Korean Journal of English Language and Linguistics, **Vol 26, February 2026, pp. 260-274**
DOI: 10.15738/kjell.26..202602.260

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Divergent Grammatical Trajectories of *Data*, *Criteria*, and *Phenomena*: A Corpus-Based Study

Siying Lyu · Ariadna Perdomo Bogliani · Isaiah WonHo Yoo (Sogang University)

Received: January 2, 2026
Revised: January 23, 2026
Accepted: January 24, 2026

Lyu, Siying (First author)
Ph.D. student, Division of English
Sogang University
Email: lvsiying0313@gmail.com

Bogliani, Ariadna Perdomo
(Co-author)
M.A. student, Division of English
Sogang University
Email: ariadnabogliani2000@gmail.com

Yoo, Isaiah WonHo
(Corresponding author)
Professor, Division of English
Sogang University
Tel: +82-2-705-8340
Email: iyoo@sogang.ac.kr

ABSTRACT

Lyu, Siying, Ariadna Perdomo Bogliani and Isaiah WonHo Yoo. 2026. Divergent grammatical trajectories of *data*, *criteria*, and *phenomena*: A corpus-based study. *Korean Journal of English Language and Linguistics* 26, 260-274.

Standard English grammar traditionally classifies *data*, *criteria*, and *phenomena* as the plural forms of *datum*, *criterion*, and *phenomenon*, respectively. However, contemporary usage frequently deviates from this prescription, showing a distinct tendency toward singularization. Based on the analyses of tokens from the Corpus of Contemporary American English (COCA) spanning from 1990 to 2019, this paper argues that these loanwords are undergoing two divergent grammatical shifts rather than a uniform process of simplification. Quantitative findings reveal that *data* exhibits a rapid and robust shift toward singular usage across Fiction, Newspaper, and Magazine registers, whereas *criteria* and *phenomena* retain a stronger adherence to traditional plural forms. Furthermore, qualitative concordance analyses demonstrate that *data* is predominantly evolving into a mass noun, evidenced by its exclusive compatibility with mass-selecting determiners (e.g., *much data*) and unitizing partitives. Conversely, *criteria* and *phenomena* are being re-atomized as singular count nouns, indicated by their reliance on singular count determiners (e.g., *a criteria*). Consequently, this study challenges the rigid prescriptive view of these nouns, suggesting that their semantic reconceptualization drives them along distinct trajectories: *data* as an unbounded aggregate and *criteria* and *phenomena* as discrete, countable entities.

KEYWORDS

loanwords, grammatical number, corpus linguistics, language change, mass-count distinction

1. Introduction

In contemporary English, a notable linguistic shift is the changing usage of nouns derived from Latin and Greek. Recent linguistic research increasingly views the traditional count–mass distinction not as a rigid binary but as a fluid continuum. Some scholars posit that this distinction is fundamentally a matter of vagueness, arguing that mass nouns are vague in a way that systematically impairs their use in counting (Chierchia 2010), while others suggest that nearly every noun can exhibit both count and mass properties through processes of sense extension (Drożdż 2020).

Against this theoretical background, this study deliberately focuses on *data*, *criteria*, and *phenomena* as its primary objects of analysis. These nouns were selected not simply because they are historically plural forms borrowed from Latin and Greek, but because they occupy a particularly revealing position at the intersection of frequency, morphology, and normativity in contemporary English. Unlike many classical loanwords that remain confined to specialized registers, these three items are highly frequent in both academic and general discourse, which makes them especially sensitive to usage-based grammatical change. High token frequency is well known to accelerate processes of reanalysis and restructuring, and thus these nouns provide an ideal testing ground for observing how prescriptive norms interact with actual usage over time.

Furthermore, *data*, *criteria*, and *phenomena* all share the plural suffix *-a*, a form that is morphologically opaque to most modern English speakers because it does not resemble the canonical English plural marker *-s*. This opacity weakens the transparency of their number marking and facilitates reinterpretation, for instance, the reanalysis of data as a singular mass noun. The coexistence of prescriptive pressure to maintain classical plural agreement and widespread usage favoring singular or mass-like patterns creates a productive tension that is central to the present study. For these reasons, these three nouns constitute particularly salient case studies for examining how the count–mass distinction is renegotiated in contemporary English through frequency-driven, usage-based change.

Although sharing the fact that they are all prescriptively plural nouns, the three nouns in question seem to have taken on different developmental pathways. *Data* seems to have long been used as a mass noun, similar in behavior to *information*, as illustrated in (1):

(1) **Much data** could be quoted as the history of baseball in the University. (Pound 1919, p. 164)

This developmental pathway can be traced to the 18th century, when *data* began to be conceptualized as a collective singular entity, distinct from its original Latin plural meaning (Rosenberg 2018). In contrast, *criteria* and *phenomena* are, at varying frequencies, being reanalyzed as singular count nouns. Consider the following examples, where *criteria* and *phenomena* are preceded by the indefinite article *a*:

(2) What is the reason behind them? If it meets **a legitimate criteria**, it's wonderful. (COCA: 1993: NEWS)

(3) It takes months between the time polar bears breed and when the female becomes pregnant because of **a phenomena** called delayed implantation. (COCA: 2017: NEWS)

While prescriptive grammar conventions often maintain the traditional plural-only usage (Ball 1928), descriptive evidence suggests a clear evolution in real-world language. This study proposes to empirically investigate this divergent phenomenon by systematically analyzing the usage patterns of *data*, *criteria*, and *phenomena* within a

large-scale corpus of contemporary American English. By comparing the changes in usage of these three words, this research aims to provide a quantitative account of this ongoing grammatical change. Accordingly, this study addresses the following three research questions:

1. What is the diachronic trend of plural vs. non-plural usage for *data*, *criteria*, and *phenomena* in English?
2. How does the plural and non-plural usage of these nouns vary in frequency across different linguistic registers, specifically Newspaper, Fiction, and Magazine?
3. What are the characteristic linguistic patterns associated with the singular usage of the target words, as observed through concordance line analysis?

2. Literature Review

Although standard grammar has traditionally treated *data*, *criteria*, and *phenomena* as the plural forms of *datum*, *criterion*, and *phenomenon*, respectively, descriptive accounts have long challenged this rigid stance. Over a century ago, Pound (1919, p. 164) observed that educated American speakers were already treating *data* and *criteria* as non-plural nouns (e.g., *Such data is misleading* and *That is no criteria*).

Given its prominence, the evolution of *data* has been particularly documented. According to the Merriam-Webster dictionary (n.d.), the word *data* is considered to have taken "a life of its own," and the literature broadly supports this characterization. The first meaning of the word *data* in Latin was "given" and was mostly used in theology and mathematics, in which it referred to "things taken for granted and thus not inquired after" (Rosenberg 2018, p. 560). Slowly, the term started to move towards a "material" meaning, until it became "quantitative facts gathered through observation, collection, and experiment, which are then subject to mathematical manipulation and scientific or social analysis" (Rosenberg 2018, p. 566).

The use of this word in different fields seems to reflect similar results: the singular word *datum* is seldom used, and some still use *data* as a singular noun. When it comes to medicine, according to McAlister (2016), the plural use of *data* once outnumbered the singular by a factor of four, while nowadays they seem to be equally used. She also argues that the word has evolved since the time when it meant "given," that the belief that *data* is plural is not a part of that evolution, and that the word *agenda* follows a similar fate, with its singular counterpart *agendum* being barely used and the word not meaning what it once meant ("to do"). Regarding the use of *data* in scholarly articles, Bordignon and Maisonobe (2022) compare how it appears in research articles in two different fields (social sciences and physical sciences). They concluded that "there is no point in trying to define the word *data*" (p. 1176) and that it is much more than what it means, since researchers will give it the meaning when they write, appropriating it by accompanying it with possessive pronouns, and using all kinds of adjectives with it.

This tension between prescriptive rules and actual usage has defined the discussion for decades. Prescriptivists like Ball (1928) argued forcefully for the simplification and Anglicization of such loanwords (e.g., preferring *datums* or *formulas*), citing the confusion caused by retaining foreign plurals. Ball (1928) specifically highlighted the chaos surrounding *data*, which was "so often used in speech and writing as a singular noun" (p. 293). This historical debate establishes the exact conflict, prescriptive ideals versus usage-based reality, that this study aims to quantify.

As Drożdż (2020) notes, many historical regularities for count-to-mass shifts were based on linguists' intuitions and focused on limited categories of nouns. In contrast, recent large-scale corpus analyses have sought to systematically map the patterns of these grammatical changes based on real-world utterances. Brown (2009), for

example, used a survey to test the retention of Latin and Greek plurals versus their Anglicized counterparts. His findings suggested that context and register are key factors; for instance, speakers preferred the Latin plural *antennae* in a scientific context but the English plural *antennas* for a common household object (Brown 2009). Drożdż's (2020) work, which found that even nouns strictly classified as count or mass by dictionaries regularly appear in senses with the reverse property, underscores the value of empirical investigation in this area.

However, despite these foundational insights, significant gaps remain in the current literature. While early descriptive studies relied primarily on manual observation and later ones, such as Brown (2009), utilized elicited survey data, there is a paucity of large-scale, diachronic corpus analyses that track the actual naturalistic production of these forms over time. Moreover, existing research often treats loanword adaptation as a uniform process, overlooking the specific divergent pathways where one term evolves into a mass noun while others evolve into singular count nouns. This study seeks to bridge this gap by utilizing a large corpus to provide a robust, quantitative account of these shifts in the words *data, criteria*, and *phenomena*.

3. Methodology

3.1 Data Collection and Sampling

This study employed a corpus-based methodology, combining quantitative and qualitative approaches to analyze the data. To provide a broader historical context, preliminary trends were observed using the Google Books Ngram Viewer. Queries were structured to capture subject-verb agreement patterns (e.g., *data is* vs. *data are*) from 1800 to 2019, offering a macro-level perspective on the singularization process. The investigation then focused on three nouns: *data* (a plural of *datum*, from Latin), *criteria* (a plural of *criterion*, from the Greek *kritḗrion*), and *phenomena* (a plural of *phenomenon*, from the Late Latin *phaenomenon*, derived from the Greek *phainómenon*). The primary data source was the Corpus of Contemporary American English (COCA), spanning the years 1990–2019. COCA was selected for its large size, balanced composition of different registers, and its representation of current language usage. To ensure data validity for native-speaker use, this study exclusively drew on the Newspaper, Fiction, and Magazine registers of the corpus.

The initial retrieval yielded a total of 56,232 tokens: 51,528 for *data*, 3,231 for *criteria*, and 1,473 for *phenomena.* Given the disproportionately high frequency of *data*, a systematic random sampling method was applied exclusively to this noun (selecting every 10th entry) to generate a representative and manageable dataset (5,153 in total). The complete datasets for *criteria* and *phenomena* were retained for analysis without sampling. Following this selection process, tokens functioning as proper nouns (e.g., the specific name of a person or organization) were excluded (63 instances for *data*, 5 for *criteria*, and 7 for *phenomena*), as only common nouns were analyzed to determine grammatical countability.

3.2 Coding and Classification

The retrieved concordance lines were manually coded into three primary categories: plural, non-plural, and unidentifiable. The classification followed a strict evidence-based protocol to avoid etymological bias. The "plural" category was restricted to tokens exhibiting explicit grammatical evidence of plurality (e.g., plural verb agreement such as *data are*; plural determiners such as *these, those, many,* and *few*; plural collocations such as *scores of* and *hundreds of*; or plural pronoun references such as *them*). Plurality was not assumed by default; positive evidence

was required. Tokens were classified as "non-plural" (encompassing singular count nouns and mass nouns) only when explicit grammatical evidence was present. To accurately capture the divergent grammatical pathways (mass vs. singular count), this evidence was stratified into seven specific sub-categories (see Table 1). Tokens addressing multiple categories were coded into more than one subcategory.

Table 1. Classification and Operational Definitions of Non-Plural Sub-Categories

Sub-category	Explanation	Examples
Singular Determiner	Determiners that are strictly restricted to singular count nouns. These serve as primary evidence for reanalysis as a singular count noun.	e.g., *a, an, this, that, one, another, every, each, either*
Mass Determiner	Quantifiers that are strictly restricted to uncountable nouns. These serve as primary evidence for reanalysis as a mass noun.	e.g., *much, little, less, a great deal of*
General Determiner	Demonstratives and determiners that indicate a singular number but are neutral regarding countability. These confirm non-plural status but do not distinguish between mass and singular count.	e.g., *the whole, the entire*
General Partitive Structure	Structures used to denote a sub-category or taxonomic class. These confirm non-plural status but do not distinguish between mass and singular count.	e.g., *a type of, a kind of, a sort of*
Mass Partitive Structure	Structures used to quantify mass nouns by assigning them a discrete unit. These strongly imply the target word is functioning as a mass noun.	e.g., *a piece of, a bit of, a large amount of, a scrap of*
Pronoun Reference	Singular pronoun substitution	e.g., *usage of it*
Singular Verb Agreement	Third-person singular verb forms	*e.g., data is, criteria has*

The "unidentifiable" category encompassed all instances in which the grammatical number remained ambiguous or indeterminable. To maintain analytical rigor, any token without positive grammatical evidence for either the plural or non-plural categories was assigned to this group. This included nouns modified by the definite article *the* without further context, as illustrated in (4), where *the phenomena* could theoretically be replaced by either singular *this* or plural *these* without violating grammaticality:

(4) We will uproot aggressive violence among us, and fight **the phenomena of racism**. (COCA: 2015: NEWS)

The category also included bare nouns appearing in headlines, labels, or lists, as shown in (5), where the lack of determiners or verbs makes number assignment impossible:

(5) # **DATA** # A SPECIES IN DECLINE # From 1972 to 2011, scientists caught more than 5,500 Antarctic toothfish in the Ross Sea (85 percent were tagged and released). (COCA: 2014: MAG)

Finally, instances modified by ambiguous quantifiers (e.g., *some, no,* and *any*) or appearing with past tense, modals, or other verb forms where number agreement is invisible were also classified as unidentifiable. In examples (6) and (7), neither the quantifiers nor the verbs provide information about countability:

(6) It's interesting that players agreed not to have **any data** found from player tracking used in contract negotiations. (COCA: 2019: MAG).

(7) Medical device executives told investors at a health care conference earlier this month that they have already examined **their internal data**, and they have not found support for the conclusion offered in the journal article. (COCA: 2019: NEWS).

3.3 Data Analysis

To ensure the reliability and consistency of the coding process, a rigorous intercoder reliability assessment was conducted. Before formal coding, a codebook was developed and refined through a pilot study to ensure shared operational understanding between the two coders. Subsequently, 10% of the analytical dataset (n=986) was randomly sampled and coded independently under blinded conditions. Intercoder agreement was assessed using Cohen's Kappa (κ). The analysis yielded a coefficient of $\kappa = 0.908$. According to Landis and Koch (1977), values between 0.81 and 1.00 indicate "almost perfect" agreement, confirming the high reliability of the coding protocol. Following this validation, the remaining corpus was divided between the two researchers for independent coding. Quantitative analysis was conducted to calculate the frequencies of each category across the three variables: year (1990–2019), register (Newspaper, Fiction, Magazine), and grammatical number (plural vs. non-plural). Within the non-plural category, the distribution of the seven evidence types was also calculated. Finally, a qualitative analysis of concordance lines was performed to examine the context of usage and select clear, illustrative examples for each pattern to support the quantitative findings.

4. Results and Discussion

4.1 Diachronic Evolution of Plural and Non-plural Usage

As illustrated in the Ngram Viewer graphs, the word *data* exhibits the most significant shift towards singular usage (see Figure 1). While the plural usage (*data are*) peaked around 1980, it has since shown a sharp decline. Conversely, the singular usage (*data is*) has risen steadily since the mid-20th century. Notably, the figure suggests a convergence of these two forms around the year 2019, where the frequency of the singular form appears to rival, and potentially overtake, the traditional plural usage.

In contrast, the trends for *criteria* and *phenomena* remain distinctly different. For both terms, the traditional plural usage (*criteria are* and *phenomena are*) continues to appear significantly more frequently than the singular forms. Although there is a discernible, albeit slight, upward trend in the singular constructions (*criteria is* and *phenomena is*) since the mid-20th century, the gap between singular and plural usage remains substantial. Similarly, *phenomena* shows a persistent preference for plural agreement, with the singular usage remaining a small minority throughout the observed period.

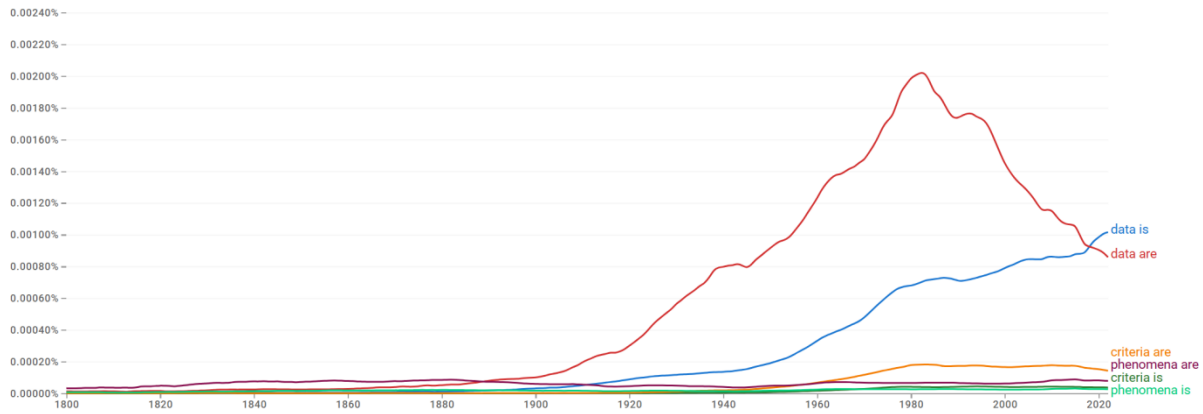

Figure 1. Diachronic Trends of "*data*/*criteria*/*phenomena*+is" and "*data*/*criteria*/*phenomena*+are" in Google Books Ngram Viewer (1800–2019)

To validate these macro-level observations within contemporary usage, the specific frequency of non-plural instances identified in the COCA dataset (1990–2019) was analyzed. As illustrated in Figure 2, the manual analysis of the COCA corpus confirms that *data* is undergoing a significantly more rapid process of singularization than the other two nouns. *Data* exhibits a pronounced upward trajectory: while the frequency of non-plural usage remained relatively stable between 1990 and 1999, it began to surge in the early 2000s, rising from approximately 50 instances per period to nearly 200 by 2015–2019. This sharp increase suggests that the acceptance of *data* as a singular or mass noun has accelerated in the 21st century. This rapid evolution of *data* compared to the lower-frequency terms *criteria* and *phenomena* aligns with Bybee's (2006) observation that high-frequency tokens are more susceptible to grammatical reanalysis.

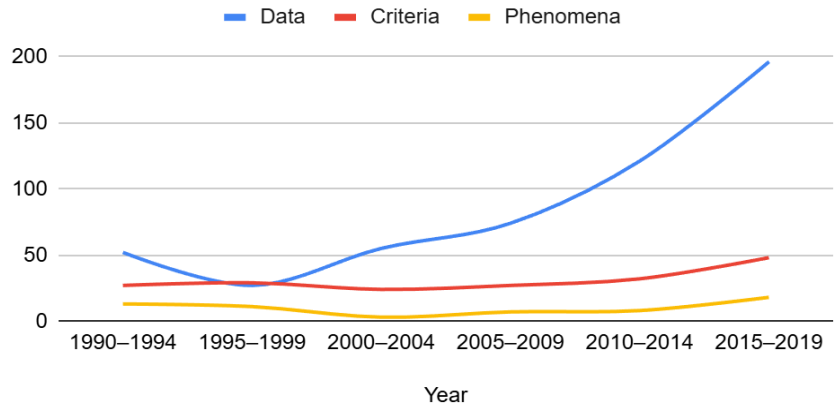

Figure 2. Frequency of Non-plural Tokens for *Data*, *Criteria*, and *Phenomena* in COCA (1990–2019)

The trends for *criteria* and *phenomena*, however, indicate a much slower rate of change. The non-plural usage of *criteria* shows a slight fluctuation but remains largely stable, with only a modest increase observed in the final five-year period. Similarly, *phenomena* consistently exhibit the lowest frequency of non-plural usage, remaining near the baseline throughout the three decades. This disparity highlights that while *data* is actively shifting toward

singular usage in standard registers, *criteria* and *phenomena* encounter stronger resistance and largely retain their traditional plural status.

4.2 Variation Across Linguistic Registers

The analysis of register variation reveals distinct patterns in how the non-plural forms of the target nouns are distributed across Newspaper (NEWS), Fiction (FIC), and Magazine (MAG). Tables 2-4 present the raw frequencies for each noun across the three coding categories. Given the substantial number of tokens classified as "unidentifiable," comparisons are best understood by examining the ratio of "non-plural" to "plural" usage within the subset of tokens that exhibited explicit grammatical marking.

Table 2. Frequency of Plural, Non-Plural, and Unidentifiable Tokens of *Data* Across Registers

Register	*Data*		
	Plural	Non-Plural	Unidentifiable
NEWS	90	181	1214
FIC	165	294	2704
MAG	9	50	383
Total	264	525	4301

For the noun *data*, the Fiction register showed a strong tendency toward non-plural usage among identifiable tokens. Specifically, 64.1% of grammatically marked instances (294 out of 459) were coded as non-plural, indicating substantial singularization in narrative contexts (see Table 2). A comparable pattern emerged in the Newspaper register, where 66.8% of identifiable tokens (181 out of 271) occurred in non-plural form. This trend aligns with McAlister's (2016) observation that journalists and other non-academic writers increasingly favor singular *data* to avoid the perceived awkwardness of the plural in everyday language. Example (8) illustrates this usage, where the singular demonstrative *this* and the singular verb *demonstrates* treat *data* as a collective entity:

(8) **This newly released data demonstrates** how aggressively Russia sought to divide Americans by race, religion and ideology. (COCA: 2018: NEWS)

The Magazine register exhibited an even stronger preference for non-plural usage. Despite a smaller sample size, 84.7% of identifiable tokens (50 out of 59) were coded as non-plural, exceeding the proportions observed in both Fiction and Newspaper. Taken together, these findings suggest that the reanalysis of *data* as a non-plural noun is established across all three registers, irrespective of differences in genre or editorial formality. Notably, however, the majority of *data* tokens in all registers lacked overt number marking. The following examples from the Magazine register typify this singular usage, showing *data* agreeing with the singular verb *is* in (9) and contrasting with the traditional plural usage in (10):

(9) For services requiring higher capacity bandwidth, **data is** $25 for an initial one gigabyte and $15 for additional gigabytes. (COCA: 2016: MAG)

(10) But if the Food &; Drug Administration's exposure **data are** correct, only one-third of one cancer risk would be so avoided. (COCA: 1990: MAG)

Table 3. Frequency of Plural, Non-Plural, and Unidentifiable Tokens of *Criteria* Across Registers

Register	*Criteria*		
	Plural	Non-Plural	Unidentifiable
NEWS	322	116	879
FIC	496	60	1158
MAG	76	11	108
Total	894	187	2145

The word *criteria* displayed a different distribution, maintaining a stronger adherence to traditional plural usage. The Newspaper register showed the highest resistance to singularization, with only 26.5% of identifiable tokens (116 out of 438) appearing in non-plural contexts (see Table 3). Magazine followed a similar pattern, with a non-plural rate of 12.6% (11 out of 87). Interestingly, while Fiction had fewer identifiable tokens overall, it showed the lowest rate of singular usage at 10.8% (60 out of 556). This indicates that across all registers, *criteria* is still predominantly treated as a plural noun when grammatical number is explicit. Example (11) demonstrates the minority singular usage with the determiner *that*, while (12) reflects the dominant plural norm:

(11) GM has 17,700 salaried employees who meet **that criteria**, and "most global executives are eligible," the memo said. (COCA: 2018: NEWS)

(12) There are **six criteria** for election under the rules of the Hall of Fame. (COCA: 2016: NEWS)

Table 4. Frequency of Plural, Non-Plural, and Unidentifiable Tokens of *Phenomena* Across Registers

Register	*Phenomena*		
	Plural	Non-Plural	Unidentifiable
NEWS	72	20	128
FIC	440	27	535
MAG	92	13	139
Total	604	60	802

Regarding *phenomena*, the analysis is limited by the smaller sample size of marked tokens. Nevertheless, the trend aligns with *criteria*. Newspaper contained a notable proportion of non-plural usages (21.7%, 20 out of 92 identifiable tokens), whereas Magazine remained highly resistant to the singular form, with only 12.4% (13 out of 105) coded as non-plural (see Table 4). Fiction demonstrated the strongest adherence to the plural norm, with non-plural usage accounting for just 5.8% (27 out of 467 identifiable tokens). The conservative nature of the Magazine register across all three nouns may reflect stricter editorial policies or a stronger adherence to prescriptive norms in semi-academic or formal periodicals, a tendency noted by Brown (2009) regarding the retention of classical

plurals in specific contexts. As shown in (13), singular usage often involves the indefinite article *a*, whereas (14) exemplifies the standard plural form:

(13) You know, I've never been to the rainforest—isn't that **a phenomena**? (COCA: 1994: FIC)

(14) A derivative of Bambi and bingo, **two trendy phenomena** when the company was started in the 1940s, the name Bimbo conjures little more than squishy happiness among Mexican consumers. (COCA: 2002: NEWS)

Overall, the results suggest that *data* has advanced significantly further in its reanalysis as a singular/mass noun, achieving majority status across all three registers. Conversely, *criteria* and *phenomena* largely retain their traditional plural status across all examined registers, with the Newspaper register showing only moderate receptivity to their singular forms.

4.3 Linguistic Patterns of Non-plural Usage

To understand the specific linguistic mechanisms driving the singularization of these nouns, the grammatical environments of all 849 identifiable non-plural tokens were analyzed. As summarized in Table 5, the distribution of grammatical categories reveals a structural dichotomy even before analyzing specific subtypes. While singular verb agreement is a common indicator across the board, the divergence lies in how these nouns are individuated. Specifically, *data* exhibits a notable reliance on partitive structures (15.67%), a syntactic environment typically required for counting mass nouns (e.g., *a piece of*). In sharp contrast, partitive structures are virtually absent for *criteria* (0.00%) and *phenomena* (3.17%). Instead, *criteria* and *phenomena* show a much higher dependency on direct determiners (44.88% and 60.32%, respectively) compared to *data* (23.58%). This macro-level distribution suggests that while *data* is being reanalyzed as a mass noun, *criteria* and *phenomena* are being treated as self-sufficient countable entities.

Table 5. Distribution of Sub-categories in Non-Plural Usage

Sub-category	*Data*	*Criteria*	*Phenomena*
Determiner	137 (23.58%)	92 (44.88%)	38 (60.32%)
Partitive Structure	91 (15.67%)	0 (0.00%)	2 (3.17%)
Pronoun Reference	41 (7.06%)	8 (3.90%)	6 (9.52%)
Singular Verb Agreement	312 (53.70%)	105 (51.22%)	17 (26.98%)
Total	581 (100.00%)	205 (100.00%)	63 (100.00%)

4.3.1 Determiners: The mass vs. count split

The distribution of determiners provides the strongest evidence for this mass-count dichotomy. As detailed in Table 6, mass-selecting determiners are exclusive to *data*, whereas *criteria* and *phenomena* exhibit a high dependency on singular count determiners.

Table 6. Distribution of Determiner Types in Non-Plural Usage

Sub-category	*Data*	*Criteria*	*Phenomena*
Singular Determiner	103 (17.73%)	92 (44.88%)	24 (38.10%)
Mass Determiner	34 (5.85%)	0 (0.00%)	0 (0.00%)
General Determiner	0 (0.00%)	0 (0.00%)	14 (22.22%)
Total	137 (100%)	92 (100%)	38 (100%)

For *data*, 5.85% of the non-plural instances (n=34) were modified by quantifiers strictly restricted to uncountable nouns, such as *much*, *little*, and *less*. This specific collocational pattern, combined with the complete absence of singular count determiners (e.g., *a/an*), confirms that speakers frequently conceptualize *data* as a homogeneous substance similar to information. Although Table 6 shows a presence of singular determiners for *data* (17.73%), these are exclusively singular demonstratives (e.g., *this data*) rather than count-specific articles. This mass-noun conceptualization is evident in (15) and (16), where *data* is modified by *little*, *much*, and *less*, which can only be used with mass nouns:

(15) There is **little data** showing safety has increased, and much [data] showing it is worse. (COCA: 2017: NEWS)

(16) Not only are they behind on follow up data with Exondys 51, they also had the hubris to think they could get accelerated approval with **less data** with Vyondys 53? (COCA: 2019: NEWS).

In contrast, mass determiners were completely absent in the non-plural datasets for *criteria* and *phenomena*. Instead, these words showed a significant reliance on singular count determiners (e.g., *a, an, one,* and *another*). For *criteria*, this category accounted for 44.88% of all non-plural evidence, serving as the second most frequent indicator after verb agreement. This usage pattern indicates that when *criteria* is singularized, it is not viewed as an aggregate mass, but rather as a distinct, countable entity synonymous with *criterion*. Examples (17) and (18) highlight this countability, featuring the indefinite article *a* and the cardinal number *one*:

(17) Administration officials have made avoiding controversy **a key criteria** for selection and have given significant thought about how their choices would be received by Senate Republicans. (COCA: 1994: NEWS)

(18) Halverson describes people break dancing, and he said **one criteria** for new employees is how well they sing. (COCA: 1999: NEWS)

Similarly, *phenomena* exhibits a parallel pattern, with Singular Count Determiners constituting 38.1% of the evidence. The presence of determiners such as *this* and *a* further reinforces its reanalysis as a singular object, as seen in (19):

(19) Ensuring full exposure of **this natural phenomena**, the Central West Astronomical Society established AstroFest to bring together local and international astronomers and enthusiasts to share their love and awe for the heavens. (COCA: 2016: MAG)

4.3.2 Partitive structures: Unitizing the mass

The analysis of partitive structures further reinforces the mass noun status of *data* versus the singular count noun status of *criteria* and *phenomena*. As shown in Table 7, mass partitive structures (e.g., *amount of* and *piece of*) constituted 9.47% (n=55) of the non-plural evidence for *data*. These structures allow speakers to count mass nouns, as in *a piece of furniture* and *a piece of information*.

Table 7. Distribution of Partitive Structures in Non-Plural Usage

Sub-category	*Data*	*Criteria*	*Phenomena*
General Partitive Structure	36 (6.20%)	0 (0.00%)	2 (3.17%)
Mass Partitive Structure	55 (9.47%)	0 (0.00%)	0 (0.00%)
Total	91 (100%)	0 (100%)	2 (100%)

Examples (20) and (21) demonstrate this function, where *data* requires a head noun like *piece* or *scrap* to be counted:

(20) CSAP scores should be viewed as **one important piece of data** in a larger body of evidence. (COCA: 2001: NEWS)

(21) However, a closer look at **that scrap of data** reveals a different picture. (COCA: 2009: FIC).

Conversely, mass partitive structures were absent for *criteria* and *phenomena*. While *data* also appeared with general partitive structures (e.g., *type of* and *kind of*) in 6.20% of cases, *criteria* did not appear with any partitive structures in the retrieved tokens. This suggests that *criteria* and *phenomena* are inherently viewed as countable items once singularized, whereas *data* requires a structural container to be counted.

4.3.3 Semantic shifts

The widespread use of singular agreement with *data* in contemporary English can also be accounted for by two recurrent usage patterns observable in corpus data. The first pattern reflects a conceptualization of *data* as an abstract asset or resource. In this construal, *data* is treated analogously to mass nouns denoting economic, informational, or infrastructural value, such as capital, information, or infrastructure. Rather than referring to discrete observations, *data* is understood as a unified entity that can be owned, managed, stored, monetized, or regulated. This aligns with Rosenberg's (2018, p. 566) historical analysis, which describes the semantic evolution of *data* from rhetorical "givens" to a material entity or "stuff" that functions as a factual resource. This asset-based interpretation is particularly prominent in technological, corporate, and institutional discourse, where singular agreement aligns with notional interpretations of *data* as an indivisible and valuable resource:

(22) In some cases, customers will be charged by **how much data** they consume. (COCA: 2003: MAG)

(23) Part of the problem with prediction is that we actually don't have **much data** on prior El Ni? (COCA: 2016: MAG)

The second pattern involves the reanalysis of *data* as a bounded collective or compound-like noun. In many contemporary contexts, *data* appears in fixed or semi-fixed multiword expressions (e.g., *data analysis, data management,* and *Big Data*), where it functions as the head of an abstract nominal unit denoting a field, process, or domain of activity. In these environments, *data* is construed less as an aggregation of individual *data* points and more as a cohesive conceptual whole. Singular agreement in such cases reflects notional agreement driven by lexicalization and compounding processes, consistent with usage patterns observed for other institutionalized or collective nouns in English:

(24) **Big Data** teaches us what's out there, not what's right. (COCA: 2013: NEWS)

(25) That can sound ominous, but **Big Data** is producing better information, not just more information, about our economy, our health and everything else, because we have better tools for slicing and dicing data, for searching, sifting and sorting through the barrage of keystrokes. (COCA: 2014: MAG).

4.3.4 Ambivalence and mixed usage

Perhaps the most compelling evidence for the divergent grammatical trajectories of *data* lies in the presence of mixed agreement patterns within single utterances. The corpus contains several instances in which speakers alternate between singular and plural realizations of *data* within the same sentence. Such hybrid constructions point to a state of grammatical variability rather than random error, in which competing representations, *data* as a singular mass noun and *data* as a plural count noun, remain simultaneously available to speakers. In example (26), the writer generates a striking syntactic mismatch: *data* is introduced by the singular demonstrative *that*, yet it is immediately paired with the plural verb *confirm* rather than the singular *confirms*.

(26) **That data confirm** the overall reduction in collisions, but add a few interesting twists, including that more serious "reportable" crashes increased slightly and less-serious "property damage only" collisions declined by nearly 30 percent. (COCA: 2016: NEWS).

Conversely, other examples exhibit a mismatch between singular or mass-oriented quantification and plural verbal agreement, revealing tension between semantic construal and inherited grammatical conventions. In (27), the speaker juxtaposes the existential singular *there is a lot of data* with the plural subject-verb agreement *The data show*:

(27) Doctors have performed the procedure on more than 1,000 breast cancer patients since 1985, and "there **is a lot of data** now," he said. **The data show** that about 20 percent of patients with late-stage cancer who received the procedure survive at least five years without a relapse, compared to less than 10 percent who received standard chemotherapy, Champlin said. (COCA: 1994: NEWS).

These mixed patterns suggest that the reanalysis of *data* as a singular mass noun is well advanced but not yet categorical. As Allan (1980) argues, countability is not always a rigid property of the noun itself but can be a feature of the noun phrase determined by context. Rather than reflecting individual inconsistency or error, such variation appears to signal an intermediate stage of grammatical change, in which both singular and plural agreement patterns coexist even within individual idiolects. From a usage-based perspective, these hybrid

constructions provide firm evidence that agreement with *data* is governed by notional factors and ongoing reanalysis rather than by stable morphological rules.

Synthesizing the diachronic trends, register variations, and linguistic patterns, this study confirms that the singularization of these loanwords follows two distinct and divergent pathways. Quantitatively, *data* exhibits a rapid and robust shift towards singular usage, particularly within Newspaper and Fiction. In contrast, *criteria* and *phenomena* demonstrate a significantly slower rate of change and remain largely plural. Qualitatively, this divergence is structurally reinforced: for *data*, the convergence of mass-selecting determiners, unitizing partitives, and semantic conceptualizations as an abstract asset or compound entity strongly corroborates its evolution towards an uncountable mass category. In contrast, *criteria* and *phenomena* lack these characteristics, instead evolving towards singular countable categories through their exclusive reliance on singular count determiners. Finally, the persistence of mixed agreement patterns within single utterances suggests that while the reanalysis of *data* is advanced, it remains a fluid, ongoing process characterized by transitional variation rather than a completed categorical shift.

5. Conclusion

This study demonstrates that *data*, *criteria*, and *phenomena*, despite sharing analogous classical origins as foreign plurals, are undergoing distinct and divergent grammatical developments in contemporary English. The findings reveal that the singularization of these nouns is not a monolithic process but rather splits into two typologically different pathways. First, the diachronic corpus evidence confirms that *data* has experienced a sustained shift toward singular usage since the mid-20th century, reaching near parity with plural forms by the late 2010s. In contrast, *criteria* and *phenomena* exhibit a much slower rate of change, with traditional plural agreement remaining the dominant norm over time. Register-based analysis further highlights this divergence. Non-plural *data* is especially prevalent in Newspaper and Fiction, suggesting that communicative accessibility and the conceptualization of *data* as an abstract asset facilitate its reanalysis. By contrast, *criteria* and *phenomena* maintain plural agreement across registers, indicating stronger resistance to change, likely due to their lower frequency and stronger association with formal or academic contexts (Brown 2009).

Second, and perhaps more critically, qualitative analysis reveals that this divergence is structural rather than superficial. *Data* is predominantly evolving into a mass-noun, evidenced by its exclusive compatibility with mass-selecting determiners, unitizing partitive structures, and asset-oriented semantics (Rosenberg 2018). Conversely, singular uses of *criteria* and *phenomena* align with singular count-noun morphology, appearing with indefinite articles and numerical markers and functioning as informal substitutes for *criterion* and *phenomenon*. This disparity aligns with Barner and Snedeker's (2005) theory of cognitive individuation, suggesting that *criteria* and *phenomena* are perceived as distinct, bounded entities requiring count syntax, whereas *data* is construed as an unbounded aggregate. Overall, these findings challenge the prescriptive view of these words as strictly plural. Instead, grammatical change in this domain appears to be driven by semantic necessity and frequency effects (Bybee 2006), with *data* emerging as a collective mass noun while *criteria* and *phenomena* showing incipient signs of becoming singular count nouns. The persistence of mixed agreement patterns involving *data* further indicates that this reanalysis, albeit advanced, remains a fluid and ongoing process.

Needless to say, this study is not without limitations. The reliance on the COCA corpus, while extensive, restricts the analysis to American English and specific written registers. Future research could extend the analysis beyond American English by incorporating other varieties (e.g., British, Australian, or World Englishes) in order

to determine whether the divergent trajectories observed here are globally stable or variety-specific. Furthermore, a notable portion of the dataset was classified as "unidentifiable" due to the lack of explicit grammatical marking, suggesting that in many contexts, the number distinction for these nouns may be neutralized or pragmatically irrelevant.

Despite these limitations, the study offers practical implications for both research and pedagogy. Future inquiries would benefit from extending this corpus-based approach to other varieties of English to ascertain whether these evolutionary paths are universal or regionally specific. Additionally, incorporating psycholinguistic experiments, including acceptability judgment tasks, could further validate whether these shifts are rooted in cognitive processing or merely surface-level conventions. Pedagogically, the findings suggest that English language instruction should acknowledge this fluidity rather than enforcing rigid prescriptive norms. Specifically, treating *data* as a mass noun in general contexts may better equip learners to navigate contemporary usage, while maintaining traditional distinctions for *criteria* and *phenomena* remains advisable for formal academic writing.

References

Allan, K. 1980. Nouns and countability. *Language* 56(3), 541-567.

Ball, C. R. 1928. English or Latin plurals for Anglicized Latin nouns? *American Speech* 3(4), 291-325.

Barner, D. and J. Snedeker. 2005. Quantity judgments and individuation: Evidence that mass nouns count. *Cognition* 97(1), 41-66.

Bordignon, F. and M. Maisonobe. 2022. Researchers and their data: A study based on the use of the word *data* in scholarly articles. *Quantitative Science Studies* 3(4), 1156-1178.

Brown, M. 2009. Of words and their plurals I sing: Latin and Greek plurals and their usage in English. *Journal of the Georgia Philological Association* 4, 112-131.

Bybee, J. L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4), 711-733.

Chierchia, G. 2010. Mass nouns, vagueness and semantic variation. *Synthese* 174, 99-149.

Drożdż, G. 2020. New insights into English count and mass noun – the cognitive grammar perspective. *English Language and Linguistics* 24(4), 833-854.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159-174.

Merriam-Webster. n.d. Can 'criteria' ever be singular?: It depends on what you base your decision. In *Merriam-Webster.com dictionary*. Available online at https://www.merriam-webster.com/grammar/criteria-vs-criterion-singular-plural-grammar

McAlister, V. C. 2016. Datum isn't; data are. *Canadian Journal of Surgery* 59(4), 220-221.

Pound, L. 1919. The pluralization of Latin loan-words in present-day American speech. *The Classical Journal* 15(3), 163-168.

Rosenberg, D. 2018. Data as word. *Historical Studies in the Natural Sciences* 48(5), 557-567.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary