UDK 811.163.42'367.623
811.163,42'42
Izvorni znanstveni članak
Prihvaćeno za tisak: 20. svibnja 2024.
https://doi.org/10.22210/suvlin.2024.097.04

Ivan Lacić
Department of Classical Philology and Italian Studies
Alma Mater Studiorum – University of Bologna
ivan.lacic2@unibo.it

An insight into the Croatian degree modifier paradigm and its clustering profiles

Degree modifiers represent linguistic items employed to alter other elements in relation to their degree. Despite being a well–studied category in English linguistics, degree modifiers in Croatian have received limited attention. This study aims to address this gap by examining a set of Croatian degree modifiers as a part of <degree modifier + adjective> construction. Initially, a corpus analysis is used, and the 29 most frequent degree modifiers of adjectives in the *hrWaC* corpus are identified. To analyse the examined modifier, we turn to the distributional hypothesis and examine collocational contexts in which modifiers occur. By employing a simple collexeme analysis, we quantify the degree of attraction between a given degree modifier and adjective for each <degree modifier + adjective> construction and its 1000 most frequent adjectival collocates. The results of simple collexeme analysis then serve as input for hierarchical agglomerative cluster analysis, shedding light on the clustering patterns of Croatian degree modifiers based on their favoured collexemes. Simple collexeme analysis reveals itself as successful in filtering out collexems that consistently appear irrespective of the context, proving its superiority over methods relying solely on raw frequencies. The subsequent cluster analysis exposes some discrepancies between the modifiers' function and their cluster profiling, resulting in clusters lacking functional homogeneity. Nonetheless, certain subclusters demonstrate perfect or almost perfect stability and empirical support, affirming the (near–)synonymy among involved modifiers.

1. Introduction

Degree modifiers constitute linguistic items employed to modify other elements in terms of degree (Bolinger 1972; Quirk et al. 1985). They are typically adverbs, as in *very cheap*, *extremely spicy*, *quite intelligent*, *somewhat interesting*.

While the exploration of English degree modifiers (Bolinger 1972; Paradis 1997; Ito and Tagliamonte 2003; Xiao and Tao 2007) and more recent investigations into German counterparts (Claudi 2006; Stratton 2020) has yielded numer-

ous empirical studies, the realm of degree modifiers in the Croatian language remains largely unexplored. Croatian degree modifiers have not undergone extensive scrutiny, and the existing analyses usually adopt a cross–linguistic perspective. For instance, Pavić Pintarić and Frleta (2014) delved into the typology of "upwards" intensifiers across English, German, and Croatian, employing a limited parallel corpus from Harry Potter novels. Similarly, Batinić, Kresić and Pavić Pintarić (2015) explored the intensifying function of German modal particles and their equivalent modal expressions in Croatian and English. Their investigation sought to ascertain whether these particles could convey varying degrees of intensity and types of intensification. In a distinct vein, Matešić and Memišević (2016) centred their focus on evaluative adjectives and their accompanying modifiers in scientific texts spanning different domains (linguistics and medicine) in both Croatian and English. Nigoević and Galić (2020) analysed contrastively the strategies of intensification in Brešan's play *Predstava Hamleta u selu Mrduša Donja* and its translation in Istro–Venetian. Furthermore, Nigoević (2020) undertook a comprehensive comparative study of intensification in Croatian and Italian, showcasing the primary linguistic tools for intensification in both languages using examples drawn from diverse corpora. Finally, Vidaković Erdeljić (2023) studied the modifier *totalno* 'totally' in a corpus of Croatian tweets from the perspective of language contact between Croatian and English. In addition to the previously mentioned references, *Grammar of the Croatian Language* (Silić and Pranjković 2007) also addresses degree modifiers, referred to by the authors as "comparing (grading) particles". Silić and Pranjković (2007) note that these elements are commonly classified as adverbs and encompass words like *vrlo* 'very', *puno* 'very', *potpuno* 'completely', *skroz* 'totally' and so on, which affect qualities expressed by adverbs or adjectives in a way that entails grading and comparison with other attributes. Among the mentioned studies, Pavić Pintarić and Frleta (2014) stands out as particularly relevant. The authors delve into the collocational patterns of various Croatian degree modifiers, identifying which modifier class predominantly attracts which semantic class of adjectives. However, the limited corpus size (only 33 intensifiers) makes it prudent to reconsider the findings.

Although the referenced studies have provided valuable insights into Croatian degree modifiers, a more comprehensive examination, particularly from a monolingual perspective, remains necessary. Particularly noteworthy is the relatively modest attention given to the exploration of collocational pairings between Croatian degree modifiers and adjectives. In addition, studies employing analytical statistics and multifactorial methods to address this inquiry remain, to the best of the author's knowledge, yet to be undertaken.

This study aims to fill the aforementioned research gaps by examining the collocational behaviour and clustering profiles of 29 Croatian degree modifiers as a part of <degree modifier + adjective> construction in the *hrWaC* corpus. While degree modifiers possess the capacity to modify various word classes, the scope of the

present inquiry is confined to the contexts in which they modify adjectives, the category whose semantics lends itself best to gradation (Dressler and Merlini Barbaresi 1994). To analyse Croatian degree modifiers, we will adopt the distributional hypothesis and extend the premise that a "difference of meaning correlates with difference of distribution" (Harris 1970: 785). This foundational concept posits that a link between distributional similarity and meaning similarity allows us to deduce the latter from the former. In this context, meaning similarity is viewed as a functional similarity. Drawing inspiration from Desagulier's (2014) methodological proposal and relying on the Behavioural Profile approach (Divjak and Gries 2006), this study seeks to delve into nuances among degree modifiers by scrutinising their collocational patterns, which can offer insights into their distinct semantics and, consequently, their roles. The frequency of the collocational patterns of Croatian degree modifiers will serve as input for hierarchical agglomerative cluster analysis. Cluster analysis involves segmenting data into groups, commonly referred to as clusters, that exhibit practical utility, intrinsic significance, or a combination of both (Tan et al. 2018). When aiming to unveil meaningful groupings inside the broader paradigm of degree modifiers, these clusters should faithfully encapsulate the inherent structure within the data. In the realm of comprehending data, clusters emerge as potential degree modifier classes.

The objectives of this paper encompass three key aspects. Firstly, we aim to examine the most frequent degree modifiers in the Croatian language. Departing from an approach that avoids pre–established lists, we rely on corpus data. Secondly, we seek to investigate the collocational preferences of these modifiers. By delving into both similarities and distinctions in the collocational profiles of degree modifiers, we intend to derive insights into their functional differences. Thirdly, we propose to assess the viability of utilising an exploratory technique such as hierarchical agglomerative cluster analysis. We aim to ascertain whether this method can discern functional paradigms of Croatian degree modifiers and whether the resulting clusters (paradigms) exhibit homogeneity or variability. The working hypothesis postulates that an overlap in collocation preferences among modifiers would suggest not only a shared semantic content but also a functional one, classifying them as members of the same (sub)cluster. As a result, it should be possible to obtain homogenous functional paradigms of Croatian degree modifiers and further analyse them to inspect their content.

The paper is structured as follows. Section 2 provides a concise overview of fundamental concepts on degree modifiers of adjectives. Section 3 outlines the methodology employed and introduces the selected corpus. Section 4 presents the obtained results. Finally, in Section 5, we engage in a discussion and interpretation of the key findings.

2. Degree modifiers of adjectives

As previously highlighted, degree modifiers, at times referred to as intensifiers, constitute a specialised subset of degree words that furnish degree specifications relevant to the words they modify (Bolinger 1972; Quirk et al. 1985; Paradis 1997). The primary role of degree modifiers is to modify adjectival constituents (Quirk et al. 1985; Paradis 1997; Klein 1998). Since the adjective modifiers addressed in this study are not only those that fittingly bear the term intensifier in its literal sense as they genuinely amplify the meaning of the adjective but also elements that convey the exact opposite of intensification, such as moderating or diminishing, this paper is going to adopt the term degree modifier (DM).

DMs differ in terms of the specific value they attribute to adjectives. After considering various classifications (Quirk et al. 1985; Allerton 1987; Paradis 1997), Paradis' (1997) was adopted as a reference model and DMs were categorised into two distinct groups: reinforcers and attenuators. The reinforcer category encompasses maximizers (e.g. *potpuno* 'completely', *sasvim* 'totally') and boosters (e.g. *vrlo* 'very', *užasno* 'terribly'), while approximators (e.g. *skoro* 'nearly', *gotovo* 'almost'), moderators (e.g. *poprilično* 'rather'), and diminishers (e.g. *malo* 'a bit', *blago* 'slightly') belong to the attenuator class.

Besides conveying a degree, one of the defining attributes of DMs rests in their pragmatic value, i.e. emotive potency, as they display active involvement and attitude, thus enhancing the emotional and subjective dimension of the conversation (Paradis 1997; Scheffler, Richter and van Hout 2023).

Regarding their grammatical status, adjective DMs do not form a confined group of elements nor an open–class category. Still, they occupy a realm of flux, characterised by a constant influx of new members and the obsolescence of older ones (Claudi 2006).

The conduct of adjective DMs is shaped by the scale configuration of the adjectives under their influence, namely by the dimensions of totality and scalarity (Paradis 1997). It is widely recognised that adjectives employ varying scale structures, which may or may not encompass a maximum and/or minimum value (Rotstein and Winter 2004; Kennedy and McNally 2005; Kennedy 2007). Consequently, a classification of adjectives can be established grounded in their distinctive scale arrangements. Kennedy and McNally (2005) and Kennedy (2007) argue four types of scales: fully open ( ), lower–closed [ ), upper–closed ( ], and fully closed [ ]:

i. Open scale adjectives (0,1): *nizak* 'short', *jeftin* 'cheap'
ii. Lower closed scale adjectives [0,1): *prljav* 'dirty', *mokar* 'wet'
iii. Upper closed scale adjectives (0,1]: *čist* 'clean', *suh* 'dry'
iv. Totally closed scale adjectives [0,1]: *zatvoren* 'closed', *prazan* 'empty'

The open scale has no minimal and maximal elements (endpoints), while the lower closed scale has only the minimal element. On the contrary, the totally closed

scale possesses minimal and maximal elements, while the upper closed scale only has maximal elements. These scales are, hence, manifested as either bounded or unbounded, and DMs play a role in shifting or anchoring the property of entities within bounded scales or stabilising them across unbounded scales.

Paradis' theory on the bidirectionality of semantic influence (1997), along with the Harmonious Configuration Hypothesis (2008), anticipates a harmonious alignment between intensifiers and the meaning structures they modify. Specifically, the nature of the adjective in a <degree modifier + adjective> pairing dictates the type of DM that can modify it. At the same time, the character of the DM influences the selection and interpretation of a compatible adjectival component. Hence, there is an apparent necessity for a certain coherence between DMs and adjectives, with limitations stemming from the distinct schematic domains of the combined elements (Cacchiani 2017). The presence of constraints becomes evident in phrases where the degree modifier and the adjective are construed based on distinct types of gradability, such as in the examples of **jako živ* 'very alive' and **potpuno visok* 'completely tall'. In theory, amplifying *scalar* modifiers, viz., boosters (such as *vrlo* 'very') alter unbounded scalar antonyms, which project along a 'more–or–less' scale (e.g. *vrlo skup* 'very expensive'). Among the downgrading *scalar* modifiers are moderators, which approximate the middle range of the scale (e.g. *poprilično skup* 'rather expensive'), and diminishers which attenuate the modified property (e.g. *donekle skup* 'somewhat expensive'). Amplifying *totality* modifiers (e.g. *potpuno* 'completely'), also known as maximizers, correspondingly associate with bounded meanings, mostly with non–scalar complementaries (e.g. *potpuno prazan* 'completely empty' – *potpuno pun* 'completely full'), which project along an 'either–or' scale. Finally, approximators scale down the property, indicating that it falls short of the boundary (e.g. *skoro prazan* 'almost empty').

Naturally, individual DMs exhibit substantial variability in their constraints concerning general patterns (Kennedy and McNally 2005). Some, like *potpuno* 'completely', have been demonstrated to predominantly collocate with bounded heads (e.g. *potpuno prazan* 'completely empty'). On the other hand, booster *jako* 'very' is more commonly associated with non–bounded heads (e.g. *jako lijep* 'very nice'). Finally, certain DMs can align with both bounded and non–bounded heads and have two distinct readings. For example, when *posve* 'totally' collocates with a bounded head, it indicates that the properties denoted by the head are construed as reaching the highest extreme on the scale, as in *posve prazan* 'totally empty'. However, when *posve* 'totally' collocates with heads of non–bounded meaning, it acts as a booster and elevates the gradable attribute up the scale, as in *posve lijep* 'very nice' (non–bounded). These peculiarities of each modifier concerning the gradability of the modified adjective, while not the primary focus of this study, can be inferred from the subsequent analysis.

3. Method

Guided by one of the fundamental maxims of corpus linguistics, "a word is known by the company it keeps" (Firth 1957), it is presumed that the context in which a variable (be it lexical or phrasal) appears provides valuable insights into its semantic characteristics (Sinclair 1991; Lenci and Sahlgren 2023).

To explore <degree modifier + adjective> constructions, investigating co–occurrences between modifiers and adjectives in a substantial corpus is a logical starting point. In this study, we analyse the co–occurrence of degree modifiers and adjectives to examine the Croatian degree modifier paradigm. To distinguish between adjectives that exhibit a notable connection with the analysed degree modifiers and those that frequently appear in the corpus regardless of context, analytical statistics circumvents a reliance solely on raw counts and basic relative frequencies. Analytical statistics involves the application of simple collexeme analysis (SCA), a technique from the collostructional analysis family (Stefanowitsch and Gries 2003), to investigate the distribution and collocational preferences of Croatian degree modifiers by quantifying the level of attraction or repulsion that words demonstrate towards constructions. The design of this method allows for the consideration of the two–way semantic influences between DMs and adjectives, effectively filtering out adjectives with high overall token frequencies in the corpus. The results obtained through SCA will be used as input for one specific multifactorial analysis method: hierarchical agglomerative cluster analysis (HACA)[1].

Considering that degree modifiers are commonly found in informal texts, books, and periodicals (Xiao and Tao 2007), the corpus of choice was *hrWaC* 2.2 (Ljubešić and Klubička 2014), a web corpus compiled from the .hr domain. Established in 2014, the corpus comprises a vast collection of written Croatian texts, amounting to 1.211 billion words. The corpus *hrWaC* was examined through SketchEngine's search interface. To identify modifiers of adjectives (adverbs) appearing in the typical context of <degree modifier + adjective>, a simple query ([tag="R.* "][tag="A.* "]) was utilised. No restriction was placed on adverbs, so every lexical element deemed by the *hrWaC* tagging regime to be an adverb was included. For reasons of practicality, the study was limited to the 300 most frequent adverbs preceding the adjectives, which were extracted and further examined.

Distinguishing whether adverbs function as degree modifiers and classifying them into subcategories is challenging to achieve solely through corpus–based techniques. The semantic context is not always sufficient for this purpose. Experimental methods, as demonstrated by Paradis (1997, 2000), are recommended. Paradis interprets <degree modifier + adjective> collocations in conjunction with intonation patterns from the prosodically annotated London–Lund Corpus of Spoken English. The author (1997: 20) employs the prosodic–semantic equivalence criterion to aid in this determination: "a modifier is defined as degree modifier if the

1 The analyses were conducted using *R* Statistical Software (version 4.2.2; R Core Team 2022).

degree meaning is predominant when it is used with contrastive focus, i.e. when the nucleus is on the modifier." Since this criterium was impossible to implement given the nature of the examined corpus, the selection was made at the author's discretion, considering the primate of the degree meaning.

Modifiers whose interpretation is ambiguous were excluded. For example, in (1 and 3), *stvarno* 'really' and *zaista* 'really' express a degree of reinforcement like that of the booster *vrlo* 'very', whereas in (2 and 4), their meaning is that of modality (with a reading of 'in truth'). Similarly, in (5), *tako* 'so' acts as a booster and conveys a degree, while in (6), its interpretation is that of quantity (cf. Paradis 1997 for ambiguous examples in English). Since a manual inspection of all examples is hardly possible given the approach and the corpus size, it was decided not to include the mentioned modifiers in the analysis.

(1) *Proces njihove izrade je stvarno dug.*
'The process of making them is really long.' (*hrWaC* 2.2)
(2) *Čak i u* […] *prirodnim žarištima bolesti koje prenose krpelji, samo mali broj krpelja je stvarno zaražen.*
'Even in areas that are considered natural hotspots for tick–borne diseases, only a small number of ticks are actually infected.' (*ibid*.)
(3) […] *vaterpolosti Jadrana i Pule priredili su zaista lijepu utakmicu.*
'The water polo teams of Jadran and Pula organised a really nice match.' (*ibid*.)
(4) *No, koliko je zaista kriva vlast?*
'But how much is the government really to blame?' (*ibid*.)
(5) *Ovdje je tako lijepo.*
'It is so nice here.' (*ibid*.)
(6) *Na rang listi vitalnih potreba ništa nije tako važno kao kisik.*
'In the ranking of vital needs, nothing is as important as oxygen.' (*ibid*.)

Furthermore, *iznimno* 'exceptionally' and *izrazito* 'extremely' were omitted since their meaning as described in dictionaries[2] (paraphrasable with "on rare occasions, in exceptional incidents; with a very clear expression") and the meaning they typically assume in actual usage (one of a booster *very*) mostly do not coincide.

After a manual inspection of the 300 most frequent adverbs in the <adverb + adjective> construction in *hrWaC*, 29 degree modifiers of adjectives whose primary reading[3] is that of degree were identified: *beskrajno* 'endlessly', *blago* 'slightly', *donekle* 'somewhat', *dosta* 'rather', *ekstremno* 'extremely', *itekako* 'very', *izuzetno* 'extremely', *jako* 'very', *krajnje* 'extremely', *maksimalno* 'as ADJ as possible', *malo* 'a bit',

2 *iznimno* (ADV) na izniman način 'in an exceptional way' < *izniman* (ADJ) koji je iznimka; osobit, poseban, 'which is an exception; particular, special'; *izrazito* (ADV) na izrazit način; jako 'in a distinct way; very' < *izrazit* (ADJ) koji je jasno izražen, prepoznatljiv, istaknut 'which is clearly expressed, recognizable, distinguished' (Šonje 2000).

3 A random sample of 500 occurrences per modifier was examined to decide its primary reading.

mnogo 'much', *podosta* 'rather', *poprilično* 'rather', *posve* 'completely', *potpuno* 'completely', *previše* 'too ADJ', *prilično* 'rather', *puno* 'a lot; much', *relativno* 'relatively', *sasvim* 'completely', *skroz* 'completely', *slabo* 'poorly', *strašno* 'frightfully', *totalno* 'totally', *užasno* 'terribly', *veoma* 'very', *vrlo* 'very', and *znatno* 'very'. The concordances were carefully reviewed to exclude examples where an adverb does not act as a degree modifier but serves as a quantifier of a noun following the adjective (e.g. *mnogo* in *mnogo skupih cipela* 'a lot of expensive shoes'). In the following steps, the 1000 most frequent adjectival collocates[4] were extracted from the corpus for each of the 29 modifiers. Collocates were extracted following a decision regarding a specific range (span) that will be analysed. This study focused on the prototypical contexts in which a degree modifier immediately precedes the adjective it modifies. To optimise the efficiency of extraction, cases of the nominal copulative predicate in which the sequence <degree modifier + adjective> is interrupted with a present or past form of verb *biti* 'to be', exemplified by (7) and (8), were not considered.

(7) *Vrlo je lijep*.
Very is 'be–3sg.PRS.' nice.
'He is very nice.'

(8) *Sasvim je bio pun*.
Completely was 'be–3sg.PST.' full.
'He was completely full.'

In fact, as pointed out by Desagulier (2014), embracing a constructional approach, i.e. treating the <modifier + adjective> sequence as a construction and limiting the semantic investigation to the syntactic frame of the construction, helps minimise the risk of obtaining irrelevant or noisy data. Nevertheless, it is important to note that due to the unique syntax of Croatian, adopting this approach could lead to the exclusion of valuable data. Consequently, we advocate for a comprehensive examination of "interrupted" constructions, which, due to practical limitations, were not addressed in this study. Finally, instances of DM "stacks" (where two or more DMs modify a single adjective), due to challenges in operationalizing them within the analysis techniques employed, were also excluded from consideration.

As one of the primary objectives of this study is to determine whether hierarchical clustering can effectively distinguish the paradigms of Croatian degree modifiers and assess whether resulting clusters demonstrate homogeneity or variability, we propose a division of the 29 degree modifiers according to the Paradis' (1997) taxonomy as a benchmark (Table 1). The most numerous class are boosters which account for 44,83% of all examined DMs. Interestingly, no approximators were found.

4 The SketchEngine platform poses a download limit of 1000 items from each list.

DEGREE	TOTALITY MODIFIERS		SCALAR MODIFIERS	
REINFORCER	**Maximizers**	*krajnje* 'extremely' *maksimalno* 'as ADJ as possible' *posve* 'completely' *potpuno* 'completely' *sasvim* 'completely' *skroz* 'completely' *totalno* 'totally'	**Boosters**	*beskrajno* 'endlessly' *veoma* 'very' *ekstremno* 'extremely' *vrlo* 'very' *itekako* 'very' *izuzetno* 'extremely' *jako* 'very' *mnogo* 'very' *previše* 'too ADJ' *puno* 'a lot; much' *strašno* 'frightfully' *užasno* 'terribly' *znatno* 'very'
ATTENUATOR	**Approximators**	/	**Moderators**	*dosta* 'rather' *podosta* 'rather' *poprilično* 'rather' *prilično* 'rather' *relativno* 'relatively'
			Diminishers	*blago* 'slightly' *donekle* 'somewhat' *malo* 'a bit' *slabo* 'poorly'

Table 1. Table of the 29 analysed DMs categorised based on Paradis' (1997) taxonomy.

4. Results

This section provides a concise overview of the outcomes stemming from the simple collexeme analysis involving 29 Croatian degree modifiers. These findings will serve as input for a more comprehensive exploration through hierarchical agglomerative cluster analysis.

4.1. Simple collexeme analysis

To quantify the degree of association (attraction or repulsion) between a linguistic unit, usually a construction[5], and its collocates, usually words that fill an empty slot of that particular construction, a method known as simple collexeme analysis (SCA) is employed[6]. In this study, for each degree modifier, we calculate the

5 Within collostructional analysis, according to the theoretical framework of Construction Grammar (cf. Hoffmann and Trousdale 2013), the concept of "construction" refers to a learned pair of form and meaning, varying from individual morphemes to extensive grammatical structures.

6 The analysis was conducted using an *R* script provided by Flach (2021).

collostruction strength of a given adjective (A) for <degree modifier + adjective> construction (C) and its 1000 most frequent adjectival collocates. As previously mentioned, instances of DM "stacks", as well as instances of non–modified adjectives were not considered.

The SCA method requires the following frequencies:

i. The raw frequency of A in the target slot of C, namely I_1
ii. The raw frequency of A in all other constructions except C, namely I_3
iii. The frequency of C with adjectives other than A, namely I_2
iv. The frequency of all other constructions then C with that of all other adjectives then A, namely I_4

These frequencies are organised in a 2x2 contingency table (Table 2), and the process is repeated for every adjective that occurs in the corpus' <degree modifier + adjective> construction (Stefanowitsch and Gries 2003).

	A	***¬ A***	**ROW TOTALS**
C	I_1	I_2	R_1
¬ C	I_3	I_4	R_2
COLUMN TOTALS	C_1	C_2	N

Table 2. 2x2 contingency table for simple collexeme analysis.

In addition to the information contained in the four subcomponents of the input contingency table, two supplementary values are necessary for SCA: the size of the units of analysis (often referred to as corpus size) (N) and the overall frequency of the examined construction (R_1). In line with the methodology proposed by Proisl (2022), the sample size N corresponds to the number of words that satisfy the restrictions on the target slot of C, i.e. the total number of adjective tokens in the corpus. Values for which the input is not provided, viz., C_1 (the number of words that are instances of A, i.e. the frequency of the examined adjective), C_2 (the frequency of all other adjectives besides the one in the examination), and R_2 (the frequency of all other constructions besides the examined one) are calculated from the input values. Upon calculating the chosen association measure using the contingency table, it becomes possible to rank all linguistic elements based on their association with the examined construction. To measure the level of association in the constructions <degree modifier + adjective>, this study adopts the log–likelihood ratio[7] (G^2) as the association measure of choice.

7 The log–likelihood ratio, together with the negative decadic log transformed *p*–value of Fisher–Yates Exact test, represent the two most frequently used association measures in the collostructional literature. For brevity, we will not delve into a detailed explanation of why one association measure is preferred over another. The advantages and drawbacks of commonly used association measures (AM) are discussed, *inter alia*, in Gries (2019). Furthermore, "much of this debate [on the right association measure] is by now probably fairly fruitless because it is likely that there simply is not one AM that fits all applications" (Gries 2023: 331).

In order to illustrate the SCA and justify the choice of the results coming from the analysis as an input for HACA over raw frequencies, the results of SCA for the construction <*veoma* 'very' + adjective> in *hrWaC* corpus will be presented. Collostructional analysis, namely SCA, has emerged as a favoured approach over raw counts and percentage–based methods as it efficiently filtered out co–occurring pairs that might have unreasonably high or low frequencies, regardless of corpus size. This filtering enables a more realistic interpretation of the results, leading to more meaningful analysis. Given that much of the research in the field of collostructions tends to concentrate on the top *n* highest–ranking items, it is logical to carefully evaluate the similarity between the top *n* lists generated by the two analyses, one based on the raw frequency and another collostruction strength (in this case, log–likelihood ratio (G^2) value). Table 3 indicates the top 15 adjectival collexemes of the construction <*veoma* 'very' + adjective>. It showcases the distinctions in ranking between raw frequency and collostruction strength. All collostructional strength results (coll.str.logl.) are significant at $p < 0.00001$.

veoma 'very'						
adjective	**frequency in corpus**	**frequency in construction**	**coll. str. logl.**	**frequency based rank**	**logl based rank**	**rank difference**
važan 'important'	550502	2751	9971,97	1	1	0
zanimljiv 'interesting'	336943	1148	3312,63	3	2	0
bitan 'important'	282950	916	2554,63	5	3	2
popularan 'popular'	133290	636	2223,76	6	4	2
težak 'heavy, difficult'	370095	937	2204,51	4	5	–1
koristan 'useful'	84653	479	1826,81	10	6	4
uspješan 'successful'	181749	605	1714,30	8	7	1
zahvalan 'grateful'	34488	302	1403,02	14	8	6

jednostavan 'simple'	176749	530	1401,62	9	9	0
interesantan 'interesting'	37561	290	1276,79	15	10	5
opasan 'dangerous'	101518	403	1269,84	13	11	2
jak 'strong'	286809	611	1252,62	7	12	–5
značajan 'significant'	153646	463	1227,96	11	13	–2
zadovoljan 'satisfied'	150896	405	990,62	12	14	–2
dobar 'good'	2156735	1670	952,34	2	15	–13

Table 3. Top 15 adjectival collexemes of *veoma* 'very' in <*veoma* 'very' + adjective> construction.

The main observation regards the difference in rankings. It can be noticed that the difference is not insignificant, and the ranking of some collexemes changes considerably. The rank scatter plot in Figure 1, based on the 400 most attracted collexemes of the construction <*veoma* 'very' + adjective>, supports this observation, revealing a strong but not perfect correlation between the rankings. Spearman's ρ of 0.8118 and Kendall's τ of 0.6304 indicate a moderate to high level of agreement between the two variables. Adjectives such as *lak* 'light; easy', *dug* 'long', *mal* 'small' and *dobar* 'good' exhibit significant ranking differences, as indicated by their positions with respect to the regression line in the scatterplot. The obtained level of correlation indicates that the SCA was successful in filtering out the collexemes that are omnipresent regardless of the context. Therefore, it is safe to confirm that a decision to use the results of the SCA of the 29 analysed modifiers as input for HACA, as noted in Gries and Stefanowitsch (2004b, 2010) and in line with Desagulier (2014), represents a methodologically sound choice.

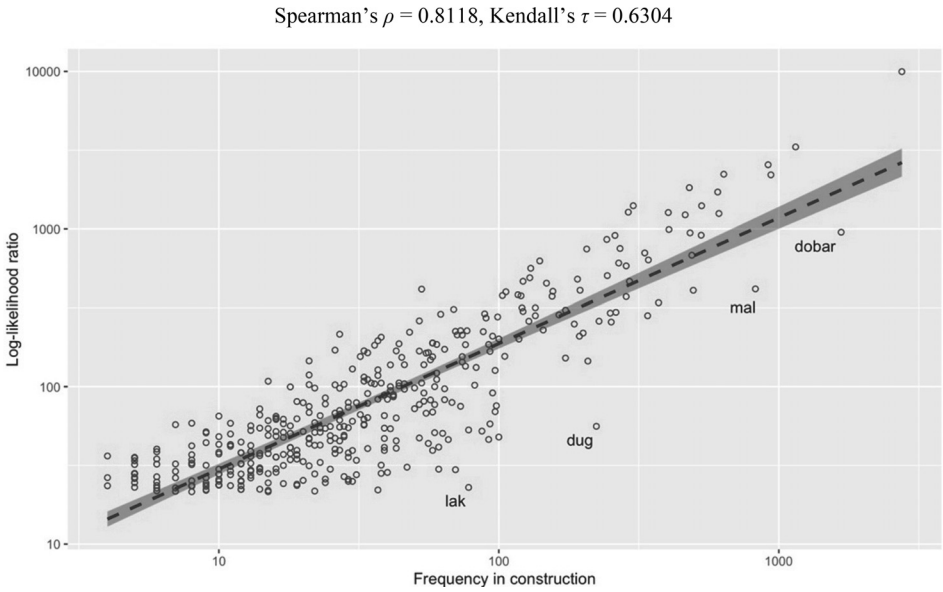

Figure 1. Scatter plot of observed frequency and G^2 scores of <*veoma* 'very' + adjective> construction, including a regression line and a 95% confidence region. Log_{10} transformed axes.

Due to space limitations, results of simple collexeme analysis (SCA) for each of the 29 examined degree modifiers cannot be presented *in extenso*. The table containing the top ten collexemes of analysed DMs can be found in Appendix 1.

4.2. Hierarchical agglomerative cluster analysis

Hierarchical agglomerative cluster analysis (HACA) encompasses a diverse array of multifactorial techniques designed to reveal underlying structures within data, specifically identifying clusters of similar objects based on their inter–object distances (Everitt et al. 2011). In HACA, entities are visualised as either leaves or branches in a dendrogram – a clustering tree that grows in the opposite direction of conventional trees (Levshina 2015). Each object – in this study, a constructional profile vector – commences as an individual cluster or "leaf" and is progressively merged with the most similar objects (i.e. those with the smallest inter–object distances) to form a cohesive and consistent paradigm, resulting in a unified tree structure.

In order to condense the dataset into more manageable proportions, the top 50 most attracted collexemes of each DM were taken further into the analysis, amounting to 1450 adjective tokens. After the cancellation of duplicate collexemes (ones appearing with two or more DMs), a list of 630 adjective types was obtained. The list was manually inspected to eliminate any unwanted elements (e.g. deviations from the Croatian standard orthography). Moreover, crystallised locutions such as *puno radno* [*vrijeme*] 'full–time' (in reference to a full–time job) and *skroz naskroz* 'completely' were excluded. Ultimately, we were left with a list of 544 adjec-

tive types. A 29–by–544 co–occurrence table reporting the frequency of each adverb–adjective pair type was submitted to HACA. HACA converts the input contingency table into a distance object, viz., a dissimilarity matrix (Table 4; Figure 2). The distances in HACA should reveal the level of (dis)similarity between constructions, determined by the proportions of contextual variable values present in the vectors. Smaller distances represent constructions with more similar vectors, while greater distances indicate more dissimilar vectors (Levshina 2015; Desagulier 2017). In this study, the Canberra distance measure[8] (Lance and Williams 1967), best suited for dealing with a substantial number of empty occurrences (Desagulier 2014, 2017; Gries 2021), was adopted using dist() function. Ward's method was applied for amalgamating clusters to produce a compact final dendrogram. The function used is hclust(). Despite being sensitive to the data points that exhibit substantial deviations with respect to the rest of the sample, the method has the benefit of producing clusters of moderate size[9] (Divjak and Fieller 2014).

The hypothesis of independence regarding the input data for HACA can be rejected: χ^2 =6963341[10], df = NA, *p*–value = 4.998e–04. Although we have established a statistically significant association between degree modifiers and adjectives, we lack information regarding the strength of this association. Since the χ^2 value is unsuitable for assessing the magnitude of the association, we need to turn to other measures that effectively capture the intensity of the association between the two variables. When working with a matrix larger than 2x2 and when the number of row variables differs from the number of column variables, Cramér's *V* emerges as the preferred measure (Levshina 2015; Desagulier 2017). Cramér's *V* ranges between 0 (no association) and 1 (perfect association). Computation of Cramér's *V* involves taking the square root of the χ^2 statistic divided by the product of the sum of all observations and the number of columns minus one. Cramér's *V* of 0.5207 in this study indicates a significative association between degree modifiers and adjectives (cf. Sheskin 2011 for guidelines on interpreting the effect size).

	previše	***prilično***	***puno***	***relativno***	***sasvim***
previše 'too ADJ'	0.0000	514.2312	520.0317	534.1755	540.6928
prilično 'rather'	514.2312	0.0000	463.7426	474.6945	524.5117
puno 'very'	520.0317	463.7426	0.0000	480.3576	527.1502
relativno 'relatively'	534.1755	474.6945	480.3576	0.0000	512.9570
sasvim 'completely'	540.6928	524.5117	527.1502	512.9570	0.0000

Table 4. A sampled dissimilarity matrix.

8 Canberra distance between vectors *x* and *y* is given by $\sum_i \frac{|x_i - y_i|}{(|x_i| + |y_i|)}$.

9 Agglomerative coefficient was calculated for "average" (0,1699), "single" (0,1630), "complete" (0,1706), and "ward" (0,4382) method. The Ward's method was chosen since it reveals the most pronounced clustering structure among the four evaluated methods.

10 Since cell counts are small, Pearson's Chi–squared test with Monte Carlo simulated *p*–value (based on 2000 replicates) was executed to get a *p*–value without assuming asymptotically normal behaviour.

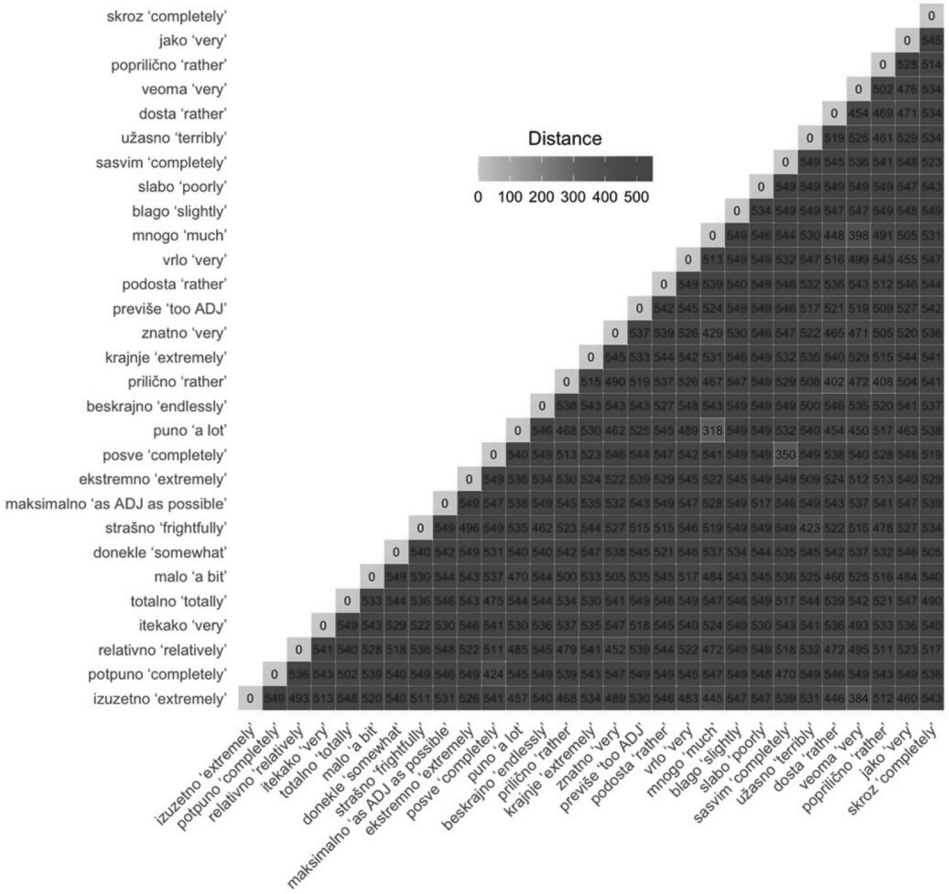

Figure 2. Dissimilarity matrix visualisation: heatmap with distance coefficients (rounded to 0 decimal places)

To validate the clusters and assess their strength, multiscale bootstrap resampling was employed, and bootstrapping–based cluster significance values[11] were computed using pvclust() package. The resulting dendrogram, displayed in Figure 3, can be examined from bottom to top. The height of the fusion indicates the grade of (dis)similarity between the two observations: the higher the merges, the less similar the observations. Each subcluster is accompanied by a rank value (from 1 to 27, as 27 clusters have been generated) and three cluster significance values. The number on the top left represents the "Selective Inference" *p*–value (SI)[12], the

11 The measure of dissimilarity is the "correlation" method.

12 In line with Shimodaira (2019), four versions of SI values were examined: the default pvclust() result; a recomputation via scaleboot compatible with pvclust(); a linear model (k=2); and a quadratic model (k=3). Although the quadratic model is anticipated to demonstrate reduced bias compared to the linear one, it tends to exhibit higher variance. Since attempts to implement scaleboot() with a broader range of scales to mitigate *p*–value variance while maintaining the distance measure and clustering method proved to be problematic, the default pvclust() results were used as considered sufficiently reliable.

one on the top right represents the "Approximately Unbiased" *p*–value (AU), meanwhile the other one indicates the "Bootstrap Probability" *p*–value (BP). The underlying principle across all three measures remains consistent: the closer the value approaches 100, the greater the robustness of the cluster. Since the SI *p*–value takes into account that clusters are selected based on data, contradicting the traditional approach of selecting null hypotheses before analysing the data, it is often favoured over AU and BP values for evaluating the stability and robustness of clusters (Shimodaira and Terada 2019). In any scenario, thoroughly assessing the reliability of clusters in HACA requires considering all values collectively to ensure informed decision–making.

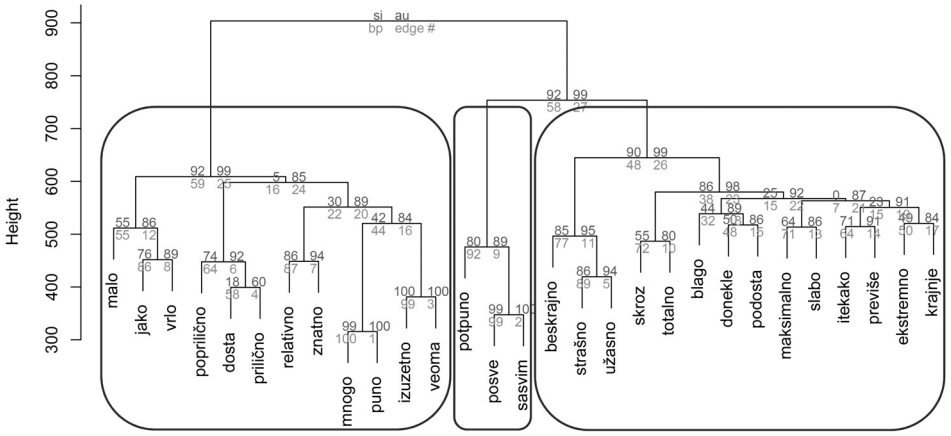

Figure 3. Cluster dendrogram of 29 Croatian adjectival degree modifiers, clustered according to their adjectival collexemes (distance: Canberra; cluster method: Ward.D).

As outlined in the Introduction, one of the primary objectives of this study was to assess whether hierarchical clustering can identify functional paradigms of Croatian degree modifiers that align with those presented in Paradis' (1997) taxonomy, as depicted in Table 1. The obtained dendrogram exhibits a moderate level of homogeneity within Croatian adjective degree modifiers, and certain discrepancies between the function of the modifiers and their clustering are visible. For interpretation, determining the most suitable number of clusters within a clustering solution is often advantageous. The elbow method and the average silhouette method (Kaufman and Rousseeuw 1990) were adopted, and it was concluded that the optimal number of clusters is three. Therefore, the dendrogram will be inspected as a composition of three main clusters, viz., clusters 25, 9 and 26, as indicated in the borders.

Several remarks can be made. Cluster 25 consolidates seven boosters (*izuzetno* 'extremely', *jako* 'very', *mnogo* 'much', *puno* 'a lot; much', *veoma* 'very', *vrlo* 'very', *znatno* 'very'), four moderators (*dosta* 'rather', *poprilično* 'rather', *prilično* 'rather',

relativno 'relatively'), and one diminisher (*malo* 'a bit'). It further subdivides into four subclusters. Two of these subclusters demonstrate functional homogeneity (one subcluster of sole moderators (subcluster 6; *dosta* 'rather, *poprilično* 'rather', *prilično* 'rather') and one of sole boosters (subcluster 12; *izuzetno* 'extremely', *mnogo* 'very', *puno* 'a lot; much', *veoma* 'very'), while the remaining two subclusters are of "mixed" nature (subclusters 7 and 13). Subcluster 7 comprises one booster (*znatno* 'very') and one moderator (*relativno* 'relatively'). Notably, subcluster 13 is particularly intriguing, encompassing two boosters (*jako* 'very', *vrlo* 'very'; forming a subcluster 8) and one diminisher (*malo* 'a bit'). Diminisher *malo* 'a bit' is a semantical and functional antonym of the two aforestated boosters[13], as it attenuates the properties it applies to. Considering that DMs are grouped based on their top 50 most preferred adjectival collexemes, it is reasonable to hypothesise that these three DMs, despite their distinct modification functions, likely have an overlap in favoured collexemes. This suggests their capacity to modify or describe similar types of entities, confirming that antonyms can operate within the same semantic domains. To examine this statement, the similarities between the three top 50 most attracted collexemes lists of *jako* 'very', *vrlo* 'very', and *malo* 'a bit' were quantified via the Dice similarity coefficient[14] (DSC), also known as the Sørensen–Dice index. DSC measures the overlap between *n* (usually two) sets, i.e., it tells us how many candidates from set X we can find in set Y or vice versa. In this study, DSC was computed between all pairs of sets (set1 (*vrlo* 'very') and set2 (*jako* 'very'), set1 (*vrlo* 'very') and set3 (*malo* 'a bit'), set2 (*jako* 'very') and set3 (*malo* 'a bit')). Additionally, the DSC providing a measure of overall similarity among all three sets was calculated (DSC_all). DSC ranges from 0 to 1, where the maximum value of 1 indicates that two sets share all collexemes, and a null value indicates that two sets do not have any collexeme in common. The results are as follows: DSC 1_2 = 0.4356; DSC 1_3 = 0.1818; DSC 2_3 = 0.2857; DSC_all = 0.3011. DSC_all of 0.3011 indicates a moderate level of similarity between the three compared sets, implying that approximately 30% of collexemes are shared. To gain deeper insights, it is recommended to conduct additional analyses, encompassing the calculation of DSC values for all sets and various clustering types[15]. These explorations are deferred to future studies.

Moving forward, the subsequent cluster, cluster 9, emerges as the most internally cohesive. It unites the maximizers *posve* 'completely', *potpuno* 'completely' and *sasvim* 'completely', with *posve* and *sasvim* forming a distinct, robust subcluster

13 To be more specific, the association between *vrlo* 'very' or *jako* 'very' and *mnogo* 'very' could be categorised as an instance of gradable antonymy (cf. Saeed 2008). Within this framework, negating one gradable term does not inherently validate the opposite term, allowing for additional states beyond those suggested by the antonyms.

14 $\mathrm{DSC} = \frac{2|A \cap B|}{|A| + |B|}$, where |A| and |B| represent the cardinalities of the two sets.

15 The clustering depicted in Figure 3 is exclusive, meaning each element is allocated to a single cluster. However, in various scenarios, a data point might reasonably belong to multiple clusters simultaneously. Exploring non–exclusive and fuzzy clustering methods to address this aspect is advisable. To deepen this topic, refer to, *inter alia*, Tan et al. (2018).

(SI value of 99). One plausible explanation for this can be attributed to the formality of these modifiers. *Potpuno* 'completely', *posve* 'completely', and *sasvim* 'completely' can be perceived as more formal degree modifiers, while *skroz* 'completely' and *totalno* 'totally' are more common in informal registers. A comparable scenario is observed in English, where the usage patterns of maximizers like *entirely*, *perfectly*, and *fully* are more prevalent in formal written genres. Conversely, maximizers such as *completely*, *absolutely*, and *totally* tend to be more frequent in informal contexts (Kennedy 2003). Moreover, distinct collexemes of *skroz* 'completely' and *totalno* 'totally' differ from those of *potpuno* 'completely', *posve* 'completely', and *sasvim* 'completely' regarding their semantics and semantic prosody (Lacić 2024). This aligns with the usage of *totally* in English which, according to Bordet (2017), tends to co–occur with colloquial adjectives that express intense emotions or judgments (e.g. *cool*, *awesome*, *hot*, *lame*, *psyched*). Further analysis is required for *krajnje* 'extremely' and *maksimalno* 'as ADJ as possible', as they were not included in the aforesaid study of Croatian maximizers.

Finally, cluster 26 brings together four maximizers (*krajnje* 'extremely', *maksimalno* 'as ADJ as possible', *skroz* 'completely', *totalno* 'totally'), six boosters (*beskrajno* 'endlessly', *ekstremno* 'extremely', *itekako* 'very', *previše* 'too ADJ', *strašno* 'frightfully', *užasno* 'terribly'), three diminishers (*blago* 'slightly', *donekle* 'somewhat', *slabo* 'poorly'), and one moderator (*podosta* 'rather'). It breaks into five subclusters. Two of the five subclusters exhibit functional homogeneity (subcluster 10 of sole maximizers (*skroz* 'completely', *totalno* 'totally') and one of sole boosters (subcluster 11; *beskrajno* 'endlessly', *strašno* 'frightfully', *užasno* 'terribly'), while the remaining three are of mixed nature (subclusters 14, 18 and 20). Subcluster 14 comprises one maximizer (*maksimalno* 'as ADJ as possible') and one diminisher (*slabo* 'poorly'). Subcluster 18 is comprised of two diminishers (*blago* 'slightly', *donekle* 'somewhat') and one moderator (*podosta* 'rather'). Subcluster 20 comprises three boosters (*ekstremno* 'extremely', *itekako* 'very', *previše* 'too ADJ') and one maximizer (*krajnje* 'extremely'). Similar to cluster 25, specific DMs within this group demonstrate intriguing clustering patterns. Subcluster 14, much like the previously mentioned subcluster 13, combines two degree modifiers that lie at opposite ends of the DM spectrum: the maximizer *maksimalno* 'as ADJ as possible' and the diminisher *slabo* 'poorly'. Once more, the connection exhibits a (partially) antonymic character. In this instance, *maksimalno* 'as ADJ as possible' denotes the utmost degree to which a property can be carried out, i.e. an exact match with a boundary. At the same time, *slabo* 'poorly', as a diminisher, points to the minimum extent of a certain property and slightly above that point. Nevertheless, the function of *slabo* 'poorly' extends beyond just indicating a specific degree of a scalar adjectival property; it also sets a boundary at the lower end of the zero–oriented scale (Paradis 2000). DCS between *maksimalno* 'as ADJ as possible' and *slabo* 'poorly' amounts to 0,1379, implying that approximately 14% of collexemes are shared. The fact that *maksimalno* 'as ADJ as possible' and *slabo* 'poorly' are clustered together with a relatively low DSC value

suggests that the clustering process may be capturing additional facets of similarity. In fact, upon closer examination of the shared collexemes of the two modifiers in question, it becomes evident that among the six shared collexemes, four of them (*iskoristiv* 'usable', *iskorišten* 'used', *popunjen* 'filled', *pripremljen* 'prepared') are distinct to the two modifiers, not appearing within the top 50 collexemes of any other degree modifier. The remaining two (*motiviran* 'motivated', *zaštićen* 'protected') are shared with only one additional modifier (viz., *motiviran* 'motivated' with *jako* 'very' and *zaštićen* 'protected' with *donekle* 'somewhat'). Given this reasoning, it is likely that the two modifiers are clustered together for this specific motive. Subcluster 11, on the other hand, captures attention since it reunites DMs associated with less advanced stages of grammaticalization with respect to other modifiers. DMs are typically sourced from the realm of adverbs, often originating from adjectives that undergo a process of adverbialization. Following Lehmann's (2015) parameters of grammaticalization in both syntagmatic and paradigmatic dimensions, each step along the continuum witnesses a reduction in paradigmatic weight or semantic integrity as the constructions undergoing grammaticalization experience a form of semantic attrition. Consider the modifiers *strašno* 'frightfully' and *užasno* 'terribly'. The original adjectives *strašan* 'frightful' and *užasan* 'terrible' underwent a process of semantical bleaching, partially shedding their adjectival meaning and acquiring a new degree–related sense of 'very'. In these instances, intensity is interpreted metaphorically. Nevertheless, their connection to the initial meaning remains more intact compared to, for instance, booster *vrlo* 'very'[16], deriving from an adjective *vrli* 'virtuous', that is, from the substantive *vrlina* 'virtue'. In contemporary use, the connection with the initial meaning is completely lost, unlike for modifiers *strašno* 'frightfully' and *užasno* 'terribly'.

While qualitative observations provide some reassurance, assessing the accuracy of the methodology presented is challenging without an alternative grouping for comparison. Ideally, this comparison would involve contrasting automatically derived clustering with non–automatic clustering. Drawing inspiration from Guzmán Naranjo and Bonami (2023), one potential approach could be to compare agglomerative cluster analysis with linguist–made clusters based on the observation of preferred collexemes of each modifier. That being said, future analyses are advised to (re)assess the clustering presented in this study.

16 Curiously, a parallel can be drawn with the English degree modifier *very*. It can be traced back to Middle English *verray* originally functioning as an adjective meaning 'true'. Yet, through a process of grammaticalization over time, it evolved into its current usage in Modern English as a degree modifier (booster) denoting a substantial degree.

5. Conclusions

This usage–based study has proposed and integrated two statistical methods to analyse the most attracted collexemes as well as clustering profiles of the 29 most frequent adjectival degree modifiers in the *hrWaC* corpus.

In adherence to the empirical cycle, the process commenced with the formulation of a theory–informed question and advanced to establish a research hypothesis. The hypothesis assumed that employing a cluster analysis method would enable the visualisation of functionally homogenous paradigms of Croatian degree modifiers, with Table 1 as a reference point. The hypothesis was put into action through operationalisation, followed by the analysis of corpus data. The 300 most frequent adverbs immediately followed by an adjective were extracted from the *hrWaC* corpus, and 29 degree modifiers with a primary degree reading were identified. The collected data then underwent testing utilising a simple collexeme analysis, the results of which were used as input for hierarchical agglomerative cluster analysis. The derived results were subjected to interpretation and critical examination of their significance.

Several observations can be drawn. Firstly, from a methodological perspective, the importance of incorporating statistical techniques in corpus–based analyses, often absent in studies of Croatian collocations, was emphasised. It was noted that collostructional analysis emerges as a superior approach to raw counts and percentage–based methods due to its ability to effectively mitigate the impact of adjectives with disproportionately high frequencies regardless of corpus size. In the case of the intensifying construction <*veoma* 'very' + adjective>, a significant disparity in rankings was noticed between raw frequency and collostruction strength (G^2) values for the top 15 most attracted collexemes. Additionally, a robust yet not flawless correlation was demonstrated by Spearman's ρ of 0.8118 and Kendall's τ of 0.6304, underscoring the significance of such approaches, particularly in studies that concentrate on the top n most attracted collexemes, as they enhance the precision and meaningfulness of result interpretation. Secondly, as shown by different preferred collexemes and, consequentially, by different clustering profiles of the degree modifiers belonging to the same paradigm (class), Paradis' (1997) perspective on the cognitive synonymy of English modifiers, showcasing similarities and differences, has been confirmed. However, the initial hypothesis regarding the potential use of HACA to identify functionally homogeneous paradigms of Croatian degree modifiers has not been supported. The cluster analysis based on preferred collexemes of each modifier did not yield clusters entirely driven by functional motivations, i.e. already known classifications such as Paradis' (1997) did not emerge from the data. This contrasts with the outcomes of Desagulier's (2014) research on 19 English DMs, which employed a comparable approach and achieved clusters with a notably higher grade of functional cohesion. The reasons for these disparities, particularly with an exploratory method like HACA, remain unclear and potentially range from

noise in the data to actual linguistic–based differences. Therefore, they warrant further, more nuanced investigations. A potential solution, as proposed in Lacić (2024), involves employing a technique such as multiple distinctive collexeme analysis (Gries and Stefanowitsch 2004a) subsequent to hierarchical clustering. This approach scrutinises the idiosyncratic modifiers of each DM in the analysis and could potentially elucidate the underlying reasons behind the generated clusters. All considered, it is safe to affirm that solely depending on cluster analysis to categorise Croatian degree modifiers into functional classes, or to derive hypothetical functional classes from generated clusters, may pose risks. However, due to a rather large number of examined DMs, the result does not come as unexpected as it is known that while DMs are cognitive synonyms (Paradis 1997) that share a fundamental functional basis, they may not always modify the same classes of adjectives and can function within distinct conceptual domains. Although the interchangeable use of modifiers may not be universally applicable, their cognitive synonymy finds substantiation through a noticeable convergence of collocational preferences observed within corpora. In this study, the three subclusters most strongly supported by data are subclusters 1 (*mnogo* 'very' – *puno* 'very' (DSC of 0,78, i.e. 39 shared collexemes)), 2 (*posve* 'completely' – *sasvim* 'completely' (DSC of 0,68, i.e. 34 shared collexemes)), and 3 (*izuzetno* 'extremely' – *veoma* 'very' (DSC of 0,62, i.e. 31 shared collexemes)). Evidently, these pairs exhibit an extensive number of shared collexemes, arguably indicating a robust status of their cognitive synonymy.

As previously highlighted, cluster analysis is an exploratory method, not an explanatory or predictive one, meaning it is used to investigate uncharted relationships, uncover novel connections, and formulate forthcoming research inquiries. The aspiration is that these objectives have been met, at least to some extent. To further deepen our understanding of the behaviour of the Croatian DMs, additional multifactorial techniques to analyse data regarding the whole DM paradigm should be put in place. Furthermore, it is advisable to integrate supplementary tiers of annotated data, such as details concerning grading force, boundedness, and the semantic categorisation of collexemes. Finally, the study should be expanded to include additional word classes that are subject to processes of degree modification. Optimistically, the methodology and the results presented in this study will serve as stimuli for a whole array of analyses to come regarding the Croatian degree modifiers and reference modulation in general.

References

Allerton, David J. (1987). English Intensifiers and Their Idiosyncrasies. Steele, Ross and Terry Threadgold, eds. *Language Topics: Essays in Honour of Michael Halliday*. *Vol. 2*. Amsterdam – Philadelphia: John Benjamins, 15–31

Batinić, Mia, Marijana Kresić and Anita Pavić Pintarić (2015). The intensifying function of modal particles and modal elements in a cross–linguistic perspective. *Rasprave* 41(1): 1–27

Bolinger, Dwight (1972). *Degree Words*. The Hague: De Gruyter Mouton

Cacchiani, Silvia (2017). Cognitive motivation in English complex intensifying adjectives. *Lexis: Journal in English Lexicology* 10(1): 1–21, https://doi.org/10.4000/lexis.1079

Claudi, Ulrike (2006). Intensifiers of adjectives in German. *Language Typology and Universals* 59(4): 350–369, https://doi.org/10.1524/stuf.2006.59.4.350

Desagulier, Guillaume (2014). Visualizing distances in a set of near–synonyms: *rather*, *quite*, *fairly*, and *pretty*. Glynn, Dylan and Justyna A. Robinson, eds. *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*. Amsterdam – Philadelphia: John Benjamins, 145–178, https://doi.org/10.1075/hcp.43.06des

Desagulier, Guillaume (2015). Forms and meanings of intensification: a multifactorial comparison of *quite* and *rather*. *Anglophonia – French Journal of English Linguistics* 20, https://doi.org/10.4000/anglophonia.558

Desagulier, Guillaume (2017). *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Berlin: Springer, https://doi.org/10.1007/978–3–319–64572–8

Divjak, Dagmar and Nick R. Fieller (2014). Cluster analysis: Finding structure in linguistic data. Glynn, Dylan and Justyna A. Robinson, eds. *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*. Amsterdam – Philadelphia: John Benjamins, 405–441, https://doi.org/10.1075/hcp.43.16div

Divjak, Dagmar and Stefan Th. Gries (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1): 23–60, https://doi.org/10.1515/cllt.2006.002

Dressler, Wolfgang U. and Lavinia M. Barbaresi (1994). *Morphopragmatics. Diminutives and Intensifiers in Italian, German, and Other Languages*. The Hague: De Gruyter Mouton

Everitt, Brian S., Sabine Landau, Morven Leese and Daniel Stahl (2011). *Cluster Analysis. 5th Edition*. Hoboken (NJ): Wiley–Blackwell, https://doi.org/10.1002/9780470977811

Firth, John R. (1957). *Papers in Linguistics, 1931–1951*. Oxford: Oxford University Press

Flach, Susanne (2021). Collostructions: An R implementation for the family of collostructional methods (v. 0.2.0.) [*R* Script]

Gries, Stefan Th. (2019). *Ten Lectures on Corpus–linguistic Approaches: Applications for Usage–based and Psycholinguistic Research*. Leiden – Boston: Brill

Gries, Stefan Th. (2021). *Statistics for Linguistics with R. 3rd Edition*. The Hague: De Gruyter Mouton, https://doi.org/10.1515/9783110718256

Gries, Stefan Th. (2023). Quantitative corpus methods in cognitive semantics/linguistics. Li, Fuyin Thomas, ed. *Handbook of cognitive semantics*. Leiden – Boston: Brill, 328–350, https://doi.org/10.1163/2773–1952_hcso_com_0210

Gries, Stefan Th. and Anatol Stefanowitsch (2004a). Extending collostructional analysis: a corpus–based perspective on alternations. *International Journal of Corpus Linguistics* 9(1): 97–129, https://doi.org/10.1075/ijcl.9.1.06gri

Gries, Stefan Th. and Anatol Stefanowitsch (2004b). Co–varying collexemes in the into–causative. Achard, Michel and Suzanne Kemmer, eds. *Language, Culture, and Mind*. Stanford: CSLI Publications, 225–236

Gries, Stefan Th. and Stefanowitsch, Anatol (2010). Cluster analysis and the identification of collexeme classes. Rice, Sally and John Newman, eds. *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford: CSLI Publications, 73–90.

Guzmán Naranjo, Matías and Olivier Bonami (2023). A distributional assessment of rivalry in word formation. *Word Structure* 16(1): 87–114, https://doi.org/10.3366/word.2023.0222

Harris, Zellig S. (1970). *Papers in Structural and Transformational Linguistics*. Dordrecht: Reidel

Hoffmann, Thomas and Graeme Trousdale (eds) (2013). The Oxford Handbook of Construction Grammar. Oxford: Oxford University Press, https://doi.org/10.1093/oxfordhb/9780195396683.001.0001

Ito, Rika and Sali Tagliamonte (2003). *Well* weird, *right* dodgy, *very* strange, *really* cool: Layering and recycling In English intensifiers. *Language in Society* 32(2): 257–279, https://doi.org/10.1017/s0047404503322055

Kaufman, Leonard and Peter Rousseeuw (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*. New York: John Wiley

Kennedy, Christopher (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30(1): 1–45, https://doi.org/10.1007/s10988–006–9008–0

Kennedy, Christopher and Louise McNally (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2): 345–381, https://doi.org/10.1353/lan.2005.0071

Kennedy, Graeme (2003). Amplifier Collocations in the British National Corpus: Implications for English Language Teaching. *TESOL Quarterly* 37(3): 467–487, https://doi.org/10.2307/3588400

Klein, Henny (1998). *Adverbs of degree in Dutch and related languages*. Amsterdam – Philadelphia: John Benjamins, https://doi.org/10.1075/la.21

Lacić, Ivan (2024). *A corpus–based study of maximizer–adjective patterns in Croatian*. *Language Sciences* 102, https://doi.org/10.1016/j.langsci.2023.101603

Lance, Godfrey N. and William T. Williams (1967). Mixed–data classificatory programs I – Agglomerative systems. *Australian Computer Journal* 1(1), 15–20

Lehmann, Christian (2015). *Thoughts on grammaticalization. 3rd Edition*. Berlin: Language Science Press, https://doi.org/10.26530/oapen_603353

Lenci, Alessando and Magnus Sahlgren (2023). *Distributional Semantics*. Cambridge: Cambridge University Press, https://doi.org/10.1017/9780511783692

Levshina, Natalia (2015). *How to do Linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam – Philadelphia: John Benjamins, https://doi.org/10.1075/z.195

Ljubešić, Nikola and Filip Klubička (2014). {bs,hr,sr}WaC –Web corpora of Bosnian, Croatian and Serbian. Bildhauer, Felix and Roland Schäfer, eds. *Proceedings of the 9th Web as Corpus* Workshop. Gothenburg: Association for Computational Linguistics, 29–35

Matešić, Mihaela and Anita Memišević (2016). Pragmatics of adjectives in academic discourse: from qualification to intensification. *Jezikoslovlje* 17(1): 179–206

Nigoević, Magdalena (2020). *Intenzifikacija u jeziku: s primjerima iz hrvatskog i talijanskog jezika*. Split: Filozofski fakultet Sveučilišta u Splitu

Nigoević, Magdalena and Josip Galić (2020). Strategije intenzifikacije u Brešanovoj drami "Predstava Hamleta u selu Mrduša Donja" i njezinu prijevodu na istromletački dijalekt. Levanat–Peričić, Miranda and Ana Gospić Županović, eds. *Prvi Brešanov svibanj. Zbornik radova s međunarodnog znanstvenog skupa o djelu Ive Brešana*. Zadar: Sveučilište u Zadru, 139–163

Paradis, Carita (1997). *Degree Modifiers of Adjectives in Spoken British English*. Lund: Lund University Press

Paradis, Carita (2000). It's well weird. Degree modifiers of adjectives revisited: the nineties. Kirk, John M., ed. *Corpora galore: Analyses and techniques in describing English*. Amsterdam: Rodopi, 147–160, https://doi.org/10.1163/9789004485211_014

Paradis, Carita (2001). Adjectives and boundedness. *Cognitive Linguistics* 12(1): 47–65, https://doi.org/10.1515/cogl.12.1.47

Paradis, Carita (2008). Configurations, construals and change: expressions of degree. *English Language and Linguistics* 12(2): 317–343, https://doi.org/10.1017/s1360674308002645

Pavić Pintarić, Anita and Zrinka Frleta (2014). Upwards Intensifiers in the English, German and Croatian Language. *Journal for Foreign Languages* 6(1): 31–48, https://doi.org/10.4312/vestnik.6.31–48

Proisl, Thomas (2022). *Use words, not constructions!* A new perspective on the unit of analysis in collostructional analysis. *International Journal of Corpus Linguistics* 27(3): 349–379, https://doi.org/10.1075/ijcl.20072.pro

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rotstein, Carmen and Yoad Winter (2004). Total adjectives vs. partial adjectives: Scale structure and higher–order modifiers. *Natural Language Semantics* 12(3): 259–288, https://doi.org/10.1023/b:nals.0000034517.56898.9a

Saeed, John I. (2008). *Semantics. 3rd edition*. Chichester: Wiley–Blackwell

Scheffler, Tatjana, Michael Richter and Roeland van Hout (2023). Tracing and classifying German intensifiers via information theory. *Language Sciences* 96, https://doi.org/10.1016/j.langsci.2022.101535

Sheskin, David J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures. 5th Edition*. Boca Raton: CRC Press, https://doi.org/10.1201/9780429186196.

Shimodaira, Hidetoshi (2019). *Computing selective inference p–values of clusters using pvclust and scaleboot*. Available on: http://stat.sys.i.kyoto–u.ac.jp/prog/ scaleboot/pvclust.pdf

Shimodaira, Hidetoshi and Yoshikazu Terada (2019). Selective inference for testing trees and edges in phylogenetics. *Frontiers in Ecology and Evolution* 7: 1–15, https://doi.org/10.3389/fevo.2019.00174

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press

Šonje, Jure (ed.) (2000). *Rječnik hrvatskoga jezika*. Zagreb: LZMK – Školska knjiga

Stefanowitsch, Anatol and Stefan Th. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243, https://doi.org/10.1075/ijcl.8.2.03ste

Stratton, James M. (*2020*). Adjective intensifiers in *German*. *Journal of Germanic Linguistics*, 32(2): 183–215, https://doi.org/10.1017/s1470542719000163

Tan, Pang–Ning, Michael Steinbach, Anuj Karpatne and Vipin Kumar (2018). *Introduction to Data Mining. 2nd Edition*. London: Pearson

Vidaković Erdeljić, Dubravka (2023). *Msm da sam totalno cool for the summer*: uporaba intenzifikatora u korpusu hrvatskih tweetova. Mikić Čolić, Ana and Maja Glušac, eds. *Standardni i nestandardni idiomi*. Zbornik radova s međunarodnoga znanstvenog skupa Hrvatskog društva za primijenjenu lingvistiku. Zagreb: HDPL, 251–262

Xiao, Richard and Hongyin Tao (2007). A Corpus–Based Sociolinguistic Study of Amplifiers in British English. *Sociolinguistic Studies* 1(2): 241–273, https://doi.org/10.1558/sols.v1i2.241

Uvid u paradigmu hrvatskih stupanjskih modifikatora i njezine klaster profile

Stupanjski modifikatori predstavljaju lingvističke jedinice korištene za preinaku drugih leksičkih elemenata s obzirom na njihov stupanj. Unatoč tome što je riječ o dobro proučenoj kategoriji u engleskoj lingvistici, stupanjski modifikatori u hrvatskome su jezikoslovlju nisu dobili dovoljno pozornosti. Ovaj rad ima za cilj doprinijeti proučavanju ovih jezičnih jedinica ispitivanjem skupine hrvatskih stupanjskih modifikatora kao dijela konstrukcije <stupanjski modifikator + pridjev>. Rad započinje analizom korpusa i identificiranjem 29 najčešćih stupanjskih modifikatora pridjeva u *hrWaC* korpusu. Kako bi se promatrani modifikatori analizirali, priklanja se distribucijskoj hipotezi i analiziraju se kolokacijski konteksti u kojima se modifikatori pojavljuju. Koristeći se jednostavnom koleksemskom analizom (engl. *simple collexeme analysis*), jednom od temeljnih metoda kolostrukcijske analize (engl. *collostructional analysis*), mjeri se asocijativna snaga pojedinačnih pridjeva s ciljanim konstrukcijama, odnosno kvantificira stupanj privlačnosti između stupanjskog modifikatora i pridjeva za svaku od 29 konstrukcija <stupanjski modifikator + pridjev> i njezinih 1000 najčešćih pridjevskih kolokata. Rezultati jednostavne koleksemske analize služe zatim kao ulazna informacija za hijerarhijsku aglomerativnu klastersku analizu (engl. *hierarchical agglomerative cluster analysis*) koja grupira stupanjske modifikatore na temelju njihovih preferiranih koleksema. Jednostavna koleksemska analiza pokazala se uspješnom u filtriranju kolokata koji su sveprisutni bez obzira na kontekst, dokazujući svoju prednost nad metodama koje se oslanjaju isključivo na "sirove" frekvencije. Nadalje, klasterska analiza otkrila je razlike između modifikatorskih funkcija i njihovog klaster profiliranja, što rezultira klasterima ograničene funkcionalne homogenosti. Unatoč tome, određeni potklasteri pokazuju savršenu, odnosno gotovo savršenu stabilnost i empirijsku potporu, potvrđujući (blisku) sinonimiju među analiziranim stupanjskim modifikatorima.

Keywords: collostructional analysis, corpus analysis, degree modifiers, hierarchical cluster analysis, Croatian

Ključne riječi: kolostrukcijska analiza, korpusna analiza, stupanjski modifikatori, pridjevi, hrvatski jezik

Appendix 1

beskrajno		***blago***		***donekle***		***dosta***		***ekstremno***		***itekako***	
collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.
zahvalan	2667,52	*povišen*	2698,72	*razumljiv*	1504,32	*dobar*	6690,31	*visok*	2513,99	*svjestan*	6587,50
tužan	891,81	*nagnut*	1391,59	*sličan*	1356,53	*skup*	4869,18	*nizak*	2073,16	*potreban*	3865,92
dosadan	789,34	*retardiran*	1279,18	*normalan*	849,62	*jak*	4153,38	*desni*	1938,22	*važan*	3798,43
dug	783,94	*zaobljen*	1139,76	*točan*	458,15	*težak*	3823,59	*desničarski*	1100,63	*bitan*	2469,53
sretan	605,40	*valovit*	992,66	*podnošljiv*	408,97	*jeftin*	3177,47	*sušan*	722,15	*prisutan*	1697,25
duhovit	470,19	*kiseo*	654,96	*ublažen*	317,72	*loš*	3123,90	*topao*	581,66	*vidljiv*	1668,09
strpljiv	455,72	*ojačan*	554,22	*pristojan*	295,78	*velik*	2958,32	*hladan*	394,70	*zanimljiv*	1378,69
šarmantan	423,03	*savinut*	526,41	*prihvatljiv*	294,52	*slab*	2473,63	*lijevi*	345,02	*dobrodošao*	1365,61
odan	313,05	*kiselkast*	404,53	*izmijenjen*	289,07	*zahtjevan*	2103,37	*brz*	328,56	*koristan*	1192,01
zabavan	299,11	*lužnat*	377,93	*opravdan*	288,85	*kompliciran*	1890,74	*težak*	321,28	*zadovoljan*	1027,95

izuzetno		***jako***		***krajnje***		***maksimalno***		***malo***		***mnogo***	
collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.
važan	20219,04	*dobar*	76442,69	*jednostavan*	3293,02	*koncentriran*	2860,68	*drugačiji*	23740,19	*velik*	12391,91
težak	6592,89	*bitan*	22979,37	*neizvjestan*	2149,31	*dopušten*	2051,56	*čudan*	13267,25	*dobar*	9344,04
zanimljiv	5545,43	*važan*	22608,32	*neozbiljan*	2152,54	*dozvoljen*	1847,11	*vjerojatan*	11369,57	*važan*	3731,08
kvalitetan	5375,89	*loš*	16816,96	*neodgovoran*	1909,96	*spreman*	1739,01	*drukčiji*	6085,67	*malen*	3456,73
uspješan	4860,36	*lijep*	14974,37	*nekorektan*	1838,79	*moguć*	1504,02	*kompliciran*	5300,86	*širok*	2816,83
povoljan	4505,03	*zanimljiv*	14844,08	*bezobrazan*	1736,49	*motiviran*	1484,84	*jak*	5157,91	*težak*	2636,13
bitan	3838,40	*zadovoljan*	14810,22	*neukusan*	1629,07	*angažiran*	1307,56	*star*	4374,39	*jak*	2561,65
zadovoljan	3641,91	*težak*	11441,29	*neprimjeren*	1442,71	*korektan*	855,49	*težak*	4028,12	*složen*	2113,49
jak	3522,22	*drag*	8788,76	*sumljiv*	1414,50	*pojednostavljen*	585,26	*velik*	3877,34	*lak*	1895,69
atraktivan	3009,85	*zahvalan*	7551,49	*ozbiljan*	1264,48	*oprezan*	562,75	*dobar*	3133,40	*dubok*	1642,58

podosta		***poprilično***		***posve***		***potpuno***		***previše***		***prilično***	
collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.
drugačiji	146,78	*jednostavan*	3524,99	*jasan*	11789.74	*nov*	30770.16	*optimističan*	1756,33	*jednostavan*	3254,99
udaljen	123,41	*loš*	2314,53	*drugačiji*	11155.40	*drugačiji*	29330.85	*zaokupljen*	1187,07	*loš*	2314,53
jeftin	81,88	*jasan*	2312,77	*drukčiji*	9613.77	*drukčiji*	9907.56	*kompliciran*	1125,74	*jasan*	2312,77
star	78,29	*zahtjevan*	1562,75	*nov*	7872.12	*različit*	9213.97	*zahtjevan*	1107,29	*zahtjevan*	1562,75
kompliciran	77,10	*težak*	1461,74	*siguran*	4261.34	*pogrešan*	8451.11	*zabrinut*	1061,10	*težak*	1461,74
zanimljiv	75,80	*neugodan*	1339,39	*normalan*	3971.87	*besplatan*	7362.07	*pametan*	1040,98	*neugodan*	1339,29
zahtjevan	75,75	*siguran*	1246,92	*različit*	3597.15	*nepoznat*	7359.73	*opterećen*	965,31	*siguran*	1246,92
izjednačen	73,89	*skup*	1136,90	*razumljiv*	3412.17	*jasan*	6939.81	*oduševljen*	963,96	*skup*	1136,90
izmijenjen	68,53	*dobar*	1112,53	*suprotan*	2553.16	*suprotan*	6892.05	*popustljiv*	925,95	*dobar*	1112,53
neobičan	63,46	*skroman*	1090,44	*nepotreban*	2076.03	*nepotreban*	6633.94	*zainteresiran*	777,64	*skroman*	1090,44
puno		***relativno***		***sasvim***		***skroz***		***slabo***		***strašno***	
collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.
dobar	55524,23	*malen*	20123,02	*normalan*	35155.33	*drugačiji*	1945.86	*poznat*	3769,23	*važan*	883,47
velik	24095,27	*kratak*	13244,64	*drugačiji*	34529.17	*normalan*	1530.26	*plaćen*	3429,23	*ljut*	859,08
malen	9044,39	*nizak*	7737,86	*dovoljan*	30057.49	*drukčiji*	894.35	*naseljen*	2701,88	*jak*	749,54
lak	6898,97	*nepoznat*	5564,09	*solidan*	20072.73	*okej*	765.84	*razvijen*	2528,15	*razočaran*	625,18
jeftin	6402,40	*visok*	5150,45	*jasan*	17104.28	*mokar*	589.93	*osvijetljen*	1763,00	*težak*	599,09
težak	6261,60	*jednostavan*	4186,66	*drukčiji*	12890.00	*lud*	577.81	*obrazovan*	1639,11	*skup*	524,49
jak	5638,74	*jeftin*	3916,73	*pristojan*	11192.59	*nebitan*	553.61	*pokretan*	1535,64	*loš*	464,07
važan	4501,67	*malen*	3487,14	*nov*	11067.66	*simpa*	538.80	*naoružan*	1206,74	*pametan*	353,73
brz	3832,99	*skroman*	3147,75	*logičan*	9520.23	*cool*	504.48	*zastupljen*	1162,19	*brz*	326,34
kvalitetan	3492,25	*mlad*	2665,96	*običan*	6647.76	*zadovoljan*	417.83	*informiran*	1003,99	*dosadan*	299,59

totalno		***užasno***		***veoma***		***vrlo***		***znatno***	
collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.	collexeme	coll. str.
nebitan	3827.50	*težak*	1697,58	*važan*	9971,97	*važan*	56852,96	*velik*	18510,06
drugačiji	3669.47	*dosadan*	1546,33	*zanimljiv*	3312,63	*jednostavan*	48366,94	*malen*	14178,94
glup	2485.75	*loš*	1252,17	*bitan*	2554,63	*zanimljiv*	40595,92	*nizak*	11185,01
lud	2072.74	*skup*	1128,76	*popularan*	2223,76	*dobar*	31297,76	*jeftin*	5630,13
promašen	2018.83	*naporan*	887,78	*težak*	2204,51	*kratak*	22651,07	*smanjen*	3858,33
drukčiji	1914.44	*glup*	824,71	*koristan*	1826,81	*uspješan*	22019,77	*povoljan*	3730,44
nepotreban	1658.65	*vruć*	673,78	*uspješan*	1714,30	*sličan*	19724,80	*dobar*	3005,34
nesposoban	1657.70	*ljut*	660,40	*zahvalan*	1403,02	*popularan*	17357,20	*visok*	2928,27
kriv	1567.09	*hladan*	592,17	*jednostavan*	1401,62	*bitan*	16762,46	*slab*	2635,42
zbunjen	1534.50	*ružan*	528,45	*interesantan*	1276,79	*težak*	16676,67	*oslabljen*	2315,50

Appendix 1. Top 10 collexemes of 29 Croatian degree modifiers with respective collostructional strength log-likelihood (G^2)