DNA as a Language: A Linguistic Comparison with Human Spoken and Written Systems

Jack Cohen* and Barak Akabayov

Department of Chemistry and Data Science Research Center, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel

Received: Feb 08, 2026 **Revised**: Jun 04, 2026 **Just Accepted Online**: Jun 11, 2026 **Published**: Xxx

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record.

Please cite this article as:

J. Cohen, B. Akabayov (2026) DNA as a Language: A Linguistic Comparison with Human Spoken and Written Systems. **Substantia**. *Just Accepted*. DOI: 10.36253/Substantia-3980

Accepted Manuscript

Abstract

The traditional metaphor of DNA as a language is examined through systematic analysis, using formal linguistic categories rather than relying on heuristic analogies. By exploring aspects of phonology, morphology, syntax, semantics, pragmatics, and diachrony, the study investigates the structural parallels between genetic sequences and human spoken and written languages, while also identifying the fundamental limitations of this comparison. The analysis indicates that DNA displays distinct symbolic units, a combinatorial organization, strict ordering constraints, context dependence, and historical evolution. These characteristics resemble written language more closely than they do speech. However, DNA fundamentally differs from human language in several ways: it lacks intentionality, reference, generativity,

and representational meaning. Genetic “meaning” is defined by its operational and causal roles, focusing on biochemical functions rather than semantic interpretations. This study emphasizes that DNA should not be viewed as a literal language but rather as a system of symbols grounded in chemistry. The comparison with linguistics is most useful for establishing distinctions rather than for expanding metaphorical interpretations.

Keywords: DNA language, morphemes, codons, gene expression, linguistics

The metaphor of DNA as a “language” has circulated in molecular biology since the mid-twentieth century, often as a heuristic for explaining how genetic sequences encode and transmit information. When examined through the formal methodology of linguistics rather than metaphor alone, this comparison becomes both more precise and more constrained. Linguistics offers a multi-level analytical framework—encompassing phonology, morphology, syntax, semantics, pragmatics, and diachrony—that allows for systematic comparison between symbol systems. Applying these categories to DNA clarifies where the analogy with human spoken and written languages is structurally informative and where it breaks down in principle.

Units of Expression: Phonology and Orthography

Human spoken languages are organized around phonology: a finite inventory of discrete, contrastive sound units (phonemes) that are meaningless in isolation but meaningful in combination [1]. Written language introduces an orthographic layer, in which visual symbols represent phonological units or morphemes depending on the writing system [2].

DNA lacks phonology in the acoustic sense, but it does possess a discrete symbolic inventory analogous to an orthography. The four nucleotide bases—adenine, thymine, cytosine, and guanine—form a finite set of contrastive units whose sequence determines downstream

Accepted Manuscript

effects in protein synthesis [3]. Unlike phonemes, however, nucleotides are not abstract cognitive categories; they are chemically instantiated entities whose physical properties are inseparable from their informational role [4].

In this respect, DNA more closely resembles written language than speech, yet the analogy remains partial. Orthographic symbols conventionally stand for sounds or meanings external to the script, whereas DNA bases directly participate in the biochemical processes that "read" them. The symbol and its material realization cannot be cleanly separated, a point emphasized in discussions of the symbol–matter relationship in biology [5]. For instance, in the human β-globin gene, a single nucleotide substitution (A→T) changes the DNA codon from GAG to GTG. This alteration results in the incorporation of a different amino acid—valine instead of glutamic acid—into the hemoglobin protein, leading to the development of sickle-cell anemia [6]. This example illustrates that, much as changing a single letter in a word can significantly alter the "meaning" of a word, altering a single nucleotide can significantly change the "meaning" in the language of genetics.

Morphology: From Morphemes to Codons and Genes

Morphology in linguistics concerns the internal structure of words and the smallest meaning-bearing units, or morphemes. These units combine according to systematic rules and often productively generate new lexical items [7].

DNA exhibits a comparable intermediate level of organization. Codons—triplets of nucleotides—function as minimal functional units, each specifying an amino acid or a termination signal in the genetic code [8]. In this sense, codons resemble morphemes more closely than letters, as they map systematically to biochemical outputs. Genes, in turn,

Accepted Manuscript

resemble complex lexical items or fixed expressions, with internal organization that includes coding regions, non-coding regions, and regulatory elements [4].

The crucial divergence lies in productivity and agency. Human speakers can generate novel morphological forms intentionally and interpret them on the fly. Genetic systems cannot invent new codon–amino acid correspondences or morphological rules; innovation occurs only through mutation, recombination, and selection, processes that are blind rather than generative in the linguistic sense [9]. For example, the codon "ATG" in DNA corresponds to the amino acid methionine and acts as a universal "start" signal for protein synthesis. This codon can be considered a morpheme-like unit of meaning in DNA because its presence at the beginning of a gene indicates the initiation of translation. A complete gene, such as the one encoding DNA primase, can be viewed as a larger lexical unit. It consists of coding sequences (exons) that code for the primase protein, as well as regulatory sequences (such as promoters and enhancers) that control when and where primase is produced. This is analogous to a complex word formed by a root and affixes that modify its usage.

Syntax: Linear Order and Structural Constraints

Syntax governs how words combine into larger structures, imposing both linear and hierarchical constraints. Human syntax is recursive and generative, enabling an unbounded range of expressions from finite resources [10].

DNA also exhibits syntactic properties in a limited but non-trivial sense. Linear order is critical: frame shifts, misplaced stop codons, or altered regulatory sequences can radically change or eliminate function [3]. There are positional constraints analogous to syntactic rules—promoters must precede genes, splice sites must occur at specific boundaries, and regulatory motifs interact across defined spatial relationships [4].

Accepted Manuscript

However, DNA syntax is procedural rather than representational. It lacks recursion, hierarchical embedding, and ambiguity resolution. Genetic sequences do not express propositions; they trigger biochemical processes. The “grammar” of DNA specifies what reactions occur, not what states of affairs are described. For example, when a single nucleotide is deleted from a coding sequence, it shifts the reading frame of the triplet codons, resulting in a nonsensical protein sequence or a premature stop. This type of frameshift mutation is observed in the human ABO blood group gene. For example, a one-base deletion (ΔG at position 261) in exon 6 alters the reading frame and produces a truncated, nonfunctional enzyme, which is responsible for the O blood type [11]. Similarly, specific sequence motifs—akin to grammar—are essential for proper DNA function. The bacterial primase enzyme, for instance, does not initiate RNA primer synthesis at random DNA sequences. E. coli primase (DnaG) ignores about 97% of possible trinucleotide sites and initiates primers only at specific recognition sequences roughly every 1.5 to 2 kilobases apart [12]. This process is analogous to the necessity of placing certain words or punctuation correctly in a sentence for it to make sense. A misplaced stop codon, for instance, is like inserting a period in the middle of a sentence; it prematurely ends the message. In DNA, an incorrectly positioned stop codon aborts the protein synthesis, adhering to the "syntax" rules of the genetic code.

Semantics: Meaning Without Reference

In linguistics, semantics concerns meaning and reference: how linguistic expressions relate to concepts, entities, and states of affairs. Human language is intentional, referential, and capable of truth or falsity [13].

DNA semantics operates under a fundamentally different logic. The “meaning” of a genetic sequence is exhausted by its functional consequences within a cellular context—protein

Accepted Manuscript

structure, expression timing, or regulatory interaction [14]. A gene does not refer to a protein in the way a word refers to an object or concept; it causally participates in its production. This distinction has been emphasized in philosophical critiques of biological information that caution against importing semantic notions wholesale from linguistics into biology [15].

Accordingly, DNA meaning cannot be false, metaphorical, or hypothetical. A sequence can only function properly, malfunction, or remain unexpressed. This marks a categorical difference between representational language and operational biological coding systems. For example, the DNA sequence of a gene is considered "meaningful" only to the extent that it produces a functional product. For example, the gene that encodes DNA primase doesn't merely describe the primase; it actually serves as a template for assembling the primase enzyme. If a mutation introduces a premature stop codon in that gene, the result is a truncated, nonfunctional enzyme, which can lead to certain genetic diseases. However, we wouldn't label the message of the mutated gene as "false"; we would simply describe it as nonfunctional. In human language, a sentence can be false or nonsensical, but in DNA, a "nonsensical" sequence merely fails to produce a viable outcome. For instance, when a codon changes from a sense codon to a nonsense (stop) codon, it doesn't carry metaphorical meaning; instead, it halts protein synthesis entirely, eliminating the gene's effect rather than misrepresenting it.

Pragmatics: Context Without Intention

Pragmatics studies how context shapes meaning in language use, including speaker intention, implicature, and social norms. While DNA lacks speakers and intentions, context nonetheless plays a decisive role in genetic expression. The same sequence can yield different outcomes depending on cellular environment, epigenetic modification, developmental stage, or external conditions [4].

Accepted Manuscript

This context sensitivity is sometimes described in pragmatic terms within biosemiotics [16], but the analogy remains limited. DNA interpretation involves no inference, negotiation, or communicative goals. Context alters outcomes mechanically rather than conversationally. For instance, the lac operon in E. coli contains genes responsible for lactose metabolism and is expressed only when lactose is present and glucose is absent in the cell's environment. In any other context, these genes—though present in the DNA—remain unexpressed or "silent." The DNA sequence itself does not change; rather, it is the cellular context (the availability of lactose and the absence of glucose) that enables the sequence to be read and acted upon. Similarly, the gene for DNA primase is active only during the S-phase of the cell cycle, when DNA replication occurs. The presence of specific replication signals and proteins, such as helicase unwinding the DNA, provides the context in which the primase gene is transcribed and its protein product utilized. Like a sentence that only makes sense in a particular situational context, a genetic sequence's effect relies on biochemical context—albeit without any conscious intention behind it.

Diachrony: Evolution and Language Change

Historical linguistics examines how languages change over time through sound shifts, semantic drift, grammaticalization, and contact [17]. Genetic systems likewise change through mutation, drift, recombination, and natural selection [8].

The parallel is strongest at this level. Both systems exhibit descent with modification, conservation of core structures, and divergence into families. Yet the driving forces differ fundamentally. Language change is mediated by cognition, culture, and social interaction [18], whereas genetic change is governed by differential survival and reproduction. For example, the DNA primase enzyme in bacteria and the primase found in humans have diverged significantly in both sequence and subunit structure over billions of years. However,

Accepted Manuscript

they still share the core function of synthesizing RNA primers. This situation is similar to the evolution of languages from Latin; for example, the Latin word "mater" evolved into "madre" in Spanish and "mère" in French. Although these words sound different, they retain the same meaning: "mother." In biology, there are distinct families of primases that exhibit no obvious sequence similarity, suggesting they have taken independent evolutionary paths. Nevertheless, even these divergent primases often share essential catalytic features due to the functional constraints of their roles [19]. In both linguistic and biological contexts, the process of inheritance with variation produces families—language families and gene families—that preserve a core structure or meaning while diversifying over time.

Conclusion

Applying linguistic methodology to DNA reveals a system that is language-like in structure but not linguistic in kind. DNA employs discrete symbols, combinatorial organization, constrained order, context sensitivity, and historical evolution. At the same time, it lacks intentionality, reference, productivity, and representation—the defining properties of human spoken and written languages.

The value of the comparison lies not in treating DNA as a literal language, but in using linguistic categories to specify precisely where the analogy holds and where it fails. DNA is best understood as a chemically instantiated, operational symbol system—one that converges with language at the level of structure while diverging decisively at the levels of meaning, use, and cognition.

References

1. Hockett, C., *The Origin of Speech.* Scientific American, 1960. **203**: p. 88-96.
2. Saussure, F.d., *Course in General Linguistics* ed. T. W. Baskin. 1916/2011: Columbia University Press.

Accepted Manuscript

3. Crick, F.H.C., *On protein synthesis.* Symposia of the Society for Experimental Biology, 1958. **12**: p. 138–163.
4. Alberts, B., et al., *Molecular Biology of the Cell* 7th ed. ed. 2022: Garland Science.
5. Pattee, H.H., *Cell psychology: An evolutionary approach to the symbol–matter problem.* Cognition and Brain Theory, 1982. **5**: p. 325–341.
6. Inusa, B.P.D., et al., *Sickle Cell Disease-Genetics, Pathophysiology, Clinical Presentation and Treatment.* Int J Neonatal Screen, 2019. **5**(2): p. 20.
7. Jackendoff, R., *Foundations of Language (Brain, Meaning,Grammar, Evolution)*. 1st ed. 2003: Oxford University Press. 477.
8. Crick, F.H., *The origin of the genetic code.* J Mol Biol, 1968. **38**(3): p. 367-79.
9. Maynard Smith, J., *The concept of information in biology.* Philosophy of Science, 2000. **67**(2): p. 177–194.
10. Chomsky, N., *Aspects of the Theory of Syntax.* . 1965: MIT Press. 251.
11. Kitano, T., A. Blancher, and N. Saitou, *The functional A allele was resurrected via recombination in the human ABO blood group gene.* Mol Biol Evol, 2012. **29**(7): p. 1791-6.
12. Soffer, A., et al., *Inferring primase-DNA specific recognition using a data driven approach.* Nucleic Acids Res, 2021. **49**(20): p. 11447-11458.
13. Saussure, F.d.W.B., Trans.). Columbia University Press., *Course in General Linguistics (W. Baskin, Trans.)*. 1916/2011: Columbia University Press.
14. Barbieri, M., *The Organic Codes: An Introduction to Semantic Biology.* . 2003: Cambridge University Press.
15. Godfrey-Smith, P., *Information, arbitrariness, and selection: Comments on Maynard Smith.* . Philosophy of Science, 2000. **67**(2): p. 202–207.
16. Hoffmeyer, J., *Biosemiotics: An Examination into the Signs of Life and the Life of Signs.* 2008: University of Scranton Press.
17. Labov, W., *Principles of Linguistic Change, Volume 1: Internal Factors.* . 1994: Blackwell.
18. Croft, W., *Explaining Language Change: An Evolutionary Approach.* 2000: Longman.
19. Kuchta, R.D. and G. Stengel, *Mechanism and evolution of DNA primases.* Biochim Biophys Acta, 2010. **1804**(5): p. 1180-9.

Accepted Manuscript