CHARITON CHARITONIDIS
Independent researcher
ORCID id: orcid.org/0000-0003-0298-2629

Fields of research: word formation, multiword expressions, lexical semantics, word recognition, emotion.

DOI: doi.org/10.35321/all90-11

EXPLORING SCALED AIC WITHIN ENGLISH CLOSED COMPOUNDS

Anglų kalbos uždarųjų junginių (sudurtinių žodžių) skalės AIC tyrimas

ANNOTATION

The Akaike Information Criterion (AIC) is an established goodness-of-fit measure for selecting models in the analysis of empirical data. However, AIC is sensitive to sample size. Author's previous research has shown that Scaled AIC, i.e. AIC divided by sample size, is an effective tool for assessing model fit and hierarchizing regression models. The present study explores further properties of this variable. The object of investigation are 66 multiple regression models referring to the processing of closed (concatenated) English compounds taken from Gagné et al.'s (2019) Large Database of English Compounds (LADEC). In particular, Scaled AIC is juxtaposed to the English Lexicon Project (ELP) and British Lexicon Project (BLP) as sources of response times, the lexical decision and naming tasks, compound length, and transparency norms. One-way ANOVA, main effects analysis, and non-parametric tests are used as methods. The findings suggest that Scaled AIC is responsive to experimental design, the source of response times, and the lexical decision and naming tasks. At the same time, the results of this study offer empirical support for the validation of methods employed by Gagné et al. (2019).

KEYWORDS: English compounds, Scaled AIC, lexical decision, naming.

ANOTACIJA

Akaikės informacijos kriterijus (angl. AIC) yra pastovus modelių tinkamumo matas, taikomas empirinių duomenų analizei. Tačiau AIC yra jautrus imties dydžiui. Ankstesni

autoriaus tyrimai parodė, kad skalės AIC, padalytas iš imties dydžio, yra veiksminga priemonė modelio tinkamumui įvertinti ir regresijos modeliams hierarchizuoti. Šiame tyrime nagrinėjamos tolimesnės šio kintamojo ypatybės. Tyrimo objektas – 66 daugialypės regresijos modeliai, susiję su uždaruųjų (sudurtinių) anglų kalbos junginių, paimtų iš Gagné'ės ir kitų (2019) Anglų kalbos sudurtinių žodžių (junginių) didžiosios duomenų bazės (angl. LADEC), apdorojimu. Pirmiausia AIC sugretinamas su Anglų kalbos žodyno projektu (angl. ELP) ir Britų kalbos žodyno projektu (angl. BLP), kaip atsako laiko, leksinių sprendimų ir įvardijimo užduočių, junginių ilgio ir skaidrumo normų šaltiniai. Naudojami metodai – vienpusė ANOVA (angl. Analysis of variance), pagrindinių rezultatų analizė ir neparametriniai testai. Išvados rodo, kad skalės AIC reaguoja į eksperimentinį projektą, atsako laiko šaltinį ir leksinių sprendimų bei įvardijimo užduotis. Tuo pačiu šio tyrimo rezultatai suteikia empirinį pagrindą Gagné'ės ir kitų (2019) taikomų metodų patvirtinimui.

ESMINIAI ŽODŽIAI: anglų kalbos junginiai (sudurtiniai žodžiai), skalės AIC, leksinis sprendimas, įvardijimas.

1. THE LARGE DATABASE OF ENGLISH COMPOUNDS (LADEC: GAGNÉ ET AL. 2019)[1]

The Large Database of English Compounds (LADEC: Gagné et al. 2019) is the largest existing database of compound words. It contains over 8000 nonspaced ("closed" or "concatenated") compounds (=nouns) selected from various sources including, among others, the CELEX database (Baayen et al. 1995), the English Lexicon Project (ELP; Balota et al. 2007), the British Lexicon Project (BLP; Keuleers et al. 2012), the British National Corpus (BNC), and Wordnet. From the full set of LADEC entries, 7,804 compounds can be uniquely parsed into two free morphemes constituents.[2] A vast variety of compounds is considered, for instance noun-noun compounds, e.g. *buttercup*, *shipyard*, compounds with a second constituent derived from a verbal stem, e.g. *pacemaker*, *painkiller*, etc. (for definitions of compound classes see Lieber 2004: 46). The first non-head constituent refers to a wide range of grammatical categories. Figure 1 contains a brief sample of LADEC entries.

Gagné et al.'s (2019) multiple-regression models include a wide range of predictor (=independent) variables, such as compound length, bigram frequency

[1] This section was adopted from Chariton Charitonidis (2022) with slight alterations.

[2] LADEC includes plurals of already listed compounds as separate entries.

at the morpheme boundary, family size, word frequency, probability and association (vector-based) measures, emotional/sentiment norms computed from participant ratings, etc. The log response times for the compounds from ELP (lexical decision, naming) and BLP (lexical decision) are used as dependent variables. For the most part, compound length (number of characters) and log compound (=word) frequency from the SUBTLEX-US corpus (Brysbaert, New 2009)[3] and BNC (BLP) are used as control variables. In Gagné et al.'s (2019) models, the predictor variables mentioned above had significant effects on lexical decision and naming times.

FIGURE 1. LADEC entries: sample

afterlife aircraft ashtray	daydreaming dimwit drawback	pacemaker padlock painkiller
backboard ballplayer buttercup	earthquake egghead eyebrow	shipyard shoelace shotgun
caretaker castaway crossfire	offspring outcasts overdrive	textbook throwback turnaround

The primary focus in Gagné et al.'s (2019) study was placed on various measures of *semantic transparency*. Gagné et al. (2019) asked participants to rate compounds considering how predictable the meaning of the compound is from its parts (*meaning predictability* ratings, compound-based) and how much of the meaning of each of the constituents is retained in the compound (*meaning retention* ratings, constituent-based). The authors found that the distribution of transparencies for the *second* constituent was much more peaked and higher than the distribution of transparencies for the first constituent (M_{C1}: 64.80 [SD: 19.59] vs. M_{C2}: 71.00 [SD: 16.46]. N = 8115). However, the rating for the

[3] The SUBTLEX-US corpus is a 51-million-token corpus based on subtitles from US films and television programs. Several recent studies have provided evidence indicating that frequency norms obtained from subtitles of movies and television programs tend to be more effective than those derived from printed texts when it comes to explaining the differences in lexical processing time and, in some cases, accuracy among native speakers of various languages (see Chen et al. 2018: 2 and the references therein).

first constituent was more strongly correlated with the rating for the entire compound than was the rating for the second constituent (c1~cmp: r = 0.75, p < .001 vs. c2~cmp: r = 0.66, p < .001. N = 429).[4] Most notably, the meaning retention rating for the first constituent and the meaning predictability rating for the compound predicted all three types of response times, i.e. ELP lexical decision, BLP lexical decision, and ELP naming times.

To conclude, the peaked and higher distribution of transparencies for the second constituent and the first constituent's better association with the compound's meaning predictability appear to be immediately mapped onto the head operations in English compounds. The second constituent, i.e. the head, is a unit whose transparency is enhanced categorially and semantically (as for the semantic aspect, see the relations of entailment and hyponymy). The first constituent, i.e. the modifier, is the most critical factor in establishing compound reference. As a result, its transparency covaries with the transparency of the compound most strongly.[5]

2. AKAIKE INFORMATION CRITERION (AIC)

In 1973, Hirotugu Akaike developed a method to estimate the relative expectation of *Kullback-Leibler distance* (Kullback 1959) using Fisher's *maximized log-likelihood* (Fisher 1922; see also Aldrich 1997). This measure, commonly referred to as the *Akaike Information Criterion* (AIC; Akaike 1973), introduced a novel framework for selecting models in the analysis of empirical data, marking a significant paradigm shift (Burnham, Anderson 2002).

AIC is typically calculated as follows: −2lnL + 2k, in which 'lnL' refers to the maximized/full log-likelihood of the model and 'k' refers to the number of parameters including the constant. A smaller set of predictors is typically associated with more efficient models (models with a lower information loss). The lower (=more negative) the AIC value, the better the fit of the model. In this context, AIC penalizes, as a goodness-of-fit measure, the use of a large number of predictors that, potentially, result in higher AIC values (see the '+2k' part of the AIC equation).

[4] Steiger's (1980) *z* test showed that this difference was significant, z = 27.71, p < .0001 (Gagné et al. 2019).

[5] By referring to previous research, Gagné et al. (2019) report that "the modifier (the first constituent in English) tends to play a larger role in the ease-of-relation selection during the processing of compounds and noun phrases."

AIC is sensitive to sample size. AICc, a corrected version of AIC, incorporates sample size through the formula $2k\,(k + 1)/(n - k - 1)$. However, it specifically addresses small sample sizes and is not recommended for models based on large sample sizes such as that in Gagné et al. (2019).[6] It should be noted that researchers such as Kenneth P. Burnham & David R. Anderson (2002) do not offer a definitive solution for comparing AIC values of models fitted on both different and large sample sizes.[7]

In particular, Burnham & Anderson (2002: 80–85, 334–335) provide a comprehensive discussion of the implications of unequal sample sizes for model comparison. They argue that employing information criteria to compare models with different sample sizes can lead to misleading results. Similarly, as noted in an online discussion by Svetunkov in 2016 (see reference after the bibliography), all information criteria are based on the likelihood function that, in turn, depends on sample size. Specifically, as the sample size increases, the likelihood decreases. Consequently, information criteria will also increase in such cases.[8]

3. PREVIOUS RESEARCH

In Charitonidis (2022) the AIC values for 44 multiple regression models with different combinations of emotion variables (valence, arousal, and concreteness for (a) words and (b) word contexts) were divided by sample size (N) to yield Scaled AIC (AIC/N) values.[9] Subsequently, these values were utilized to assess

[6] For further information on AICc, the reader is referred to Burnham & Anderson (2002: 374–380).

[7] One of the solutions that Burnham & Anderson (2002) propose refers to the transformation of the AIC values to "Akaike weights" that are defined as "the relative likelihood of the model, given the data" (Burnham, Anderson 2002: xiii; see also Wagenmakers, Farrell 2004).

[8] Available at: https://stats.stackexchange.com/questions/94718/model-comparison-with-aic-based-on-different-sample-size [accessed 16.06.2023]. The reader can comprehend Svetunkov's statement by substituting different values for the 'lnL' component in the AIC equation, while maintaining the '2k' component constant. A decrease in the lnL value will result in a higher, i.e. inferior, AIC value.

[9] In the literature, Scaled AIC is also referred to as "mean AIC". According to Svetunkov (personal communication), the practice of dividing the Akaike Information Criterion by the sample size is not novel. For instance, Hastie et al. (2009: 230–231) define AIC in a non-canonical manner, employing N as the denominator in the formula. While this deviation from the conventional AIC formula is not without its critics, it remains a prevalent approach, as exemplified by its inclusion in the statistical software package Stata. For instance, Stata reports "AIC divided by N" in its model output, as evidenced by various examples available online (Gratitude is extended to I. Svetunkov for providing this information).

and compare the models' goodness-of-fit. The insertion of key predictors into *global,* i.e. general, models showed that the BLP lexical decision times called for a better goodness-of-fit than the ELP lexical decision times. The fit of the ELP naming models fell within the range of those observed for the ELP and BLP lexical decision models. Most notably, *context concreteness for the second constituent* emerged as a significant predictor in all models with SUBTLEX-US frequency, across lexical decision and naming.

In Charitonidis (2024), all significant coefficients from the *global* models with SUBTLEX-US frequency were juxtaposed to the *hyponymy* variable (Gagné et al. 2020). It was found that models including both hyponymy and context concreteness for the second constituent were always associated with the lowest (=best) Scaled AIC value as compared to *nested*, i.e. reduced, models omitting either of these two variables. The subsequently applied Wald tests showed that nested models, always referred to a significant reduction (=deterioration) of the coefficient of determination (R^2). Tables 1 and 2 display the Scaled AIC values and the results of the corresponding Wald tests, respectively.

TABLE 1. Scaled AIC values for nested models omitting hyponymy ('Model 2') or context concreteness for the second constituent ('Model 3') from full models ('Model 1') to predict English Lexicon Project (ELP) lexical decision (LD) times, British Lexicon Project (BLP) lexical decision times, and ELP naming times

Model	Scaled AIC	AIC	N
ELP LD			
1	-3.36281[a]	-3557.85	1058
2	-3.30375	-4169.334	1262
3	-3.34845	-3700.038	1105
BLP LD			
1	-3.79552[a]	-2903.574	765
2	-3.76618	-3920.592	1041
3	-3.76718	-2987.37	793
ELP naming			
1	-3.58686[b]	-7396.108	2062
2	-3.54304	-8418.27	2376
3	-3.58379	-7389.784	2062

a. Predictors: (Constant), hyponymy judgement, length of compound, SUBTLEX-US frequency, representation valence (cmp), context concreteness (c2)

b. Predictors: (Constant), hyponymy judgement, length of compound, SUBTLEX-US frequency, context valence (cmp), context arousal (c1), context arousal (c2), context concreteness (c2)

TABLE 2. Wald tests for nested models omitting hyponymy ('Model 2') or context concreteness for the second constituent ('Model 3') from full models ('Model 1') to predict English Lexicon Project (ELP) lexical decision (LD) times, British Lexicon Project (BLP) lexical decision times, and ELP naming times

Model	R2 square	F change	df1	df2	p
ELP LD					
1	.184[a]	47.506	5	1052	.000
2	-.004	4.562	1	1052	.033
3	-.010	13.175	1	1052	.000
BLP LD					
1	.211[a]	40.479	5	759	.000
2	-.013	12.091	1	759	.001
3	-.015	14.007	1	759	.000
ELP naming					
1	.288[b]	118.711	7	2054	.000
2	-.003	8.286	1	2054	.004
3	-.003	8.309	1	2054	.004

a. Predictors: (Constant), hyponymy judgement, length of compound, SUBTLEX-US frequency, representation valence (cmp), context concreteness (c2)

b. Predictors: (Constant), hyponymy judgement, length of compound, SUBTLEX-US frequency, context valence (cmp), context arousal (c1), context arousal (c2), context concreteness (c2)

In conclusion, two different effect-size measures, namely Scaled AIC and R^2, hierarchized the same regression models identically while demonstrating the same preference for the best model. Thus, there is strong evidence that the Scaled AIC measure is a qualitative tool for assessing model fit.

4. THE PRESENT STUDY

The present study builds upon the author's previous research presented in section 3. The research subjects are 66 lexical decision and naming models for the English closed (concatenated) compounds built by Gagné et al. (2019). All models include SUBTLEX-US frequency as control variable. Our objectives are twofold and run in parallel. First, we assess the characteristics of Gagné et al.'s (ibid.) models. Second, we explore essential properties of the Scaled AIC measure.

The research questions are:

1. Is Scaled AIC sensitive to the model design in Gagné et al. (2019)? Which model groups are favoured?

2. What is the impact of the control variables 'compound frequency' and 'compound length' on Scaled AIC?

3. How is morphological transparency related to Scaled AIC?

Our study is structured as follows: Section 5 provides an overview of our methods. Section 6.1 provides descriptive statistics for Scaled AIC referring to the models under consideration. Emphasis is given to the parametric versus non-parametric characteristics of model categories. Section 6.2 explores the relationship between the source of response times and the lexical processing tasks. Section 6.3 juxtaposes Scaled AIC to the control variables 'compound frequency' and 'compound length'. In section 6.4 the significance levels of the transparency coefficients from Gagné et al.'s (2019) models are mapped onto the Scaled AIC values. The key findings are summarized in section 7, followed by a discussion of the results in section 8.

5. METHODS

Our general method was the comparative analysis of the main parameters and characteristics of Gagné et al.'s (2019) models, using Scaled AIC as the dependent variable. Independent variables included sample characteristics (e.g. response time source and the lexical processing tasks), study design (e.g. control variables), and the significance level of transparency coefficients, among other factors.

The specific statistical methods employed were as follows: (a) descriptive statistics pertaining to means and medians, along with the application of the Shapiro-Wilk test to assess the central tendency, variability, and distribution of Scaled AIC across different model categories and groups (sections 6.1 and 6.2), (b) main effects analyses conducted for the source of response times (ELP/BLP) and the lexical processing tasks (lexical decision/naming) (section

6.2), (c) distinct ANOVAs performed on response time source and the lexical processing tasks, incorporating compound length as a covariate (section 6.3), and (d) utilization of the Kruskal-Wallis and the Jonckheere-Terpstra tests to explore differences among the ranks of ordinally-recoded coefficients for semantic transparency (section 6.4). For more information on methods, the reader is referred to the analyses in sections 6.1–6.4.

6. ANALYSES

6.1. Scaled AIC vs. model categories

The 66 AIC values from Gagné et al.'s (2019) multiple-regression models with SUBTLEX-US frequency as a control variable were divided by each model's sample size to yield a set of 66 Scaled AIC values.

Table 3 below provides the descriptive statistics for Scaled AIC and Figure 2 displays the corresponding boxplot referring to the ordered set of values.[10] There were no outliers in the sample. The skewness (Sk) and kurtosis (Ku) values were tolerable.[11]

The mean value for Scaled AIC was -3.50030. The standard deviation was 0.19052, that is the observations were relatively tightly clustered around the mean. The minimum and maximum values were -3.85502 and -3.15754, respectively. The median value was -3.55732, i.e. slightly lower than the mean value.[12] The middle 50% of the data ranged between -3.66575 (first quartile, Q1) and -3.30959 (third quartile, Q3). Accordingly, the interquartile range (IQR) was 0.35616.

[10] The lower or *first quartile* line of the box (Q1) marks the boundary below which the bottom 25% of the data extends. Similarly, the upper or *third quartile* line of the box (Q3) marks the boundary above which the upper 25% of the data extends. The shaded area shows the boundaries of the middle 50% of the data or *interquartile range* (IQR), which can be computed by subtracting the first quartile from the third quartile (Q3-Q1). The horizontal line inside the box shows the median or *middle quartile* (Q2), i.e. the value that falls in the middle of the dataset.

[11] With reference to the SPSS environment, the values between -1 and +1 for skewness and between -2 and +2 for kurtosis are generally considered acceptable for normal distribution assessment. It is worth noting, however, that skewness and kurtosis alone do not provide a conclusive proof of normality (see also the discussion on https://www.researchgate.net/post/What_is_the_acceptable_range_of_skewness_and_kurtosis_for_normal_distribution_of_data).

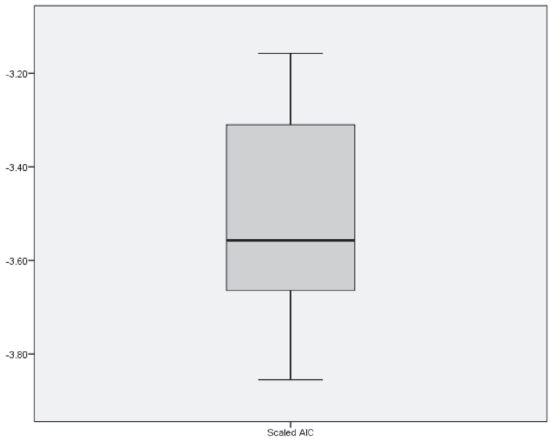
[12] In line with this pattern, there was a small amount of positive skew in the data (0.494).

TABLE 3. Descriptive statistics for Scaled AIC (full set)

Mean	SD	Min	Max
-3.50030	0.19052	-3.85502	-3.15754
Median	IQR	Q1	Q3
-3.55732	0.35616	-3.66575	-3.30959

N=66. Sk=0.494, Ku=-1.101

FIGURE 2. Distribution of Scaled AIC (full set): Boxplot



It was expected that the full set of Scaled AIC values would not show the normal distribution because they encompassed different tasks (lexical decision, naming) and were derived from different sources of response times (ELP, BLP). The Shapiro-Wilk test and the Kolmogorov-Smirnov test confirmed our assumptions.[13] In particular, the statistics for the Shapiro-Wilk test were $W(66) = 0.90$, $p < .001$, and the statistics for the Kolmogorov-Smirnov test were $D(66) = 0.17$, $p < .001$. It should be noted that the same full set did not exhibit a normal distribution, even after applying the natural logarithm and the square root transformations to the absolute values.

Table 4 below provides the descriptive statistics for Scaled AIC with reference to the main model categories in Gagné et al. (2019), i.e. ELP lexical decision, BLP lexical decision, and ELP naming (each group contains 22 models). Figure 3 displays the corresponding boxplots. As can be seen, the means and medians

[13] The Kolmogorov-Smirnov test applied the Lilliefors Significance Correction.

for BLP lexical decision and ELP naming referred to similar Scaled AIC values. ELP lexical decision referred to higher (=inferior) values that were, additionally, *disjoint* from the BLP lexical decision values.[14] In summary, BLP lexical decision and ELP naming always called for better models than ELP lexical decision.

TABLE 4. Descriptive statistics for Scaled AIC by the main model categories

	ELP lexical decision	**BLP lexical decision**	**ELP naming**
Mean	-3.25338	-3.62255	-3.62495
SD	0.06605	0.06778	0.08710
Min	-3.35102	-3.75547	-3.85502
Max	-3.15754	-3.52945	-3.50150
Median	-3.27165	-3.62174	-3.61886
IQR	0.13762	0.13704	0.13758
Q1	-3.31025	-3.68840	-3.69419
Q3	-3.17263	-3.55136	-3.55662
Sk	0.274	-0.097	-0.559
Ku	-1.444	-1.244	0.725
N	22	22	22

It was expected that the non-parametric profile of the full set of Scaled AIC values would be less likely to occur within the main model categories, i.e. the main combinations of response times and tasks. As will become apparent, this expectation was confirmed.

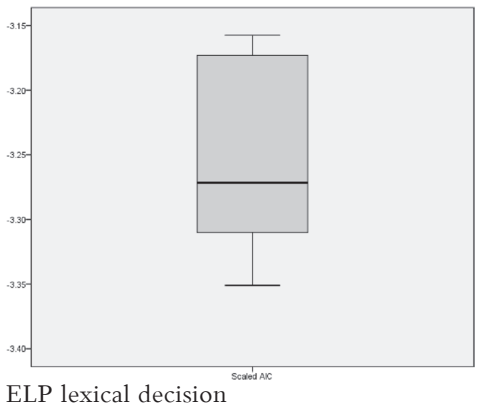
The Shapiro-Wilk test showed that the ELP lexical decision data deviated significantly from normality. In particular, we calculated a test statistic of $W(22) = 0.90$, $p < .05$.[15]

The BLP lexical decision data were normally distributed. The respective statistics were $W(22) = 0.92$, $p > .05$ (Shapiro-Wilk).

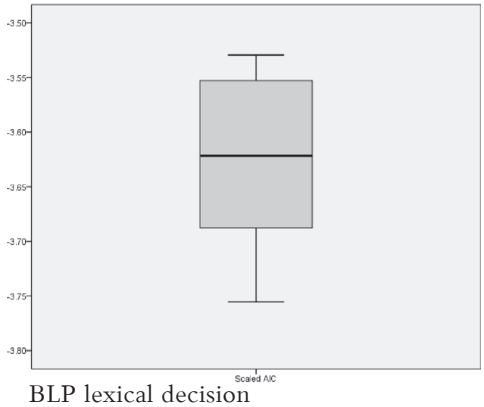
[14] A t-test on the summary data for ELP lexical decision and BLP lexical decision (Table 4) indicated a highly significant difference between sample means, $t = 18.296$, $p < .0001$.

[15] It should be noted that a set of six ELP lexical decision models referred to Scaled AIC values near the maximum value of -3.15754, indicating the poorest fit in the entire dataset. These models contributed to a prominent peak towards the higher end of the distribution (ranging from -3.17304 to -3.15754), resulting in an almost bimodal distribution pattern. This observation is supported by a relatively high negative kurtosis value of -1.444.

FIGURE 3. Boxplot representations of Scaled AIC by the main model categories



ELP lexical decision



BLP lexical decision

ELP naming

Similarly, the ELP naming data were normally distributed. The respective statistics were W (22) = 0.93, p > .05 (Shapiro-Wilk).

Summarizing, both the full set of models and the ELP lexical decision models were associated with a non-parametric distribution of Scaled AIC. On the other hand, the BLP lexical decision and ELP naming data were normally distributed. Let us now try to uncover the influence of individual factors on Scaled AIC.

6.2. Scaled AIC vs. response time source and tasks

Table 5 below contains the descriptive statistics for the Scaled AIC values for the groups categorized under the main factors 'response times' and 'tasks', i.e. (a) ELP response times, (b) BLP response times, (c) naming task and (d) lexical decision task. The ELP response times and the lexical decision task

are overarching groups referring to 'lexical decision & naming', and 'ELP & BLP', respectively. The groups 'BLP' and 'Naming' are identical to the model categories 'BLP lexical decision' and 'ELP naming', respectively that were already presented in Table 4 (section 6.1). Figure 4 displays the respective boxplot representations.

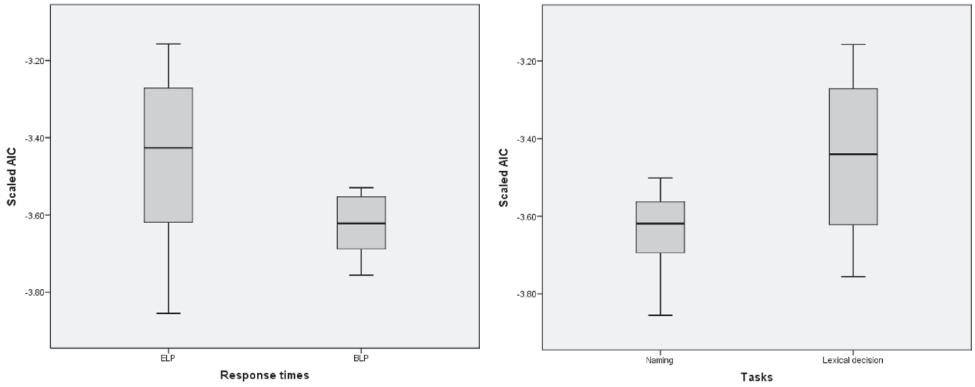
As can be seen in Table 5, both 'BLP' and 'Naming' referred to (a) lower (=better) main and median values for Scaled AIC, and (b) smaller standard deviation (SD) and interquartile range (IQR) values, in contrast to the overarching groups 'ELP response times' and 'lexical decision'.[16] The same overarching groups have relatively high kurtosis values, i.e. -1.403 and -1.616, respectively (in the histograms for these groups – not given here – two clear peaks emerged, similar to that for a bimodal distribution of values). These patterns suggest that the Scaled AIC measure was uniquely associated with the well-defined experimental categories BLP lexical decision (see 'BLP' group) and ELP naming (see 'Naming' group).

TABLE 5. Descriptive statistics for Scaled AIC regarding the groups categorized under 'response times' and 'tasks'

	Response times		**Tasks**	
	ELP	**BLP**	**Naming**	**Lexical Decision**
Mean	-3.43917	-3.62255	-3.62495	-3.43797
SD	0.20286	0.06778	0.08710	0.19808
Min	-3.85502	-3.75547	-3.85502	-3.75547
Max	-3.15754	-3.52945	-3.50150	-3.15754
Median	-3.42626	-3.62174	-3.61886	-3.44024
IQR	0.35152	0.13704	0.13758	0.36196
Q1	-3.62127	-3.68840	-3.69419	-3.63171
Q3	-3.26975	-3.55136	-3.55662	-3.26975
Sk	-0.126	-0.097	-0.559	-0.006
Ku	-1.403	-1.244	0.725	-1.616
	44	22	22	44

[16] Lower means and medians together with smaller standard deviation and interquartile range values indicate greater reliability and a reduced susceptibility to outliers or extreme values.

FIGURE 4. Boxplot representations for Scaled AIC regarding the groups categorized under 'response times' and 'tasks'



As with the entire set of Scaled AIC values discussed in section 6.1, it was expected that a non-parametric profile would be apparent for the overarching groups 'ELP' and 'lexical decision', not constrained by specific experiments. The normality tests confirmed our expectations.

With reference to the Shapiro-Wilk test, the Scaled AIC data for the ELP group deviated significantly from normality. We calculated a test statistic of $W(44) = 0.91$, $p < .01$. Similarly, the Scaled AIC data for the lexical decision group deviated significantly from normality. We calculated a test statistic of $W(44) = 0.89$, $p < .001$.

Let us now proceed to the main effects analysis. The primary objective was to ascertain whether the BLP lexical decision and the ELP naming groups continue to predict better models when controlling for the effects of either group.

In the statistical tests to follow, the groups categorized under the main factors 'response times' and 'tasks' were assigned nominal values. The Scaled AIC mean for the BLP group, i.e. -3.623, was used as the reference cell or intercept.

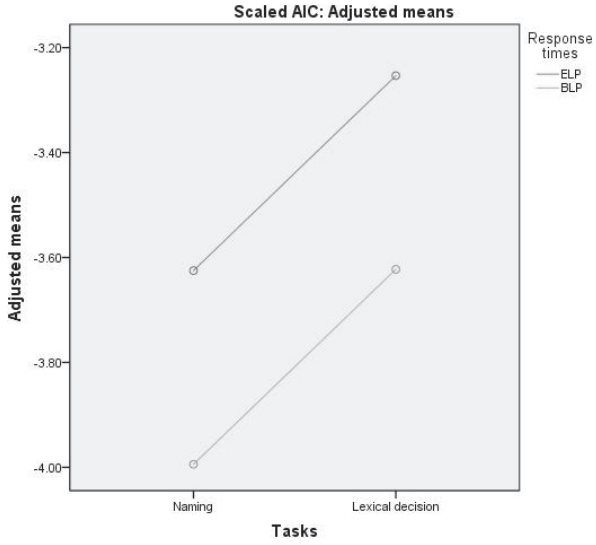
The results showed that both response times and tasks had a main effect on Scaled AIC, but there was no interaction. In particular, there was a significant main effect of response times, $F(1, 63) = 271.87$, $p < .001$. In regression terms, the coefficient for ELP predicted higher, i.e. inferior, Scaled AIC value, $b = 0.37$, $SE = 0.02$, $t = 16.49$, $p < .001$. In addition, there was a significant main effect of tasks, $F(1, 63) = 275.42$, $p < .001$. In regression terms, the coefficient for naming predicted a lower, i.e. better, Scaled AIC value, $b = -0.37$, $SE = 0.02$, $t = -16.60$, $p < .001$.

Figure 5 provides an overview of the main effects by means of a point and line plot. The intercept or reference cell refers to both the lowest ELP naming and highest BLP lexical decision value, i.e. -3.623. The lines are stacked

vertically because the ranges of Scaled AIC values for ELP lexical decision and BLP lexical decision were disjoint (see Table 4). On top of this, the lines are parallel because the absence of a 'BLP naming' group within tasks eliminated any interaction effects.

In summary, the BLP response times and the naming task persistently predicted better scaled AIC values, even after mutually controlling for relevant factors. These findings enhance the accuracy and effectiveness of the respective models.

FIGURE 5. Main effects plot for Scaled AIC



It should be noted that the main effects analysis presented in this section is more applicable to the source of response times than task performance because, as already mentioned, the combination 'BLP naming' was not available. This fact can limit the generalizability of the results beyond the specific samples.

6.3. Scaled AIC vs. control variables

In this section, the emphasis will be placed on the control variables compound length (in characters) and SUBTLEX-US frequency, i.e. the log compound frequency derived from the SUBTLEX-US corpus (Brysbaert, New 2009). Both variables were extensively used in Gagné et al.'s (2019) regression models.

To detect the influence of compound length and compound frequency on Scaled AIC, the respective regression coefficients were recoded into ordinal values according to their positivity and significance level, see Table 6. The significance levels were mapped onto *ordinal* scales because they represented conventional cut-off points based on the exact significance values.

TABLE 6. Ordinal recoding chart for regression coefficients

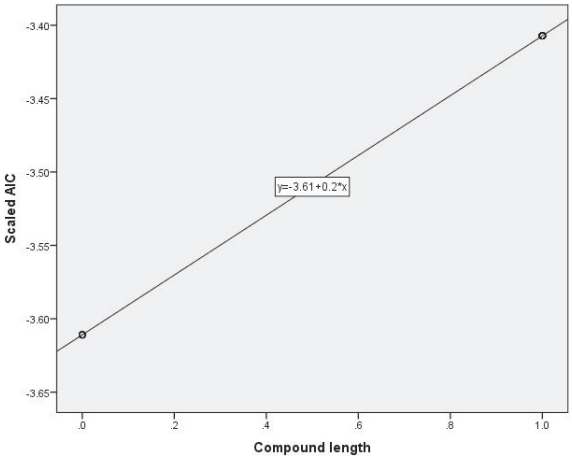
Significance	Positivity	Ordinal values	Description
$p < .001$	negative	-3	large negative effect
$p < .01$	negative	-2	moderate negative effect
$p < .05$	negative	-1	small negative effect
$p > .05$	negative/positive	0	non-significant effect
$p < .05$	positive	1	small positive effect
$p < .01$	positive	2	moderate positive effect
$p < .001$	positive	3	large positive effect

In Gagné et al.'s (2019) models, *SUBTLEX-US frequency* was always associated with negative (=latency-reducing) coefficients with a large effect, $p < .001$. Accordingly, all coefficients were recoded as -3, a value that was perfectly collinear with the outcome variable, Scaled AIC. For this reason, SUBTLEX-US frequency was excluded from the present analysis.

As for *compound length,* all significant regression coefficients from Gagné et al.'s (2019) models had a large positive (=latency-inducing) effect, $p < .001$. In contrast to the SUBTLEX-US variable, several non-significant coefficients showed up. Given these patterns, a categorical variable was created with the values '1' for positive effect (=interference of compound length) and '0' for no effect (=no interference of compound length). The resulting sample contained 39 Scaled AIC values. The Pearson correlation test between compound length and Scaled AIC yielded a highly significant correlation coefficient of 0.51, $p = .001$, indicating a moderate-to-strong correlation between the two variables.

Compound length was included as a single independent variable in a linear regression model. It was found that the predicted Scaled AIC mean for no interference of compound length was 3.611 (=the intercept). The interference of compound length resulted in a higher (=inferior) value of -3.407 ($b = 0.204$, $p = .001$). Figure 6 below illustrates these patterns. In a nutshell, Scaled AIC deteriorates when compound length becomes relevant within models.

FIGURE 6. Compound length and Scaled AIC



We conducted separate ANOVAs for response time source (ELP/BLP) and task performance (lexical decision/naming), taking into account compound length as a covariate. The primary objective was to identify disparities in means that were previously adjusted to accommodate the controlling effects of compound length.

(a) ANOVA for response time source. BLP was assigned the value '0' and ELP was assigned the value '1'. The Pearson correlation test revealed a strong collinearity between response time source and compound length, r = 1 (N = 39). The following evidence supports our finding: First, the ELP group consistently showed significant positive correlations with compound length, indicating a large effect. Second, the BLP group consistently displayed non-significant correlations with compound length.[17] Consequently, the predicted Scaled AIC mean for response time source was the same with or without compound length in the analysis (M = 3.407 in both cases). In summary, compound length did not have a significant effect on Scaled AIC when the response time source was included.

(b) ANOVA for task performance. Naming was assigned the value '0' and lexical decision was assigned the value '1'. The Pearson correlation test revealed a negative correlation between task and compound length, indicating a moderate effect, r = -.5, p = .001 (N = 39). This result suggests that compound length is more relevant to naming than to lexical decision.

[17] It should be noted that 11 of the 13 BLP coefficients were negative.

When considering task as the primary variable, compound length was a significant predictor of Scaled AIC, F (1, 145.030), p = .000. Similarly, when considering compound length as the primary variable, task was a significant predictor of Scaled AIC, F (1, 125.30), p = .000. The predicted Scaled AIC mean for task alone was significantly different from the predicted Scaled AIC mean when compound length was taken into account (3.584 vs. 3.203, respectively). Likewise, the predicted Scaled AIC mean for compound length alone (3.584) was significantly different from the predicted Scaled AIC mean when the task was taken into account (-3.23). Summarizing, in terms of covariate adjustment, both the lexical decision task and compound length predicted inferior models.

6.4. Scaled AIC vs. transparency norms

This section investigates the effect of regression coefficients for semantic transparency in Gagné et al. (2019) models on Scaled AIC. These regression coefficients were coded on three ordinal scales, each corresponding to one of the three morphological levels, i.e. compound, first constituent, and second constituent. The ordinal recoding chart can be found in Table 6.

Table 7 below displays the medians and ranges of the ordinally-transformed transparency coefficients for all three morphological levels. The medians provide useful information about the central tendency and dispersion of ordinal values and can help inform analyses based on ordinal variables.

TABLE 7. Ordinally-transformed transparency coefficients: Medians and ranges

	Median	Minimum	Maximum
Compound	-3	-3	-1
First constituent	2	-3	3
Second constituent	0	-2	3

N = 18

As can be seen, the median for the compound was '-3', the median for the first constituent was '2', and the median for the second constituent was '0'. These findings suggest that in Gagné et al.'s (2019) models with SUBTLEX-US frequency, transparency for the compound was associated with a large negative effect (shorter response times), transparency for the first constituent was associated with a moderate positive effect (longer response times), and transparency for the second constituent did not have a significant effect or had

an uncertain role. The higher transparency ratings for the second constituent, reported by Gagné et al. (ibid.), suggest an inherent bias favouring it, leading to the overall transparency of the compound being dependent on the transparency of the first constituent. In this context, the positive, latency-inducing, median for the first constituent indicates its mediating, perhaps reference-establishing, role in this relationship (see also section 1). It remains to be demonstrated which are the semantic functions that sufficiently represent, in processing terms, the inherent bias of the second constituent.[18]

The research question to be addressed now is whether the positivity and significance level of transparency coefficients influence Scaled AIC. Our method primarily aims at detecting overfitting effects. As Daniel J. Navarro & Jay I. Myung (2005) argue, *overfitting* occurs when "a complex model with many parameters and highly nonlinear form can often fit data better than a simple model with few parameters even if the latter generated the data" (Navarro, Myung 2005: 1240). Accordingly, a large number of parameters have the potential to capture noise or unique characteristics of the available data but may hinder the model's ability to *generalize to new data*. AIC mitigates the issue of overfitting by introducing a penalty on the inclusion of numerous parameters in a model, see the '+2k' part of the AIC equation in section 2.

Regarding the analysis to follow, it is postulated that models exhibiting higher (=inferior) Scaled AIC values may possess significant, systematically derived, coefficients, i.e. coefficients that are relevant according to the LADEC dataset alone. In this context, our conjecture suggests that a contrasting trend might emerge in the connection between transparency and Scaled AIC, as compared to the indication provided by the medians in Table 7.

In particular, lower (=better) Scaled AIC values may be associated with (a) positive coefficients (longer response times) concerning the whole compound, (b) negative coefficients (shorter response times) concerning the first constituent, and (c) positive or negative coefficients (longer or shorter response times, respectively) concerning the second constituent. It should be noted that, regarding the second constituent, the median in Table 7 suggests no effect.

To answer the research question, two non-parametric measures will be employed, i.e. the Kruskal-Wallis test and the Jonckheere-Terpstra test. The Kruskal-Wallis test, also known as the 'H test', is a non-parametric test based on

[18] In Charitonidis (2024) it is argued that both hyponymy and context concreteness for the second constituent are significant semantic predictors in lexical decision and naming. The analysis presented therein shows that including both of these predictors results in an improvement in Scaled AIC and R^2 as compared to models that omit either of these variables.

the chi-square distribution. It requires that the dependent variable be ordinal or continuous. This test is designed to determine whether there are significant differences between the medians of two or more groups and is used as an alternative to one-way ANOVA. Concerning the procedure, the values of the continuous dependent variable, i.e. Scaled AIC, were ordered from lowest to highest and the scores were assigned ranks. The resulting ranks were entered back into the groups of significance level (the independent variable) and the ranks for each group were summed. The formula for calculating 'H' involved, among others, squaring the sum of ranks for each group and then dividing this value by sample size.[19] Tables 8–10 contain the input data considered and the sum of ranks for each group.[20]

TABLE 8. Compound

	Significance levels	N	Sum of Ranks
Scaled AIC	1 – small negative effect	1	2
	2 – moderate negative effect	5	46
	3 – large negative effect	12	123
	Total	18	

TABLE 9. First constituent

	Significance levels	N	Sum of Ranks
Scaled AIC	1 – large positive effect	6	59
	2 – moderate positive effect	4	34
	3 – small positive effect	1	3
	4 – no effect	1	2
	5 – small negative effect	1	10
	6 – moderate negative effect	1	11
	7 – large negative effect	4	52
	Total	18	

[19] For the rest of calculations see Field (2009: 561–562).

[20] In all three tables, the total sum of ranks is approximately 171. It is equal to the sum of the integers from 1 to 18, see sample size (N).

TABLE 10. Second constituent

	Significance levels	N	Sum of Ranks
Scaled AIC	1 – large positive effect	6	59
	2 – small positive effect	1	14
	3 – no effect	8	59
	4 – small negative effect	1	18
	5 – moderate negative effect	2	21
	Total	18	

Before delving into the results of the Kruskal-Wallis (H) test, it is important to note that this test does not provide information about the specific differences between individual groups. To address this issue, the Jonckheere-Terpstra (JT) test was additionally employed. This test provided information about whether the medians of the groups increased or decreased in the order specified by the coding (=grouping) variable, specifically from large positive effect to large negative effect. Regarding methods, the JT statistic was converted into a z-score. A positive z-value indicated a trend of ascending medians, that is the medians increased (=higher/inferior Scaled AIC) as the values of the coding variable increased. A negative z-value indicated a trend of descending medians, that is the medians decreased (=lower/better Scaled AIC) as the values of the coding variable increased. In the following, the results of the Kruskal-Wallis (H) and Jonckheere-Terpstra (JT) tests are given jointly.

(a) Scaled AIC for the compound was not significantly affected by significance level, as determined by the Kruskal-Wallis test ($H(2) = 2.226$, $p > .05$). A trend of ascending medians was found confirming our overfitting hypothesis, see the negative median for the compound in Table 7. This trend, however, was not statistically significant according to the Jonckheere-Terpstra test ($JT = 50$, $z = 1.064$, $p > .05$).

(b) Scaled AIC for the first constituent was not significantly affected by significance level, as determined by the Kruskal-Wallis test ($H(6) = 5.427$, $p > .05$). A trend of ascending medians was found rejecting our overfitting hypothesis, see the positive median for the first constituent in Table 7. This trend, however, was not statistically significant according to the Jonckheere-Terpstra test ($JT = 70$, $z = 0.549$, $p > .05$).

(c) Scaled AIC for the second constituent was not significantly affected by significance level, as determined by the Kruskal-Wallis test ($H(4) = 4.607$, $p > .05$). A trend of ascending or descending medians was not observed, rejecting

our overfitting hypothesis. In particular, the z-statistic of the Jonckheere-Terpstra test was essentially zero, in accordance with the zero median for the second constituent in Table 7 (JT = 54, z = 0.041, p > .05).

Summarizing, it can be inferred that the significance level of the transparency coefficients in Gagné et al.'s (2019) models with SUBTLEX-US frequency does not affect the magnitude of Scaled AIC. This finding indirectly supports the quality of Gagné et al.'s (2019) models with transparency predictors, specifically indicating that the overfitting hypothesis for these models is not tenable. A limitation of the present study is the small sample size used, with N = 18. To confirm our findings, more research is needed using a wider range of Scaled AIC values.

To ensure clarity and completeness in presenting our research outcomes, we have incorporated a dedicated section focused on summarizing the key findings of our study. For this comprehensive overview, please continue to Section 7.

7. KEY FINDINGS

Table 11 below presents a comprehensive analysis of model performance and relevant variables in the context of lexical decision and naming tasks, based on the findings of Gagné et al. (2019). Each section of the table delves into specific subjects, revealing which models are most effective. The ANOVA and the Kruskal-Wallis/Jonckheere-Terpstra tests (sections 6.3 and 6.4) were applied after assigning nominal (ordinal or categorical) values to the regression coefficients from Gagné et al's (2019) models. For details on the special tests applied, please refer to the respective sections.

TABLE 11. Scaled AIC within English closed compounds: Comprehensive analysis of model performance in lexical decision and naming tasks (Gagné et al. 2019)

Subjects	Statistics	Evaluation		Section
Model categories	Descriptives Normality tests	ELP lexical decision BLP lexical decision ELP naming	NPAR/~ PAR/✓ PAR/✓	6.1
Response time source Lexical processing task	Main effects	ELP lexical decision BLP lexical decision ELP naming	~ ✓ ✓	6.2

Subjects	Statistics	Evaluation		Section
Control variables	ANOVA	Compound frequency	✓	6.3
		Compound length	~	
Semantic transparency	Kruskal-Wallis Jonckheere-Terpstra	First constituent	NOF/ns	6.4
		Second constituent	NOF/ns	
		Compound	NOF/ns	

PAR: parametric data | NPAR: non-parametric data | ✓: better models (lower AIC) ~: inferior models (higher AIC) | NOF/ns: no overfitting/non-significant test

8. DISCUSSION

Previous research by Charitonidis (2022, 2024) has demonstrated that Scaled AIC is a reliable goodness-of-fit measure that can be employed in model selection, perhaps in cooperation with other measures such as the Wald test (see section 3). With reference to Gagné et al.'s (2019) multiple-regression models with SUBTLEX-US frequency, the present analysis introduced additional properties of the Scaled AIC measure. While valid concerns have been raised regarding the comparison of models fitted on different sample sizes using information criteria (see section 2), the findings of this study suggest that in certain contexts, Scaled AIC can indeed be a valuable tool for assessing model fit and hierarchizing regression models. Our research has demonstrated that Scaled AIC is responsive to experimental design, response time sources, and specific tasks. However, it is essential to recognize that the applicability of Scaled AIC may be context-dependent, and its utility should be evaluated on a case-by-case basis.

Before proceeding to the primary findings of this paper, it is important to address the research questions set up in section 4.

1. The distributions of Scaled AIC values, along with combinations of different sources of response times and processing tasks, suggest that Scaled AIC effectively identifies the presence or absence of well-defined underlying factors in experimental design and statistical modelling. In this context, BLP lexical decision and ELP naming exhibited stronger predictive power for Scaled AIC even under controlled conditions.

2. Compound frequency was unexceptionally a negative predictor of Scaled AIC, always indicating a large effect. ELP lexical decision consistently showed

significant positive correlations with compound length predicting higher (=inferior) Scaled AIC values. BLP lexical decision consistently showed non-significant correlations with compound length. Both lexical decision and compound length predicted inferior models in covariate adjustment.

3. The positivity and the significance level of transparency coefficients in Gagné et al.'s (2019) models did not affect the magnitude of Scaled AIC. This finding implies that Gagné et al.'s (2019) models with transparency predictors do not introduce overfitting bias.

By referencing specific sections of the analyses, the primary findings of this study can be summarized as follows:

In Section 6.1, our analysis focused on the comparison between Scaled AIC values across various model categories. Even after attempting the transformations 'natural logarithm' and 'square root' on the absolute values, the overall Scaled AIC sample did not conform to a normal distribution. Similarly, the ELP lexical decision models showcased a non-parametric distribution of their Scaled AIC values. On the contrary, the data related to BLP lexical decision and ELP naming followed a normal distribution pattern.

In Section 6.2, our focus shifted to examining the relationship between Scaled AIC and (a) the sources of response times and (b) task performance. Interestingly, the ranges of Scaled AIC values for the ELP and BLP lexical decision models did not overlap, signifying their distinctiveness. The test results revealed significant main effects of both response time source and task performance on Scaled AIC. Notably, the predictive capability of Scaled AIC was better for models associated with the BLP lexical decision times and the naming task. These findings contribute to the precision and efficacy of the respective models significantly.

In Section 6.3, our exploration delved into the relationship between Scaled AIC and the control variables 'compound frequency' and 'compound length'. Compound frequency was excluded from the analysis because it was perfectly collinear with Scaled AIC. On the other hand, a decline in Scaled AIC values was observed when compound length became a relevant factor within models. Concomitantly, compound length was most relevant for the naming task.

In terms of covariate adjustment, both the lexical decision task and compound length were predictive of inferior models.

In Section 6.4, our focus was placed on the relationship between Scaled AIC and semantic transparency. The research question was whether the positivity and the significance level of transparency coefficients in Gagné et al.'s (2019) models had an impact on Scaled AIC. The primary goal of our method was to detect potential overfitting effects. We postulated that models with inferior Scaled AIC values might possess significant coefficients that hold relevance

according to the LADEC dataset alone. While we observed a trend of increasing (=inferior) Scaled AIC values for a cluster of significant negative coefficients at the compound level – aligning with our overfitting hypothesis – the Jonckheere-Terpstra test showed that this trend did not achieve statistical significance.

In conclusion, the exploration of different parameters using Scaled AIC as a dependent variable has illuminated the diverse ways in which model categories, response time source, processing tasks, control variables, and semantic transparency impact the goodness-of-fit of models. By recognizing the nuanced relationships among these elements, researchers are better equipped to make informed decisions in model selection, adjustments, and interpretation.

REFERENCES

Akaike Hirotugu 1973: Information Theory and an Extension of the Maximum Likelihood Principle. – *Second International Symposium on Information Theory*, eds. B. F. Csaki, B. N. Petrov, Budapest: Academiai Kiado, 267–281.

Aldrich John R. 1997: R. A. Fisher and the Making of Maximum Likelihood 1912–1922. – *Statistical Science* 12(3), 162–176.

Baayen Harald R., Piepenbrock Richard, Gulikers Leon 1995: *The CELEX Lexical Database* (Data set, Release 2, CD-ROM), Linguistic Data Consortium, University of Pennsylvania. Available at: https://catalog.ldc.upenn.edu.

Balota David A., Yap Melvin J., Cortese Michael J., Hutchison Keith I., Kessler Brett, Loftis Bjorn, Neely James H., Nelson Douglas L., Simpson Greg B., Treiman Rebecca 2007: The English Lexicon Project. – *Behavior Research Methods* 39, 445–459.

Brysbaert Marc, New Boris 2009: Moving Beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. – *Behavior Research Methods* 41, 977–990. DOI: doi.org/10.3758/BRM.41.4.977.

Burnham Kenneth P., Anderson David R. 2002: *Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach*, 2nd edition, New York: Springer. DOI: dx.doi.org/10.1007/b97636.

Charitonidis Chariton 2022: Context Concreteness for the Second Constituent Slows Down Compound-Word Processing. – *Lexis* 20. DOI: doi.org/10.4000/lexis.6769.

Charitonidis Chariton 2024: The Role of Hyponymy and Context Concreteness in Compound Word Processing. – *Studia Neophilologica*. DOI: doi.org/10.1080/00393274.2024.2336851.

Chen Xiaocong, Dong Yanping, Yu Xiufen 2018: On the Predictive Validity of Various Corpus-Based Frequency Norms in L2 English Lexical Processing. – *Behavior Research Methods* 50, 1–25. DOI: doi.org/10.3758/s13428-017-1001-8.

Field Andy 2009: *Discovering statistics using SPSS*, 3rd edition, London: Sage Publications.

Fisher Ronald A. 1922: On the Mathematical Foundations of Theoretical Statistics. – *Philosophical Transactions of the Royal Society of London, Series A* 222, 309–368.

Gagné Christina L., Spalding Thomas L., Schmidtke Daniel 2019: LADEC: The Large Database of English Compounds. – *Behavior Research Methods* 51, 2152–2179. DOI: doi.org/10.3758/s13428-019-01282-6.

Gagné Christina L., Spalding Thomas L., Spicer Patricia, Wong Dixie, Rubio Beatriz, Perez-Cruz Karen 2020: Is Buttercup a Kind of Cup? Hyponymy and Semantic Transparency in Compound Words. – *Journal of Memory and Language* 113. DOI: doi.org/10.1016/j.jml.2020.104110.

Hastie Trevor, Tibshirani Robert, Friedman Jerome 2009: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, New York: Springer.

Keuleers Emmanuel, Lacey Paula, Rastle Kathleen, Brysbaert Marc 2012: The British Lexicon Project: Lexical Decision Data for 28,730 Monosyllabic and Disyllabic English Words. – *Behavior Research Methods* 44, 287–304. DOI: doi.org/10.3758/s13428-011-0118-4.

Kullback Solomon 1959: *Information Theory and Statistics*, New York: Wiley.

Navarro Daniel J., Myung Jay I. 2005: Model Evaluation. – *Encyclopedia of Statistics in Behavioral Science*, vol. 3, eds. B. S. Everitt, D. C. Howell, Chichester: Wiley, 1239–1242.

Steiger James H. 1980: Tests for Comparing Elements of a Correlation Matrix. – *Psychological Bulletin* 87(2), 245–251. DOI: doi.org/10.1037/0033-2909.87.2.245.

Wagenmakers Eric-Jan, Farrell Simon 2004: AIC Model Selection Using Akaike Weights. – *Psychonomic Bulletin & Review* 11(1), 192–196.

INTERNET DISCUSSIONS

Model comparison with AIC based on different sample size. Available at: https://stats.stackexchange.com/questions/94718/model-comparison-with-aic-based-on-different-sample-size [accessed 15.5.2024].

What is the acceptable range of skewness and kurtosis for normal distribution of data? Available at: https://www.researchgate.net/post/What_is_the_acceptable_range_of_skewness_and_kurtosis_for_normal_distribution_of_data [accessed 15.5.2024].

LARGE LANGUAGE MODELS

GPT-3.5 (available at: https://chat.openai.com/) and Google Gemini (available at: https://gemini.google.com [accessed 11.10.2023]) were selectively used as auxiliary proofreading and editing tools. The responses from both AI tools were further proofread and edited by the author.

Anglų kalbos uždarųjų junginių (sudurtinių žodžių) skalės AIC tyrimas

SANTRAUKA

Šiame tyrime nagrinėjamos modifikuotos Akaikės informacijos kriterijaus, pavadinto skalės kriterijumi, t. y. AIC, kaip tinkamumo rodiklio, padalinto iš imties dydžio, varianto ypatybės. Tyrimo objektas – 66 daugialypės regresijos modeliai, susiję su uždarųjų (sudurtinių) anglų kalbos žodžių junginių, paimtų iš Gagné'ės ir kitų (2019) Anglų kalbos sudurtinių žodžių (junginių) didžiosios duomenų bazės (angl. LADEC), apdorojimu.

Toliau pateikiami išsamios analizės rezultatai:

1. Modelių kategorijų modeliai pasižymi nevienareikšmiais rezultatais. Britų kalbos žodyno projekto (angl. BLP) leksinių sprendimų modeliai ir Anglų kalbos žodyno projekto (angl. ELP) įvardijimo modeliai veikia geriau (tai rodo mažesnis AIC), palyginus su Anglų kalbos žodyno projekto (angl. ELP) leksinių sprendimų modeliais.

2. Laiko šaltinio ir leksikos apdorojimo užduočių atsakymas atskleidžia reikšmingus pagrindinius skalės AIC rezultatus. Britų kalbos žodyno projekto leksinių sprendimų modeliai ir Anglų kalbos žodyno projekto įvardijimo modeliai veikia gerai, o pastarojo projekto leksinių sprendimų modeliai yra mažiau veiksmingi.
3. Įvertinus kontrolinius kintamuosius, tokius kaip junginių dažnumas ir junginių ilgis, matyti, kad modelių rezultatyvumas skiriasi. Junginių dažnumas yra stiprus veiksnys (parodo didesnis produktyvumas), o sudėtinio ilgio modelių prognozės blogesnės.
4. Kalbant apie pirmosios ir antrosios žodžių junginių sudedamųjų dalių, taip pat ir apie junginių semantinį skaidrumą, modeliai nerodo per didelio suderinamumo. Tai parodo, kad semantinis skaidrumas nesumažina modelių galėjimo apibendrinti naujus duomenis.

Įteikta 2024 m. sausio 11 d.

CHARITON CHARITONIDIS
dr.chariton.charitonidis@gmail.com