed by a noun in partitive case (*tosin kilo*/*päeva*/*õuna* ‘dozen kilo/days/apples’). It is also modifiable by degree adverbs (*vähemalt tosin* ‘at least a dozen’, *peaaegu tosin* ‘almost a dozen’) and is used predicatively. All of these patterns contribute to the high outcome and explain inter

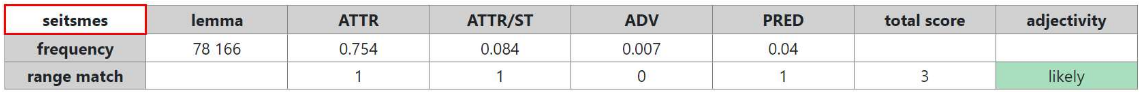
[17] The proadjectives are marked as adjectives in CombiDic.

alia why the cardinal numerals generally meet all the requirements of adjectivity set by ASC (cf. example 8c in next section).

4.2.6 Ordinals and cardinals

The Estonian ordinals are basically considered to function as adjectives (Erelt 2017: 63). In the ASC, the ordinal *seitsmes* ‘seventh’ receives the assessment “likely adjective” with three points (see 8a). The result reveals the one condition in which Estonian ordinals do not behave as adjectives: the adverb pattern.

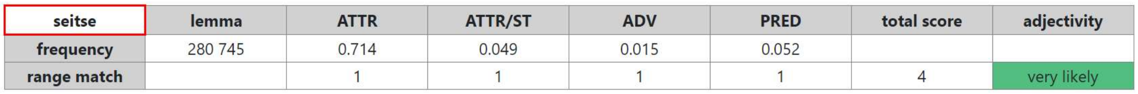
(8a) *seitsmes* ’seventh’



seitsmes	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	78 166	0.754	0.084	0.007	0.04		
range match		1	1	0	1	3	likely

Interestingly, the ordinals do not score as high as cardinals, a category assumed to belong to the quantifier class. An example of a cardinal is given in (8b); the explanation given for the pronumeral *tosin* ‘dozen’ in the previous section also applies here.

(8b) *seitse* ‘seven’



seitse	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	280 745	0.714	0.049	0.015	0.052		
range match		1	1	1	1	4	very likely

4.3 Discussion of the results

The problem to be solved by the assistance of the ASC is whether to label a particular word or word form in a dictionary as an adjective or not. The quality of the ASC can be estimated by assessing its output of both non-ambiguous and ambiguous representatives of the word classes overlapping with adjectives. Another guiding line is formed by the decisions made by lexicographers so far, as well as a closer examination of the corpus behaviour of the tested words.

Overall, the ASC results indicate that the application works as intended: the rates of the words that are clearly non-adjectival fall into the lower interval in the similarity assessments (from 0 to 2; “very unlikely”, “likely”, or “ambiguous” regarding adjectival behaviour) and the ratings of cases that can be expected to behave to some extent adjectively fall into the upper interval (3–4, with the corresponding rates “likely” and “very likely”). When it comes to the analysis of validated adjectives themselves, we can conclude that almost all tested words received the rating “very likely”, with the highest score of 4.

Exceptions prove the rule, and this is also the case with the ASC. As our previous studies of adjectival behaviour have indicated, at least some of the (perfectly common) Estonian adjectives seem to prefer either attributive or predicative constructions. This may be the reason why some quite “normal“ adjectives receive only average or even lower scores in the ASC (see Section 4.2.1). The existence and extent of this phenomenon needs closer examination, which is something the ASC can be used as a tool for.

Another factor interfering with the results are constructions typical to other classes than adjectives but (partly) overlapping with the patterns constituting the adjective profile. One question is: how can we rule out genitive attributes, the typical pattern of nouns modifying other nouns? A solution would be to work out some restrictive conditions. However, as the ASC analysis of the example noun *kool* ‘school’ (3b) showed, a noun frequently used in the attributive function still does not receive a summary value high enough to conform to the adjective profile. This outcome can even be seen as a positive aspect – the ASC allows one to study a noun's tendency to function as a genitive attribute.

An additional issue is the interference of other than predicative constructions around the copula verb *olema* ‘be’. There are different construction families clustering around *olema* in Estonian: compound tenses, existential clauses and possessive clauses. Manual checking of the corpus data regarding the words tested in this study has shown that the occurrences still mostly involve predicative clauses.

The inclusion of pronouns and numerals was mostly motivated by idle interest, as this closed class practically does not pose problems of categorisation. Still, the results of the ASC analysis were interesting, for instance, regarding the different behaviours of cardinals and ordinals: strikingly, the ordinals, regarded as adjectives, did not score as highly as the cardinals. Hence, it is surprising that the cardinals outscore ordinals in conforming adjectives: one would have expected that the meaning of a cardinal is not modifiable by scaling adverbs. This tells us, possibly, something about the practical fuzziness of the meanings. There is evidently a need for further studies in this area.

We are aware that the frequency results of the ASC directly depend on the quality of the tagging system, and we recognise that tagging and disambiguation errors affect the analysis. For instance, the morphoanalyser struggles with the form homonymy cases (e.g. *armutud* can be analysed as the nominative plural form of the adjective *armutu* ‘merciless’ or as the past participle impersonal form of the verb *armuma* ‘fall in love’). At any rate, the experienced lexicographer will discover the abnormalities and can check the results in the corpus to avoid problems.

The analysis in this study is solely based on morphosyntactic patterns, but adjectivity also undoubtedly has a distinctive semantic dimension. A direction for future studies could be the inclusion of semantic aspects in the adjectivity assessment battery. In addition, the semantic effect on the attributive-predicative prevalence noted in Section (4.2.1) is an interesting topic to explore further.

5. Conclusions

The ASC is a web-based application accessible to everyone. It takes a word whose similarity to adjectival behaviour is to be measured as input from the user and retrieves corpus data (the frequencies of the word form in requested positions – corpus patterns – and the total frequency of lemma). The tool calculates the relative salience of the instances of patterns and compares the values to the ranges of adjectival behaviour (cf. section 2.2). The ASC provides the outcome both in terms of numerical measures and verbal labels (as described in section 3.1). The calculator can be used to explore the syntactic behaviour of any word.

The constituency of the set of automatically searchable corpus patterns was tested to find the optimal solution, and the thresholds of adjectival behaviour determined on the basis of the results were adjusted. Decisions about previously tried methods for calculating the distance of a word from the adjectival profile (see Tuulik et al., 2022, Paulsen et al., 2022, and Vainik et al., 2023) were made. The ASC described in this study is the prototype of the application; the development process is still ongoing. Consultations with lexicographers who will test the ASC in actual use will be an important part of the further application design.

This study proved that corpus data can be used to establish the prototypical behaviour of a word class by creating a corpus profile of the central close-context patterns characteristic to the category. At least the adjective profile was confirmed to be operational as a template for comparing particular words or word forms. The study also showed that the patterns constituting the profile work in combination: no pattern alone can be used as proof of adjectivity.

6. Abbreviations

Glossing: ADE – adessive case; ALL – allative case; COM – comitative case; GEN – genitive case; INE – inessive case; NOM – nominative case; PART – partitive case; PAST – past tense; PL – plural; SG – singular; TER – terminative case; TRA – translative case.

7. Acknowledgements

This research was supported by the Estonian Research Council grant PRG1978. We thank the anonymous reviewers for their helpful suggestions.

8. References

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

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

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. Available at https://sonaveeb.ee.

Erelt, M. (2017). 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.

Erelt, M. (2009). Typological overview of Estonian syntax. *STUF – Language Typology and Universals*, 62.

Fábregas, A. & Marín, R. (2017). Problems and questions in derived adjectives. *Word Structure*, 10 (1), pp. 1–26.

Geeraerts, Dirk (1989). Prospects and problems of prototype theory. *Linguistics* 27, pp. 587—612.

Haugen, T. A. (2013). Adjectival valency as valency constructions: Evidence from Norwegian. *Constructions and Frames*, 5 (1), pp. 35–68. DOI: https://doi.org/10.1075/cf.5.1.02hau

Helbig, G. (1992). *Probleme der Valenz- und Kasustheorie* [Problems in Valency and Case Theory]. Tübingen: Niemeyer.

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.

Kilgarriff, A. & Rychlý, P. (2008). Finding the words which are most X. In: *Proceedings of the XIII EURALEX International Congress* (Barcelona, 15-19 July 2008). 2008. p. 433-436.

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.

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, pp. 7–36.

Koppel, K. & Kallas, J. (2022a). Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu [Estonian National Corpus 2013–2021: the largest collection of Estonian language data.]. *Estonian Papers in Applied Linguistics*, 18, pp. 207–228. http://dx.doi.org/10.5128/ERYa18.12.

Koppel, K. & Kallas, J. (2022b). *Eesti keele ühendkorpus 2021* [Estonian National Corpus 2021]. https://doi.org/10.15155/3-00-0000-0000-0000-08E60L.

Langemets, M. (2010). *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras* [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources]. PhD thesis. Tallinn: Eesti Keele Sihtasutus.

Lassiter, D. (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, 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). Available at: https://aclanthology.org/2020.lrec-1.0.pdf.

Lindström, L. (2005). *Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles* [The position of the finite verb in a clause: word order and the factors affecting it in spoken Estonian]. PhD thesis. Tartu Ülikooli kirjastus, Tartu.

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. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_18.pdf.

Paulsen, G., Tuulik, M., Lohk, A. & Vainik, E. (2022a). From verbal to adjectival. Evaluating the lexicalisation of participles in an Estonian corpus. *Slovenščina 2.0*, 10(1), pp. 65–97.

Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2022b). The morphosyntactic profile of prototypical adjectives in Estonian. Presentation held at the XX EURALEX International Congress 12–16 July 2022 in Mannheim, Germany.

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, pp. 279—302. http://dx.doi.org/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. Available at: https://euralex.org/publications/a-typology-of-lexical-ambiforms-in-estonian/

Vainik, E., Paulsen, G., Tuulik, M. & Lohk, A. (2023). Towards the Morphosyntactic Corpus Profile of Prototypical Adjectives in Estonian. *Estonian Papers in Applied Linguistics*, 19, pp. 225–244. DOI: http://dx.doi.org/10.5128/ERYa19.13

Vare, S. (2006). Adjektiivide substantivatsioonist ühe tähendusrühma näitel. [On substantivisation of adjectives: Analysing a semantic group] E. Niit. *Keele ehe.* Tartu: Tartu Ülikool. Tartu Ülikooli eesti keele õppetooli toimetised; 30, pp. 205—222.

Viht, A. & Habicht, K. (2019). Eesti keele sõnamuutmine [The Estonian inflection]. Eesti keele varamu IV. Tartu: Tartu University Press.

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.

Appendix 1. Inclusive list of adverbs used as a filter while searching for the ADV pattern (with English translations)

väga	'very, highly'
üsna	'quite, fairly'
päris	'quite, right'
piisavalt	'enough, sufficiently'
niivõrd	'so, insofar as'
suhteliselt	'relatively, comparatively'
üpris	'very, much, greatly'
võimalikult	'possibly, as possible'
suht	'relatively' (colloquial)
liiga	'too, excessively'
äärmiselt	'extremely, utterly'
küllaltki	'rather, fairly'
täiesti	'entirely, wholly'
erakordselt	'outstandingly, exceedingly'
võrdlemisi	'comparatively, relatively'
täitsa	'completely, quite'
tõeliselt	'positively, truly'
küllalt	'sufficiently, enough'
ülimalt	'infinitely, immeasurably'
sedavõrd	'inasmuch, so'
liialt	'excessively'
endiselt	'as before, still'
üllatavalt	'surprisingly, amazingly'
üksnes	'merely, only'
igati	'to the outmost, in every way'
palju	'much, a lot of, many'
vähem	'less, fewer'
ääretult	'boundlessly, infinitely'
väga-väga	'very, greatly, highy'
vähemalt	'at least, at any rate'
kuivõrd	'insofar as'
peamiselt	'chiefly, principally'
enam-vähem	'more or less'
tohutult	'infinitely, vastly'
uskumatult	'incredibly, unbelievably'
niigi	'already, as it is'
hästi	'very, greatly'

peaaegu	‘almost, nearly’
hoopis	‘instead, entirely’
hirmus	‘very, greatly’
mõnusalt	‘pleasurably’
enamvähem	‘more or less’
suuresti	‘greatly, largely, highly’
erinevalt	‘variously, unlike, differently’
kaugeltki	‘by far’
natuke	‘a little’
kindlasti	‘for sure, certainly’
niisama	‘just so; for nothing’
iseenesest	‘unintentionally, by itself’
jätkuvalt	‘continually’
valdavalt	‘predominantly’
kahtlemata	‘undoubtedly, definitely’
eeskätt	‘primarily, mainly’
absoluutselt	‘absolutely’
tõenäoliselt	‘probably, likely’
meeletult	‘deliriously, wildly’
tõepoolest	‘indeed, actually’
kaunis	‘pretty’
täielikult	‘completely’
eriliselt	‘specially, particularly’
iseäranis	‘particularly, exclusively’
pisut	‘a little, slightly’
ülemäära	‘excessively’
parajalt	‘moderately’
veidi	‘a bit’
mõnevõrra	‘somewhat’