Beyond Wierzbicka: LLM-Powered Analysis of the Natural Semantic Metalanguage Boundaries

Bartosz Naskręcki[1], Michał Szczyszek[2]

[1]Adam Mickiewicz University, Poznań, Poland
bartosz.naskrecki@amu.edu.pl
[2]Adam Mickiewicz University, Poznań, Poland
szczysze@amu.edu.pl

Abstract

This study presents a novel computational approach to analyzing the boundaries of Anna Wierzbicka's Natural Semantic Metalanguage (NSM) using Large Language Models (LLMs). Through a comprehensive analysis of 22 concepts across five semantic domains (mathematical, legal, physical, social-emotional, and hybrid), we demonstrate that fundamental mathematical concepts systematically exceed NSM's expressive capacity. Using Claude Sonnet 4.5, we developed a 22-dimensional analytical framework that reveals a 45.5 percentage point difficulty gap between mathematical concepts (79.1% NSM difficulty) and socialemotional concepts (33.6% NSM difficulty). Our findings empirically confirm that while NSM excels at describing human social and emotional experiences—its intended domain—it is fundamentally inadequate for expressing formal mathematical abstractions. This work introduces the first LLM-based NSM evaluation system and establishes quantitative boundaries for semantic metalanguage applicability.

Keywords: Anna Wierzbicka's Natural Semantic Metalanguage, Large Language Models, Semantic Analysis, Mathematical, Legal and Social Concepts, Computational Linguistics

1. Introduction

1.1 Background

Anna Wierzbicka's Natural Semantic Metalanguage (NSM) theory posits that all languages share a core set of approximately 65 semantic primitives (Wierzbicka, 1996, 2006). These primitives — concepts like *I*, *YOU*, *SOMEONE*, *SOMETHING*, *THINK*, *FEEL*, *GOOD*, *BAD* — are claimed to be:

a. Universal across all human languages
b. Indefinable (semantically atomic)
c. Sufficient to define all other concepts through reductive paraphrase.

These elementary units in Wierzbicka's concept are, in a sense, semantic axioms.

The NSM framework has been successfully applied to diverse domains including emotion concepts (Wierzbicka, 1999), cultural scripts (Goddard & Wierzbicka, 2004), and cross-cultural pragmatics (Wierzbicka, 2003). However, critical questions remain about its boundaries: Which concepts can NSM adequately express, and which fundamentally exceed its capacity?

1.2 Research Gap

While NSM has been extensively validated for social-emotional concepts, its applicability to formal domains — particularly mathematics — remains underexplored. Previous critiques (Riemer, 2006; Drobnak, 2009) have suggested limitations in handling abstract concepts, but lacked systematic, quantitative evaluation.

1.3 Research Questions

1. Can fundamental mathematical concepts be adequately expressed using NSM's 65 primitives?
2. How do mathematical concepts compare quantitatively to social-emotional concepts in NSM difficulty?
3. Can Large Language Models (LLMs) provide expert-level semantic analysis of NSM explications?
4. What dimensions differentiate concepts that fit NSM from those that exceed it?

1.4 Contribution

This study makes four key contributions:

a. First LLM-based NSM evaluation system using Claude Sonnet 4.5 and GPT-5
b. 22-dimensional analytical framework for concept analysis (extending prior 12-dimension models)
c. Quantitative evidence of NSM boundaries through batch analysis of 18 concepts
d. Empirical validation that mathematical concepts systematically exceed NSM capacity.

2. Theoretical Framework

2.1 NSM Theory Overview

We base our analyses on the concept of the Natural Semantic Metalanguage (NSM) by Anna Wierzbicka and Cliff Goddard. According to this theory, there are 65 elementary units (universals / basic concepts / semantic axioms / indefinable semantic elements) in natural language that have lexical equivalents in every language. Below we present a list of these elements, grouped thematically, in accordance with the classical NSM approach – Table 1.

Bartosz Naskręcki, Michał Szczyszek, *Beyond Wierzbicka: LLM-Powered Analysis of the Natural Semantic Metalanguage Boundaries*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 73–79. Poznań: Adam Mickiewicz University Press, 2025. © Authors, 2025. https://doi.org/10.14746/amup.9788323245414.14. Open Access chapter, distributed under the terms of the CC licence (BY-NC--ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

Categorical Group	Semantic Primes
Substantives	I, YOU, SOMEONE/PERSON, PEOPLE, SOMETHING/THING, BODY
Determiners	THIS, THE SAME, OTHER
Quantifiers	ONE, TWO, MANY/MUCH, FEW/LITTLE, SOME, ALL
Evaluators / Descriptors	GOOD, BAD, BIG, SMALL
Mental Predicates	KNOW, THINK, WANT, FEEL, SEE, HEAR
Speech	SAY, WORD, TRUE
Actions / Events	DO, HAPPEN, MOVE
Relations / Existential	BE (exist) →THERE IS, BE (somewhere), BE (someone/something), HAVE
Life / Time	LIVE, DIE, WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT
Place / Space	WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE, TOUCH
Logic / Miscellaneous	NOT, MAYBE, CAN, BECAUSE, IF, VERY, MORE, LIKE

Tabela 1. List of 65 elementary units (NSM Primes)

2.2 Semantic Molecules

Beyond primitives, NSM recognizes semantic molecules—universal but complex concepts like MOTHER, Family, Sad — that serve as intermediate building blocks (Goddard, 2012). While semantic elements (universals, primes) are the absolute, indefinable bases of meaning (like atoms), semantic molecules are:

- Complex, yet highly organized and common (though not necessarily universal) concepts.
- Constructs definable using a combination of the primitive semantic elements (NSM).
- Concepts that function as "intermediate products" in complex definitions.

They can be compared to molecules in chemistry – they are built from atoms (semantic elements), but they themselves constitute stable and recognizable units, which are then used to build larger and more complicated structures (the meanings of complex words).

Semantic molecules are crucial in NSM because:

- **They facilitate defining:** They allow complex words to be defined in a more concise and readable way, instead of breaking down every concept directly into the 65 primitive elements.
- **They point to cultural similarities/differences:** Although some molecules are universal, many are culturally or regionally specific, which helps to capture and describe differences in ways of thinking between cultures (e.g., concerning emotions, social relations, or body-part vocabulary).

Examples of Semantic Molecules – Table 2.

Semantic Molecule	NSM Definition - Simplified Example	Note
Mother	someone kind of person, someone knows: this is my body part for some time before, someone thinks: "I want this child to live and to happen good things to it"	Many social and biological definitions based on simple primes.
Family	a group of people, who live together for a long time and think about each other like parts of one whole thing	Definition based on primes concerning place, time, relations, and thinking.
Sad	someone feels something bad in their body, someone thinks: "something bad happened", does not want this to happen	An internal state with elements of thought, feeling, and evaluation.

Table 2. Examples of Semantic Molecules

2.3 NSM's Intended Domain

Wierzbicka explicitly designed NSM for analyzing:

- Human emotions and feelings
- Social relationships and interactions
- Cultural norms and values
- Everyday experiences

Critical observation: Mathematical and formal concepts were never claimed to be NSM's strength.

3. Methodology

In this section we will explain our methodology and tools used to study the NSM system.

3.1 LLM-Based Analysis System (GPT-5 Architecture)

Our analytical pipeline was designed around GPT-5, enabling multi-pass semantic reasoning, internal self-critique, and long-context validation unavailable in earlier models. The methodological architecture consists of three complementary tiers that build on each other to provide increasingly precise analysis of NSM explications.

Tier 1: Heuristic Baseline (v1)

This stage provides deterministic, rule-based grounding for later probabilistic reasoning.

• TF-IDF weighting of NSM primitives
• Automated detection of non-NSM lexical units
• Primitive distance heuristics (primitive–concept mapping)
• Baseline semantic similarity checks using lexical surfaces

These heuristics serve as an anchor for later comparisons with GPT-5 outputs.

Tier 2: Machine Learning Enhancement (v2)

The second tier uses domain-neutral ML models to extract latent semantic structure.

• Bayesian category classification across four semantic domains
• Embedding-based clustering approximating conceptual proximity
• Dependency-network construction mapping co-occurrence of primitives
• PCA reduction uncovering principal semantic axes

This produces numerical representations independent of GPT-5 reasoning, which supports cross-validation.

Tier 3: GPT-5 Expert System (v3)

GPT-5 functions as a semantic expert capable of evaluating explications in a multi-stage reasoning pipeline:

· Deep parsing of the NSM explication with primitive-level validation
· Domain inference using GPT-5 structured reasoning mode
· Multi-dimensional scoring across the full 22-dimension metric
· Self-critique cycle identifying circularity, hidden presuppositions, ambiguous structures, and illicit semantic imports
· Final synthesis including:
  o recommended corrections,
  o classification of violations,
  o consistency-checked numerical scores,
  o confidence estimates.

GPT-5's chain-of-thought and error-checking capabilities ensure that the results are semantically consistent, detect conceptual leakage, and remain stable across multiple runs.

3.2 The 22-Dimensional NSM Evaluation Framework

The evaluation framework extends earlier models from 12 to 22 dimensions, capturing structural, semantic, categorical, and cognitive properties of NSM explications. These dimensions fall into three groups – see Teble 3.:

Quality and Structure	difficulty, definability, semantic coherence, structural depth
Semantic Domain Placement	emotional loading, abstract-formal score, cultural-linguistic profile
Cognitive and Cross-Linguistic Factors	conceptual distance, metaphorical complexity, developmental age, cognitive load

Table 3. Three groups of dimensions

GPT-5 uses a multi-pass verification loop to ensure the internal consistency of scores—for instance, high abstraction typically co-occurs with high conceptual distance and low emotional loading.

3.3 Prompt Engineering and GPT-5 Reasoning Protocol

The prompt used for GPT-5 was engineered to maximize stability, transparency, and reproducibility. It includes:

· A structured reasoning chain divided into semantic, syntactic, and domain-specific subtasks
· A self-critique loop where GPT-5 revisits and corrects its own evaluations
· Representative examples of good and bad NSM explications
· A circular-definition detection subroutine
· A domain-sensitive inference schema for mathematical and legal concepts
· Score-consistency checks enforcing bounded variance

The final prompt (~1,200 words) yields high output stability (<3% variance), significantly outperforming earlier LLM generations.

3.4 Dataset

The dataset includes 22 concepts across five categories – table 4:

Mathematical	entropy, function, limit, set, infinity
Physical	velocity, energy, quantum
Legal	solicitation, rape, murder, robbery
Social-Emotional	sadness, friendship, gratitude, jealousy
Hybrid	time, mind, growth

Table 4. Five categories of concepts

Each concept has a manually prepared NSM explication serving as ground truth input for analysis. The dataset was selected to cover both domains where NSM is well-known to succeed and domains where it is expected to fail.

3.5 Analysis Procedure

Step 1: Human Explication

Authors prepare explications strictly adhering to NSM primitives and canonical syntactic frames.

Step 2: GPT-5 Semantic Analysis

GPT-5 evaluates each explication across all 22 dimensions, detects semantic violations, and produces a structured JSON report.

Step 3: Dimensional Scoring

Quantitative metrics (difficulty, definability, coherence, etc.) are extracted and validated using GPT-5 internal consistency mechanisms.

Step 4: Statistical Testing

We compare categories using ANOVA, two-sample t-tests, and effect size measurements (Cohen's d), contrasting domains such as mathematical vs. social-emotional.

Step 5: Multi-Model Validation

Results are cross-checked with:

· heuristic baseline (v1)
· ML intermediate model (v2)
· trace-level reasoning consistency in GPT-5

This three-layer comparison ensures robustness of findings.

3.6 Implementation in the NSM Analyzer

The analysis described above is operationalized in the application implemented in app.py. The app provides:

· Streamlit interface allowing users to load or enter NSM explications
· model selection (GPT-5, GPT-5-mini, GPT-5-nano)
· automatic invocation of GPT-5 with the calibrated prompt
· visual presentation of metrics including:
  o primitive coverage
  o semantic categories
  o lexical statistics
  o 22-dimension evaluation results
· export functionality generating a complete Markdown report

Internally, application performs the following steps:

1. Reads user explication input
2. Preprocesses and validates syntax
3. Sends the prompt to GPT-5 via the OpenAI API
4. Parses the returned JSON
5. Generates human-readable summaries and charts
6. Stores analysis history for reproducibility

This implementation enables reliable, repeatable NSM analysis without requiring users to manually construct prompts or handle GPT-5 output parsing.

3.7 Automatic Improvement of NSM Definitions (generate_improvement_prompt.py)

The auxiliary script generate_improvement_prompt.py automates the process of improving NSM explications. It works by:

1. Reading an existing analysis file (e.g., a Markdown report from the application)
2. Extracting key metadata such as:
   a. list of used primitives
   b. list of missing primitives
   c. identified semantic violations
   d. errors (e.g., circularity, non-NSM lexicon)
3. Preparing a specialized GPT-5 prompt that instructs the model to:
   a. rewrite the explication using only NSM-approved primitives
   b. fix structural, semantic, and syntactic issues
   c. preserve intended meaning while eliminating non-NSM elements
   d. return the improved explication in structured JSON

The resulting prompt can be fed into GPT-5 to generate a corrected NSM explication with high definability, low difficulty, and minimal semantic leakage. The tool therefore acts as a second-stage refinement layer, enabling iterative improvement of explications and supporting rapid prototyping of NSM-compliant definitions.

4. Results

4.1 Primary Analysis: Mathematical vs. Social-Emotional Concepts

Category	n	Avg Difficulty	Avg Definability	Avg Formality	Beyond NSM
Mathematical	5	79.1%	28.0%	90.0%	100% (5/5)
Physical	3	68.3%	39.5%	74.2%	67% (2/3)
Legal	4	68.13%	61.25%	59.38%	83.48%
Social-Emotional	4	33.6%	75.3%	13.7%	0% (0/4)
Hybrid	3	59.5%	50.4%	49.8%	33% (1/3)

Table 5: Aggregate Statistics by Category

Key finding: Mathematical concepts show 45.5 percentage points higher difficulty than social-emotional concepts (79.1% vs. 33.6%, $p < 0.001$).

4.2 Individual Concept Analysis

Rank	Concept	NSM Difficulty	Definability	Exceeds NSM
1.	Infinity	85.7%	19.4%	+
2.	Limit	81.3%	25.8%	+
3.	**Rozbój**	80.0%	57.5%	+
4.	Entropy	78.6%	28.3%	+
5.	**Zabójstwo**	77.5%	77.5%	+
	Quantum	77.4%	27.3%	+
	Function	76.2%	31.5%	+

Table 6: Top 7 Most Difficult Concepts

In the top are all mathematical or physics and two legal concepts.

Rank	Concept	NSM Difficulty	Definability	Exceeds NSM
1.	Sadness	28.3%	81.7%	X
2.	Friendship	32.1%	77.3%	X
3.	**Zabójstwo**	77.5%	77.5	X
4.	Gratitude	35.7%	73.2%	X
5.	Jealousy	38.4%	68.9%	X
6.	Growth	42.6%	64.1%	X
7.	**Podżeganie**	55%	60%	X

Table 7: Top 5 Most Definable Concepts

In the top are all top 5 are social-emotional or hybrid and 2 legal concepts.

4.3 Case Study: Entropy

NSM Explication Attempt:

– something about something
– when I think how many different things can happen
– when I think how much I don't know what will happen
– if I know everything entropy is small
– if I know nothing entropy is large
– this says how much something is uncertain

LLM Analysis Results:

– Quality Score: 45.2%
– NSM Difficulty: 78.6%
– Definability: 28.3%
– Formality: 88.4%

Key Problems Identified by LLM:

– Circular definition: Uses "entropy" to define "entropy"
– Requires set theory: "different things" needs mathematical formalism
– Lacks precision: Doesn't distinguish thermodynamic vs. information-theoretic entropy
– Missing formal concepts: Probability, state space, measure theory

LLM Verdict: "Concept 'entropy' fundamentally exceeds NSM. Requires formal mathematical metalanguage from probability theory, measure theory, and information theory."

4.4 Case Study: Sadness

NSM Explication:

– someone feels something bad
– this person doesn't want to feel this
– because something bad happened
– this person thinks about this
– this person wants it didn't happen

LLM Analysis Results:

– Quality Score: 78.4%
– NSM Difficulty: 28.3%
– Definability: 81.7%
– Formality: 8.2%

LLM Assessment: "High-quality NSM explication. Uses only primitives (FEEL, BAD, WANT, BECAUSE, HAPPEN, THINK). Captures essential features of sadness. Minor improvements: add temporal aspect (FEEL for A LONG TIME) and intensity (VERY BAD)."

LLM Verdict: "Concept 'sadness' fits perfectly within NSM's core domain."

4.5 LLM vs. Heuristic Comparison

Metric	Heuristic v2	LLM v3	Difference
Avg Difficulty (Math)	74.2%	83.5%	+9.3 pp
Avg Definability (Math)	18%	21.9%	+3.8 pp
Circular Definitions Detected	0/3	2/3	+67 %
Concrete Problems Identified	0	16	+16
Constructive Improvements	0	15	+15

Table 8: Method Comparison

Key advantage of LLM: Detects logical errors (circular definitions), provides expert synthesis, and offers actionable improvements.

4.6 New Dimensions: Insights

Translatability correlates strongly with NSM fitness ($r = 0.82$, $p < 0.01$):

– Mathematical concepts: avg 35% translatability
– Social concepts: avg 93% translatability

Conceptual Distance differentiates categories:

– Mathematical: avg distance 5.2 steps from primitives
– Social: avg distance 1.1 steps from primitives

Emotional Loading predicts NSM definability ($r = 0.76$, $p < 0.01$):

– High emotional loading → high definability
– Low emotional loading → low definability

5. Discussion

5.1 Why Mathematical (and same legal) Concepts Exceed NMS

Three fundamental incompatibilities:

5.1.1 Lack of Formal Primitives

– NSM primitives describe human experience, not formal structures. Missing concepts:
– Set membership: ∈ (element of)
– Function mapping: f: A → B
– Infinity: ℵ₀, ∞
– Continuity: ε-δ definition
– Measure: μ(A)

5.1.2 Level of Abstraction

Mathematical concepts operate at meta-meta-levels:

– NSM: Maximum 2-3 levels (primitives → molecules → concepts)
– Mathematics: 6+ levels (e.g., biholomorfism requires: numbers → complex numbers → functions → holomorphic functions → bijective holomorphic functions)

5.1.3 Precision Requirements

Mathematics demands exact definitions:

– NSM: "very," "maybe," "like" (inherently vague)
– Mathematics: "∀ε>0 ∃δ>0..." (absolutely precise)

5.2 NSM's Strength: Social-Emotional Domain

Our findings validate NSM's core claim: it excellently captures human social and emotional experiences.

Evidence:

– 0% of social-emotional concepts exceed NSM
– Average definability: 75.3%
– Average quality score: 73.5%

Why NSM succeeds here?

– Primitives designed for this domain (FEEL, WANT, THINK, GOOD, BAD)
– Cross-cultural universality of basic emotions
– Experiential grounding (not abstract).

5.3 LLM as Linguistic Expert

Advantages over heuristic methods:

– Semantic understanding: Detects circular definitions
– Expert synthesis: Provides coherent assessment
– Constructive feedback: Actionable improvement suggestions
– Contextual reasoning: Understands why concepts fail.

Limitations:

– Requires careful prompt engineering
– Potential inconsistency between runs (~8% variance)
– Lacks formal mathematical knowledge (relies on training data).

5.4 Implications for Linguistic Theory

NSM is domain-specific, not universal:

– Universal within its intended scope (social-emotional)
– Non-universal outside that scope (formal, mathematical)

This doesn't invalidate NSM—it clarifies its boundaries.

5.5 The "Beyond-NSM" Threshold

Quantitative definition:

– Beyond NSM: NSM Difficulty > 70% AND Definability < 30%
– Beyond NSM: NSM Difficulty > 70% AND Definability < 30%
– Borderline: 50% < Difficulty < 70%
– Within NSM: Difficulty < 50% AND Definability > 60%

Predictors of "Beyond-NSM" status:

– Abstract-Formal category > 80%
– Conceptual Distance > 4
– Translatability < 40%
– Emotional Loading < 20%
– Developmental Age > 5 (15+ years).

6. Limitations

6.1 Sample Size

22 concepts is modest. Future work should analyze 100+ concepts across more domains (medicine, theology, sociology).

6.2 Explication Quality

Author-generated explications may not represent optimal NSM formulations. Expert NSM linguists might construct better explications.

6.3 LLM Bias

Claude Sonnet 4.5's training data may influence assessments. Cross-validation with other LLMs (GPT-5, Gemini) recommended.

6.4 Language Specificity

Analysis conducted in Polish. Cross-linguistic validation needed to confirm universality of findings.

6.5 No Expert Validation

Results not validated against NSM experts' assessments. Inter-rater reliability study required.

7. Future Work

7.1 Cross-Linguistic Validation

Test same concepts in 10+ languages (English, Mandarin, Arabic, Swahili, etc.) to verify universality.

7.2 Expert Validation Study

Compare LLM assessments with judgments from NSM-trained linguists. Calculate inter-rater agreement (Krippendorff's α).

7.3 Larger Dataset

– Analyze 100+ concepts spanning:
– Law (contract, liability, precedent)
– Medicine (diagnosis, syndrome, pathology)
– Philosophy (qualia, intentionality, ontology)
– Technology (algorithm, encryption, protocol)

7.4 Fine-Tuned LLM

Train LLM specifically on NSM corpus for improved accuracy.

7.5 Automated Explication Generation

Develop AI system to generate optimal NSM explications, not just evaluate them.

7.6 Extended Dimensions

– Add dimensions 23-26:
– Kolmogorov complexity
– Historical semantic stability
– Domain specificity

Compositionality index

8. Conclusion

This study provides the first systematic, quantitative evidence that fundamental mathematical concepts systematically exceed the Natural Semantic Metalanguage's expressive capacity. Through LLMpowered analysis of 22 concepts across four domains, we demonstrate a 45.5 percentage point difficulty gap between mathematical concepts (79.1% NSM difficulty) and social-emotional concepts (33.6% difficulty).

Key findings:

– Mathematical concepts are beyond NSM: 100% of mathematical concepts (5/5) exceed NSM capacity, with average difficulty 79.1% and definability only 28.0%.

– Social-emotional concepts fit NSM perfectly: 0% of social-emotional concepts (0/4) exceed NSM, with average difficulty 33.6% and definability 75.3%.

– LLM analysis is superior to heuristics: Claude Sonnet 4.5 detects circular definitions, provides expert synthesis, and offers constructive improvements that heuristic methods miss.

– 22-dimensional framework reveals boundaries: New dimensions (translatability, conceptual distance, emotional loading) predict NSM fitness with high accuracy.

– NSM is domain-specific: NSM excels at its intended domain (human social-emotional experience) but is fundamentally inadequate for formal mathematical abstractions.

Theoretical implication: NSM's universality claim should be reinterpreted as domain-universal (universal within social-emotional domain) rather than absolute-universal (universal across all conceptual domains).

Practical contribution: Our LLM-based evaluation system provides a scalable, systematic method for assessing NSM explication quality and identifying conceptual boundaries.

Final verdict: Anna Wierzbicka's NSM is a powerful tool for analyzing human social and emotional concepts cross-culturally. However, it was never designed for—and cannot adequately express – fundamental mathematical and formal concepts. This is not a failure but a clarification of scope.

References

Drobnak, T. F. (2009). On the merits and shortcomings of semantic primes and Natural Semantic Metalanguage in cross-cultural Translation. *English Language Oversees: Perspectives and Enquiries*, 6(1-2), 99-121.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press.