arXiv:submit/6924343 [math.LO] 25 Oct 2025

Redefining Natural Language: From $\aleph_0$ to $2^{\aleph_0}$

Ramon Padilla-Reyes Charles Woodrum

October 25, 2025

Abstract

We argue that language should organically permit expressions of countably infinite length, and thus that the size of our language is $2^{\aleph_0}$. Abstract formalisms of language have brought much technical understanding to the fields of syntax, semantics, and pragmatics. In this paper we explore, through an abundance of proofs, the size of all possible human utterances—be it a specific language or all languages. We show that the set of all possible sentences has size $\aleph_0$ unless either the length of the sentence can be infinite or the set of all possible words is uncountably infinite. This is true even under maximally permissive grammatical rules and expansive vocabularies. The fact that our arguably uncountable world must be described by a countable system of finite sentences may have interesting implications for philosophy and metaphysics, but that will not be addressed here. Our aim is to give a definitive exploration of the possible cardinalities of language, depending on the size of utterances allowed by the axioms of the language.

1 Introduction

The study of language has intrigued scholars across various disciplines for centuries. Linguistics, philosophy, and mathematics have each contributed unique perspectives. Yet the intersection of these fields presents untapped potential for new results. This paper focuses on one such intersection: the mathematical cardinality of language. Traditionally, language has been viewed as countable, bound by finiteness in grammar and vocabulary. We challenge that paradigm, arguing that language can be uncountable under specific, principled idealizations.

Karp's infinitary logic does more than expand expressive power; it hints at a reevaluation of language's potential capacity [1]. The infinitary language $L_{\omega_1,\omega}$ permits countably infinite conjunctions and disjunctions, giving a disciplined way to move from finite to infinite constructions and to challenge the default attribution of cardinality $\aleph_0$. Our hypothesis is: under theoretical conditions that allow sentences of infinite length or vocabularies of uncountable size, the cardinality of (idealized) language reaches the continuum $2^{\aleph_0}$.

Adapting this rationale to natural language, we illuminate how standard ideas from infinitary logic travel into linguistics. This is not merely a technical extension; it reconceptualizes the bounds of linguistic possibility.

The traditional view takes language to have cardinality $\aleph_0$ [3, 10]. That standpoint aligns with finite grammars over countable vocabularies. But if we lift the finiteness constraints on vocabulary or sentence length, the usual cardinality assumptions shift. Historically regarded as countably infinite ($\aleph_0$), the cardinality of natural language may, under these idealizations, approach the continuum ($2^{\aleph_0}$), cf. [2]. We therefore divide the paper into two sections. In Section I we set out scenarios where language moves beyond these boundaries. In Section II we examine more closely the case of sentences of infinite length and the role of uncountable vocabularies.

We support the claim with a series of mathematical constructions (including a diagonal argument). The upshot is a clean accounting: with finite-length sentences over a countable vocabulary, the set of sentences is countable; allow countably infinite length or an uncountable vocabulary, and it is uncountable.

2 Section I

In Section I we present scenarios where language transcends the default countable picture—particularly when permitting sentences of infinite length or uncountable vocabularies. Cantor's diagonalization then shows how, under those conditions, language attains $2^{\aleph_0}$ [2].
Convention. Unless stated otherwise, "sentence length" means number of words; "sentence character length" refers to characters.

Theorem 2.1. *Grammatical sentences can have unbounded length.*

Proof. Consider the schema *"The mother of the mother of the mother ... fell."*

Vocabulary Size	Sentence Word Length	Size of Language
Finite	Finite	$\aleph_0$
$\aleph_0$	Finite	$\aleph_0$
$2^{\aleph_0}$	Finite	$2^{\aleph_0}$
Finite	$\aleph_0$	$2^{\aleph_0}$
$\aleph_0$	$\aleph_0$	$2^{\aleph_0}$
$2^{\aleph_0}$	$\aleph_0$	$2^{\aleph_0}$

Table 1: Vocabulary/length regimes and the resulting size of (idealized) language. "Finite" here means unbounded but finite—arbitrarily large, yet not infinite.

For any $k \in \mathbb{N}$, the phrase referring to the k-th generation is grammatical. Thus for every n there is a grammatical sentence longer than n, so grammatical length is unbounded. Moreover, the k-th generation index k ranges over $\mathbb{N}$, yielding countably many such sentences. □

Theorem 2.2. *With a finite vocabulary and unbounded but finite sentence length, the set of sentences is countable.*

Proof. Let V be the vocabulary with $|V| = N \in \mathbb{N}$. Let G be the set of all grammatical sentences under the maximally permissive finite-length rule (any finite word sequence is grammatical). For each length k, let S_k be the set of length-k sentences. Then

$$|S_0| = 1, \quad |S_1| = N, \quad |S_2| = N^2, \ \ldots, \ |S_k| = N^k.$$

Hence

$$G = \bigcup_{k=0}^{\infty} S_k,$$

a countable union of finite sets, so $|G| = |\mathbb{N}|$. For any more restrictive grammar r, $G_r \subseteq G$, whence $|G_r| \leq |G| = \aleph_0$. By Theorem 2.1, we also know G_r is infinite, hence countably infinite in this regime. □

If the vocabulary is uncountable, sentences of (even) length 1 already witness uncountability, so $|\text{Language}| = 2^{\aleph_0}$. By contrast, the set of all human words actually *used* (excluding numerals) is finite, in principle listable; adjoining $\mathbb{N}$ yields a countably infinite vocabulary; adjoining $\mathbb{Q}$ keeps it countable; adjoining $\mathbb{R} \setminus \mathbb{Q}$ pushes beyond countable—but then expressing single reals as *one word* requires allowing countably infinite *character*-length encodings.

Theorem 2.3. *With a countably infinite vocabulary and unbounded but finite sentence length, the set of sentences is countable.*

Proof. Let $|V| = |\mathbb{N}|$ and G be as above (maximally permissive over finite lengths). For $k \in \mathbb{N}$, $S_k = V^k$ (Cartesian k-fold product). Since a Cartesian product of finitely many countable sets is countable, $|S_k| = \aleph_0$ for all k.

Thus

$$G = \bigcup_{k=0}^{\infty} S_k$$

is a countable union of countable sets, hence countable. Any more restrictive grammar $G_r \subseteq G$ remains countable in this regime. □

It is useful to motivate admitting infinite sentences. Consider the decimal expansion of $\pi = 3.14159\ldots$ (nonrepeating, nonterminating). One can describe it with a sentence schema: *"The first digit after 3 is 1, then 4, then 1, ..."* No finite sentence captures the entirety. If language is allowed to express the full mathematical thought, infinite sentences are a coherent idealization—just as set theory counts nonterminating decimals as legitimate objects [2]. Likewise, if the world's description truly requires uncountable detail, full descriptions will outrun any finite sentence-length bound.

Theorem 2.4. *With finite vocabulary but countably infinite phrases (and hence countably infinite sentences), the set of sentences is uncountable.*

Proof 1.

Proof. Consider sentences of the form *"The first digit is b_1 and the second digit is b_2 and the third digit is b_3 and ..."* with $b_j \in \{0, 1, \ldots, 9\}$. Such (infinite) sentences are in bijection with $[0, 1]$ in base 10, hence uncountably many.

To avoid numerals in-sentence, set up a base-2 variant using two proper names (e.g. $0 \leftrightarrow$ Alice, $1 \leftrightarrow$ Bob) and form *"b_1 thinks that b_2 thinks that b_3 thinks that ... the grandmother fell."* Mapping $0.b_1b_2b_3\ldots$ to that sentence gives a bijection with $[0, 1]$. Thus some grammar r that permits countably infinite sentences yields a set G_1 with $|G_1| = 2^{\aleph_0}$.

To compute an upper bound in three cases, let G be the maximally permissive grammar (finite or countably infinite length; any word sequence grammatical).

Case 1: Finite vocabulary. Here $|S_k| = N^k$ for finite k, and S_∞ (the countably infinite sequences over an N-ary alphabet) has cardinality $N^{\aleph_0} = 2^{\aleph_0}$ (the set of functions $\mathbb{N} \to \{1, \ldots, N\}$). Hence $|G| = 2^{\aleph_0}$ and $2^{\aleph_0} = |G_1| \leq |G_r| \leq |G| = 2^{\aleph_0}$, so $|G_r| = 2^{\aleph_0}$.

Case 2: Countable vocabulary. Each finite $S_k = V^k$ is countable, but $S_\infty = V^{\mathbb{N}}$ (the set of infinite sequences of vocabulary items) has cardinality $|\mathbb{N}^{\mathbb{N}}| = 2^{\aleph_0}$. Equivalently, a direct diagonal argument shows uncountability. Thus again $|G| = 2^{\aleph_0}$ and $|G_r| = 2^{\aleph_0}$.

Case 3: Uncountable vocabulary. Let $|V| = 2^{\aleph_0}$. Then for each finite k, $|S_k| = |V|^k = 2^{\aleph_0}$; and $S_\infty = V^{\mathbb{N}}$ also has cardinality $(2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0}$. *(Correction: a finite Cartesian product of an uncountable set is* ***uncountable****, not countable.)* Hence $|G| = 2^{\aleph_0}$ and again $|G_r| = 2^{\aleph_0}$. □

3 Section II

Universal Algebra and Natural Language

Definition 3.1 (Universal Algebra)**.** A universal algebra $\mathcal{A} = (A, F)$ has nonempty A and a set F of n-ary operations on A, each $f : A^n \to A$.

Definition 3.2 (Sentence Equality)**.** Two sentences $s_1, s_2 \in S$ are equal iff for all i, the i-th word of s_1 equals the i-th word of s_2.

Theorem 3.3. *Within a natural-language universal algebra with an adjective set A of arbitrary size, there exists a set S such that the algebra generates an infinite subset of S composed of (alongside finite) sentences of infinite length.*

Proof. Assume an adjective-stacking operation. Repeated application yields arbitrarily long strings; admitting countably many applications yields infinite length. Distinct choices of adjective (by our equality convention) produce distinct sentences. Hence infinitely many infinite-length sentences. □

Any observed boundedness in practice is due to cognitive and social constraints, not the algebra itself. The situation is analogous to nonterminating decimals: legitimate mathematically, truncated in use.

4 Sentences with Infinite Adjective Stacking

1. The black black black black black black black . . . cat leaped.
2. The grumpy old grumpy old grumpy old grumpy old . . . cat leaped.
3. The fat lazy fat lazy fat lazy fat lazy . . . cat leaped.
4. The grey happy grey happy grey happy grey happy . . . cat leaped.
5. The furry slim furry slim furry slim furry slim . . . cat leaped.
6. The tiny black tiny black tiny black tiny black . . . cat leaped.
7. The old fat old fat old fat old fat old fat . . . cat leaped.
8. The lazy grey lazy grey lazy grey lazy grey lazy grey . . . cat leaped.
9. The happy furry happy furry happy furry happy furry happy furry . . . cat leaped.
10. The slim tiny slim tiny slim tiny slim tiny slim tiny slim . . . cat leaped.

4.1 Binary Representation

- Noun (N) → `10`
- Adjective (Adj) → `01`
- Verb (V) → `00`

Words:

cat (N) : `10`
black (Adj) : `01`
grumpy (Adj) : `01`
old (Adj) : `01`
fat (Adj) : `01`
lazy (Adj) : `01`
grey (Adj) : `01`
happy (Adj) : `01`
furry (Adj) : `01`
slim (Adj) : `01`
tiny (Adj) : `01`
leaped (V) : `00`

Original Binary	Adjective	Flipped Binary	New Adjective
01 0100	black	01 1011	fat
01 0110	old	01 1001	furry
01 0111	fat	01 1000	grumpy
01 1010	happy	01 0101	tiny
01 1011	furry	01 0100	old
01 1001	grey	01 0110	slim
01 1000	lazy	01 0111	black
01 1100	grumpy	01 0011	lazy
01 1101	slim	01 0010	grey
01 1110	tiny	01 0001	happy

Not flipped:

1. The 01_0100$_1$ 01_0100 01_0100 01_0100 01_0100 01_0100 . . . 10_1001 00_0010.
2. The 01_0101 01_0110$_2$ 01_0101 01_0110 01_0101 01_0110 . . . 10_1001 00_0010.
3. The 01_0111 01_1000 01_0111$_3$ 01_1000 01_0111 01_1000 . . . 10_1001 00_0010.
4. The 01_1001 01_1010 01_1001 01_1010$_4$ 01_1001 01_1010 . . . 10_1001 00_0010.
5. The 01_1011 01_1100 01_1011 01_1100 01_1011$_5$ 01_1100 . . . 10_1001 00_0010.

6. The 01_1101 01_0100 01_1101 01_0100 01_1101 01_0100$_6$ 01_0100 . . . 10_1001 00_0010.
7. The 01_0110 01_0111 01_0110 01_0111 01_0110 01_0111 01_0110$_7$ 01_0111 . . . 10_1001 00_0010.
8. The 01_1000 01_1001 01_1000 01_1001 01_1000 01_1001 01_1000 01_1001$_8$ 01_1001 . . . 10_1001 00_0010.
9. The 01_1010 01_1011 01_1010 01_1011 01_1010 01_1011 01_1010 01_1011 01_1010$_9$ 01_1011 . . . 10_1001 00_0010.
10. The 01_1100 01_1101 01_1100 01_1101 01_1100 01_1101 01_1100 01_1101 01_1100 01_1101$_{10}$ 01_1101 . . . 10_1001 00_0010.

Diagonal:

01 0100   01 0110   01 0111   01 1010   01 1011   . . .

Flip (keeping the `01` adjective prefix fixed):

01 1011   01 1001   01 1000   01 0101   01 0100   . . .

4.2 Decoding the Sentence

Original Binary	Adjective	Flipped Binary	New Adjective
01 0100	black	01 1011	fat
01 0110	old	01 1001	furry
01 0111	fat	01 1000	grumpy
01 1010	happy	01 0101	tiny
01 1011	furry	01 0100	old
01 1001	grey	01 0110	slim
01 1000	lazy	01 0111	black
01 1100	grumpy	01 0011	lazy
01 1101	slim	01 0010	grey
01 1110	tiny	01 0001	happy

Using the flipped sequence, the n-th adjective differs from the n-th adjective in row n, ensuring the new sentence is not in the list.

1. The **fat** **black black black black** . . . 10_1001 00_0010.
2. The **old** **furry** **old old old** . . . 10_1001 00_0010.
3. The **fat lazy** **grumpy** **lazy fat** . . . 10_1001 00_0010.
4. The **grey happy grey** **tiny** **grey** . . . 10_1001 00_0010.
5. The **furry grumpy furry grumpy** **old** . . . 10_1001 00_0010.
6. The **slim black slim black slim** **slim** . . . 10_1001 00_0010.

7. The **old fat old fat old fat black** ... 10_1001 00_0010.
8. The **lazy grey lazy grey lazy grey lazy lazy** ... 10_1001 00_0010.
9. The **happy furry happy furry happy furry happy furry grey** ... 10_1001 00_0010.
10. The **grumpy slim grumpy slim grumpy slim grumpy slim grumpy happy** ... 10_1001 00_0010.

Thus the diagonal yields a new (infinite) sentence:

The fat furry grumpy tiny old slim black lazy grey happy ... cat leaped.

Definition 4.1 (Universal Algebra)**.** A *universal algebra* is a structure $\mathcal{A} = (A, F)$ with carrier A and operations F, closed under F.

Definition 4.2 (Sentence Equality in Natural Language)**.** Two sentences are equal iff they have the same word sequence.

5 The Case for Infinite Sentences

Lemma 5.1 (Existence of Infinite Sentences)**.** *In a natural-language algebra* $\mathcal{NLA} = (A, F)$ *with an adjective-stacking operation, there are infinitely many sentences of infinite length.*

Proof. Iterating the stacking operation ad infinitum yields countably infinite chains; distinct choices yield distinct sentences. □

To address completeness objections: (i) *The black cat fell.* (finite); (ii) *The big black Adj Adj Adj...* (incomplete); (iii) *The cat that fell is big black Adj Adj Adj...* (syntactically complete even if the modifier sequence does not terminate).

6 Natural Language Recursion and $2^{\aleph_0}$

6.1 The Recursive Grammatical Structure

Definition 6.1 (Recursive Grammar G)**.** We take the familiar template $[SP\ [NP]\ [VP\ [NP]]]$ (cf. Chomsky [3]); for our purposes we treat it as context-free.

Definition 6.2 (SP, NP, VP)**.** $SP = NP\ VP$. Let NP range over a finite list (e.g. John, Mary, Alice, Bob) for illustrations; VP over transitive predicates (e.g. loves [NP], hates [NP], knows [NP], sees [NP]).

7 Definition of the Context-Free Grammar

Definition 7.1 (CFG for [SP [NP] [VP [NP]]]). $G = (V, \Sigma, R, S)$ with

- $V = \{S, NP, VP, N', N, Adj, D, V\}$,
- $\Sigma = \{\text{The, neighbors, big, black, fat, grey, cat, jumped}\}$,
- R:

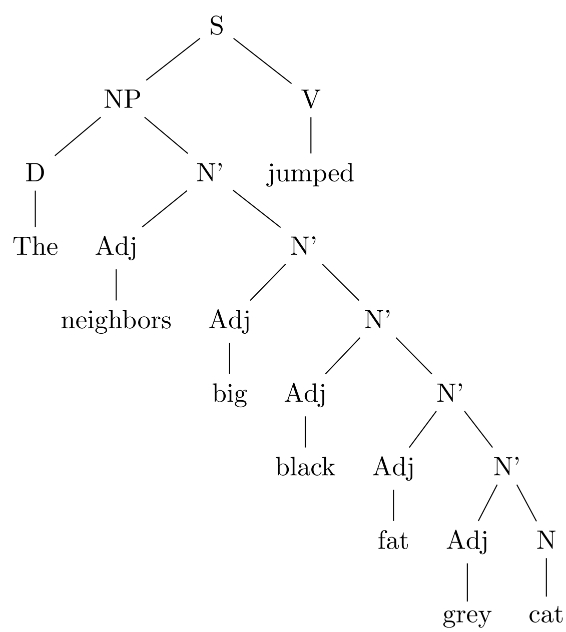
$$
\begin{aligned}
S &\to \text{NP VP} \\
NP &\to \text{D N'} \\
N' &\to \text{Adj N'} \mid \text{N} \\
N &\to \text{cat} \mid \text{dog} \mid \text{bird} \\
Adj &\to \text{neighbors} \mid \text{big} \mid \text{black} \mid \text{fat} \mid \text{grey} \\
D &\to \text{The} \\
VP &\to \text{V} \mid \text{V NP} \\
V &\to \text{jumped} \mid \text{chased}
\end{aligned}
$$

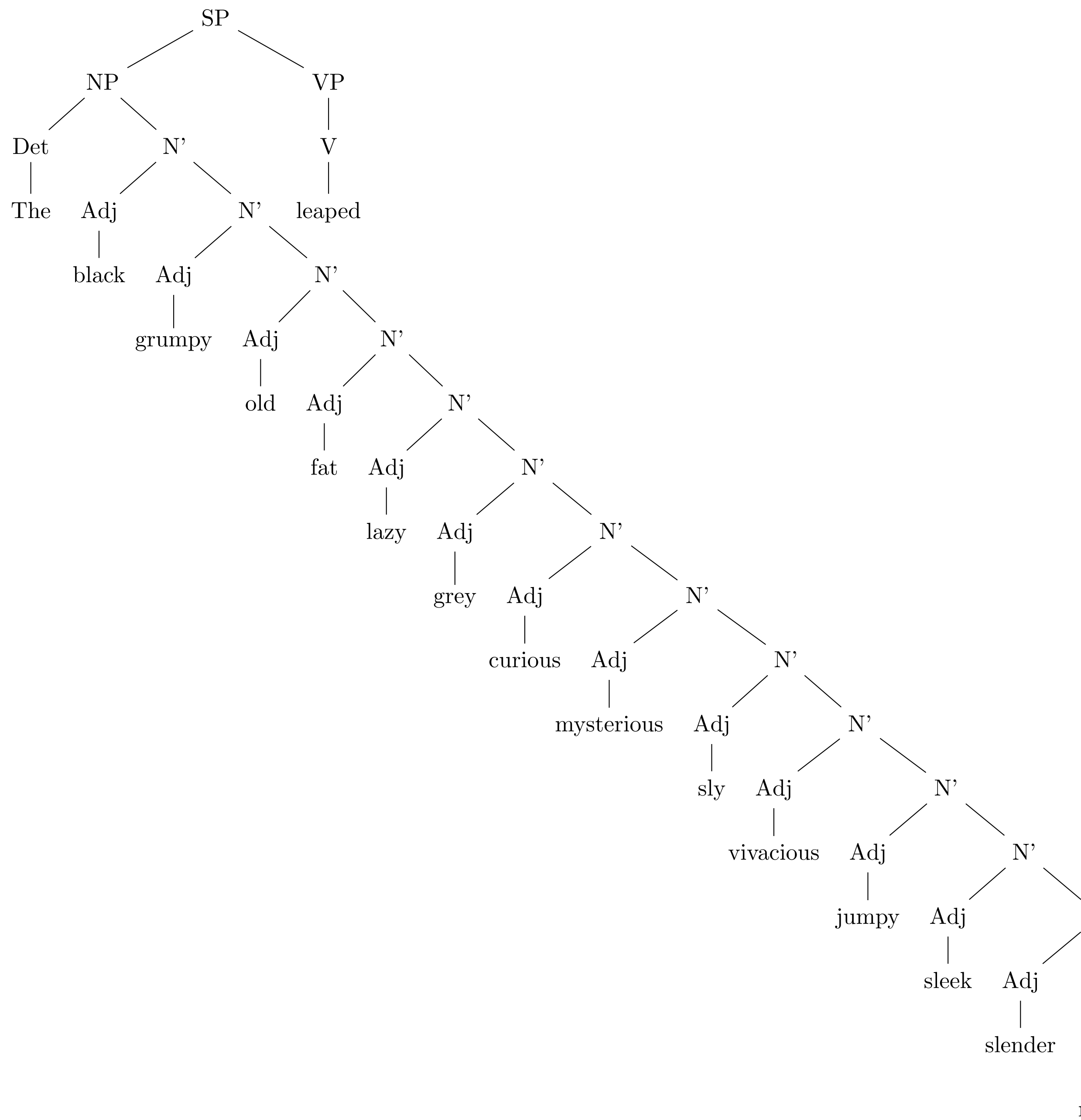
S is the start symbol.

8 Infinite Adjective Stacking

1. The cat jumped.
2. The grey cat jumped.
3. The fat grey cat jumped.
4. The black fat grey cat jumped.
5. The big black fat grey cat jumped.
6. The the neighbors' big black fat grey cat jumped.
7. The $\text{Adj}_1\text{Adj}_2\cdots\text{Adj}_{n+1}$ cat jumped.

A tree illustrating unbounded stacking:


S
NP
V
D
N'
jumped
The
Adj
N'
neighbors
Adj
N'
big
Adj
N'
black
Adj
N'
fat
Adj
N
grey
cat


SP
NP
VP
Det
N'
V
The
Adj
N'
leaped
black
Adj
N'
grumpy
Adj
N'
old
Adj
N'
fat
Adj
N'
lazy
Adj
N'
grey
Adj
N'
curious
Adj
N'
mysterious
Adj
N'
sly
Adj
N'
vivacious
Adj
N'
jumpy
Adj
sleek
Adj
slender

8.1 Encoding Adjective Stacking

We define a binary encoding B for phrases, with disjoint prefixes for categories, and unique codes (finite or infinite) for lexical items. To ensure the diagonal construction yields a new adjective, we fix the category prefix for adjectives and flip the payload bits to another valid (distinct) adjective code.

Remark. The code design avoids overlap, ensuring a bijective correspondence on the designated set of well-formed codes.

8.2 Decoding Procedure

Define a decoding map D from the set of well-formed codes to G that inverts B on its codomain. Then B is injective and D witnesses surjectivity *onto the well-formed codes.*

9 Properties of the Encoding Function

9.1 Injectivity of B

Lemma 9.1. *The encoding function* $B : G \to \{0,1\}^*$ *is injective on its codomain.*

Proof. By construction, distinct tokens/structures receive distinct non-overlapping codes. □

9.2 Surjectivity of B (on valid codes)

Lemma 9.2. *Every valid code in the designated code set decodes to some sentence of* G*; hence* B *is surjective onto that set and* D *is its inverse there.*

Proof. Immediate from the definition of the code set (well-formed codes) and the deterministic decoding scheme. □

9.3 Bijection

Corollary 9.3. B *is a bijection between sentences in* G *and well-formed codes.*

10 Main Result: Uncountability of G-sentences

10.1 Ensuring Syntactic Validity in Diagonalization

We stipulate "fallback patterns" so that diagonal flips remain within the well-formed code set, hence decode to grammatical sentences.

10.2 Cantor's Diagonalization

Theorem 10.1 (Uncountability of sentences in G). *The set of sentences generated using G with countably infinite length admitted is uncountable.*

Proof. Assume a countable listing of infinite codes for G-sentences; flip the i-th payload bit of the i-th adjective (keeping the adjective prefix) to obtain a new valid code differing at position i. This decodes to a new G-sentence not in the list. □

11 Encoding Natural Language using Decimal Numbers

Assign digits to adjectives and form infinite sequences inside a fixed sentence schema (*"The $a_1 a_2 a_3 \ldots$ cat jumped."*). A diagonal construction over any putative enumeration again yields a new sequence; hence uncountably many such sentences.

12 Uncountability in Natural Language

Theorem 12.1 (Unbounded Modifiability ⇒ Uncountability). *If a language L has a construct that can be extended indefinitely (unbounded modifiability) and infinite extensions are admitted as single sentences, then L generates uncountably many sentences.*

Proof. Diagonalization over infinite modifier sequences as above. □

13 Another Argument in Favor of Infinite Sentences

Completeness considerations (e.g. for π) motivate countably infinite sentences: *"The ratio of a circle's circumference to its diameter is 3.14159. . ."* (fixed: "its", not "it's"). Also, infinite *lists* of finite sentences (e.g. one per digit) can be conjunctivized into a single infinite sentence when conjunction is allowed to iterate transfinitely in the idealization.

14 Conclusion, Theoretical and Practical Implications

Two points. (i) Definite descriptions can be infinitely elaborated, extending classical theories of reference [11, 12, 13]. (ii) Allowing infinite sentences makes the set $\mathcal{L}$ of sentences uncountable—of size $2^{\aleph_0}$—with implications for linguistics, philosophy, and computation. The picture parallels how real numbers extend beyond rationals: an “irrational” (i.e. uncountable) layer of language beyond the finite.

References

[1] Carol R. Karp. Languages with Expressions of Infinite Length. *Studies in Logic and the Foundations of Mathematics, Vol. 60*, North-Holland, 1964.

[2] Georg Cantor. Über eine elementare Frage der Mannigfaltigkeitslehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 1891.

[3] Noam Chomsky. *Syntactic Structures*. Mouton, 1957.

[4] Richard Montague. Universal Grammar. *Theoria* 36(3):373–398, 1970.

[5] Yehoshua Bar-Hillel. A Quasi-Arithmetical Notation for Syntactic Description. *Language* 29(1):47–58, 1953.

[6] David Lewis. General Semantics. *Synthese* 22(1–2):18–67, 1970.

[7] Alfred Tarski. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* 1:261–405, 1936.

[8] Herbert B. Enderton. *Elements of Set Theory*. Academic Press, 1977.

[9] Edward L. Keenan and Jonathan Stavi. A Semantic Characterization of Natural Language Determiners. *Linguistics and Philosophy* 9(3):253–326, 1986.

[10] Barbara H. Partee. Quantification in Natural Languages. In Shalom Lappin (ed.), *The Handbook of Contemporary Semantic Theory*, Blackwell, 1995, pp. 487–517.

[11] Bertrand Russell. On Denoting. *Mind* 14(56):479–493, 1905.

[12] P. F. Strawson. On Referring. *Mind* 59(235):320–344, 1950.

[13] Keith Donnellan. Reference and Definite Descriptions. *Philosophical Review* 75(3):281–304, 1966.