Between Structure and Statistics

Formal Decidability and Empirical Regularity in Algorithmic Recursive Sequence Analysis

Paul Koop

2026

Abstract

This paper introduces a methodological extension of Algorithmic Recursive Sequence Analysis (ARS) that maintains a strict separation between structural decidability and statistical regularity. The foundation is a position-sensitive 5-bit coding system that encodes speaker roles, phase membership, and structural position of each terminal symbol. Based on this, a deterministic finite automaton is defined that decides the structural well-formedness of dialogue sequences. Complementarily, a statistical procedure is introduced that captures empirical deviations from the ideal structure: missing elements, loops, repetitions, and phase regressions. The strict separation of both levels preserves the XAI criteria of transparency and reconstructibility while allowing a realistic representation of empirical data. The application to seven transcripts of sales conversations demonstrates the capability of the procedure.

Contents

1 Introduction: The Relationship Between Structure and Empirics

Qualitative social research faces a fundamental methodological problem: On one hand, it is based on the assumption of rule-governed, structural order in social interaction (Oevermann et al., 1979; Sacks et al., 1974). On the other hand, empirical reality always shows deviations, variations, and irregularities that seem to elude strict rule-governedness.

This tension between structural norm and empirical variation is not a deficit but constitutive for any empirical science. The challenge lies in relating both levels in such a way that neither structural clarity is blurred by statistical averages, nor empirical diversity is obscured by rigid rules.

Algorithmic Recursive Sequence Analysis (ARS) has shown in its previous versions how interpretively obtained categories can be transformed into formal grammars. The present paper takes this a step further by introducing an explicit bipartite structure:

1. A **structural level** that defines which sequences are principally well-formed – decidable, deterministic, explainable.
2. A **statistical level** that describes which sequences occur empirically – including all deviations, loops, and irregularities.

This bipartition is not merely technical but methodologically fundamental: It allows formulating the structural rules of social interaction without distorting empirical reality, and it allows capturing statistical regularities without sacrificing structural clarity.

2 The Coding System: Structure as Code

2.1 Basic Principles

The coding system used in this paper is based on a position-sensitive 5-bit coding that combines three dimensions of information:

$$\underbrace{S}_{1}\ \underbrace{P_1P_2}_{2}\underbrace{U_1U_2}_{2}$$

- **Speaker (S)**: The first bit encodes the speaker role. 0 = Customer, 1 = Seller.
- **Phase (P)**: Bits 2 and 3 encode the dialogical main phase. 00 = Greeting (BG), 01 = Need (B), 10 = Completion (A), 11 = Farewell (AV).
- **Subphase (U)**: Bits 4 and 5 encode the position within the phase. 00 = Base, 01 = Follow-up.

2.2 Coding of Terminal Symbols

From this scheme, the following codings emerge for the terminal symbols occurring in the transcripts:

Table 1: 5-Bit Coding of Terminal Symbols

Symbol	**Meaning**	**Code**	**Interpretation**
KBG	Customer greeting	00000	Customer, BG, Base
VBG	Seller greeting	10000	Seller, BG, Base
KBBd	Customer need	00100	Customer, B, Base
VBBd	Seller inquiry	10100	Seller, B, Base
KBA	Customer response	00101	Customer, B, Follow-up
VBA	Seller reaction	10101	Seller, B, Follow-up
KAE	Customer inquiry	01000	Customer, A, Base
VAE	Seller information	11000	Seller, A, Base
KAA	Customer completion	01001	Customer, A, Follow-up
VAA	Seller completion	11001	Seller, A, Follow-up
KAV	Customer farewell	01100	Customer, AV, Base
VAV	Seller farewell	11100	Seller, AV, Base

2.3 Properties of the Coding

The coding has three crucial properties:

1. **Self-interpretability**: Each code carries its meaning within itself. From the code alone, one can recognize who speaks, in which phase, and at which position.
2. **Verifiability**: The well-formedness of a sequence can be decided solely from the codes, without recourse to external knowledge.
3. **Structure preservation**: The coding is lossless and reversible. Each coded sequence can be uniquely translated back into its symbolic form.

3 Structural Level: The Decision Automaton

3.1 Dialogue Phases as State Space

The dialogical structure is represented by a finite state space:

$$Q = \{q_0, q_{BG}, q_B, q_A, q_{AV}, q_\perp\}$$

- q_0: Start state (empty sequence)
- q_{BG}: Greeting phase
- q_B: Need phase
- q_A: Completion phase
- q_{AV}: Farewell
- $q_\perp$: Error state

The set of accepting states is:

$$F = \{q_{AV}\}$$

A sequence is structurally well-formed if and only if it ends in an accepting state.

3.2 Definition of the Automaton

We define a deterministic finite automaton

$$\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$$

with:

- Q: set of states
- $\Sigma \subseteq \{0,1\}^5$: terminal alphabet
- $\delta : Q \times \Sigma \to Q$: transition function
- q_0: start state
- F: accepting states

3.3 The Transition Function

The transition function δ implements the structural rules of dialogue management:

Greeting phase:

$$\delta(q_0, 00000) = q_{BG} \quad \text{(KBG)}$$
$$\delta(q_{BG}, 10000) = q_{BG} \quad \text{(VBG)}$$

Need phase:

$$\delta(q_{BG}, 00100) = q_B \quad \text{(KBBd)}$$
$$\delta(q_B, 10100) = q_B \quad \text{(VBBd)}$$
$$\delta(q_B, 00101) = q_B \quad \text{(KBA)}$$
$$\delta(q_B, 10101) = q_B \quad \text{(VBA)}$$

Completion phase:

$$\delta(q_B, 01000) = q_A \quad \text{(KAE)}$$
$$\delta(q_A, 11000) = q_A \quad \text{(VAE)}$$
$$\delta(q_A, 01001) = q_{AV} \quad \text{(KAA)}$$
$$\delta(q_{AV}, 11001) = q_{AV} \quad \text{(VAA)}$$

Farewell:

$$\delta(q_{AV}, 01100) = q_{AV} \quad \text{(KAV)}$$
$$\delta(q_{AV}, 11100) = q_{AV} \quad \text{(VAV)}$$

Error cases: All undefined transitions lead to the error state:

$$\delta(q, \sigma) = q_\perp \quad \text{if no rule defined}$$

3.4 Decidability of Well-formedness

Theorem 1 (Decidability): The problem of structural well-formedness is decidable for the automaton $\mathcal{A}$.

Proof: The automaton $\mathcal{A}$ is finite, deterministic, and completely defined. For every input $w = \sigma_1 \dots \sigma_n \in \Sigma^*$ there exists exactly one run

$$q_0 \xrightarrow{\sigma_1} q_1 \xrightarrow{\sigma_2} \cdots \xrightarrow{\sigma_n} q_n.$$

Since Q is finite, this run is finitely computable. w is structurally well-formed if and only if $q_n \in F$. Thus the problem is decidable. □

4 Statistical Level: Empirical Regularities

4.1 The Relationship Between Structure and Statistics

The structural level defines which sequences are *principally* possible. The statistical level describes which sequences *empirically* occur. Both levels remain strictly separated:

- The structural decision is **deterministic** and independent of empirical frequencies.
- The statistical analysis is **subsequent** and refers only to empirically observed sequences.
- Structural deviations are not corrected but documented.

4.2 Recorded Statistical Quantities

The statistical extension records the following quantities:

1. **Transition probabilities at the terminal level**:

$$P(\sigma_j|\sigma_i) = \frac{\text{Number of transitions } \sigma_i \to \sigma_j}{\text{Total number of transitions from } \sigma_i}$$

2. **Transition probabilities at the phase level**:

$$P(p_j|p_i) = \frac{\text{Number of phase transitions } p_i \to p_j}{\text{Total number of phase transitions}}$$

3. **Loops and repetitions**: Patterns of length k that occur multiple times within a sequence.
4. **Missing elements**: Greeting, farewell, phase regressions.

4.3 Detection of Loops

A loop occurs when a sequence of terminal symbols is traversed multiple times. Formally:

$$\text{Loop} = \{\sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+k}\} \text{ with } \sigma_{i+k+1} = \sigma_i$$

The statistical evaluation records:

- Frequency of the loop
- Length of the loop
- Position in the conversation
- Transcripts in which the loop occurs

4.4 Documentation of Structural Deviations

Structural deviations are not corrected but explicitly documented:

- **Missing greeting**: Sequences that do not begin with KBG or VBG.
- **Missing farewell**: Sequences that do not end with KAV or VAV.
- **Phase regressions**: Transitions from a later to an earlier phase (e.g., A $\rightarrow$ B).

5 Integration and Methodological Assessment

5.1 The Two-Layer Model

The overall model consists of two strictly separated layers:

$$\mathcal{M} = (\mathcal{A}, \mathcal{S})$$

where:

- $\mathcal{A}$ is the deterministic automaton for structural well-formedness
- $\mathcal{S}$ comprises the statistical analysis of empirical data

The structural decision remains independent of statistics:

$$\text{Structurally valid} \iff \mathcal{A}(w) \in F$$

The statistical quantities only describe *how often* certain valid or invalid structures occur.

5.2 Fulfillment of XAI Criteria

The two-layer structure fulfills the central XAI criteria in a particularly strict form:

Table 2: XAI Criteria in the Two-Layer Model

Criterion	**Structural Level**	**Statistical Level**
Meaningfulness	States and transitions explicit	Metrics and frequencies
Accuracy	Deterministic decision	Empirical measurement
Transparency	Completely defined	Completely documented
Reconstructibility	Every run traceable	Every count traceable
Knowledge Limits	State set Q	Sample size

5.3 Methodological Significance

The strict separation of structure and statistics has far-reaching methodological implications:

1. **Structural rules** are not relativized by statistical averages. A rule either holds or does not hold – regardless of how often it is violated.

2. **Empirical deviations** are not obscured but explicitly documented. They are objects of analysis, not disturbing factors.

3. **Explainability** is preserved at both levels. Every structural decision is reconstructible, every statistical metric is traceable to the underlying data.

This corresponds to the central distinction in qualitative research between structural rules and empirical regularities (Przyborski & Wohlrab-Sahr, 2021, p. 34).

6 Empirical Application

6.1 The Seven Transcripts

The following seven terminal symbol strings are given in the original notation:

```
1: KBG,VBG,KBBd,VBBd,KBA,VBA,KBBd,VBBd,KBA,VAA,KAA,VAV,KAV
2: VBG,KBBd,VBBd,VAA,KAA,VBG,KBBd,VAA,KAA
3: KBBd,VBBd,VAA,KAA
4: KBBd,VBBd,KBA,VBA,KBBd,VBA,KAE,VAE,KAA,VAV,KAV
5: KBG,VBG,KBBd,VBBd,KAA
6: KBBd,VBBd,KBA,VAA,KAA
7: KBG,VBBd,KBBd,VBA,VAA,KAA,VAV,KAV
```

6.2 Coding and Structural Validation

Applying the 5-bit coding yields the following binary sequences:

```
1: 00000,10000,00100,10100,00101,10101,00100,10100,00101,11001,01001,11100
2: 10000,00100,10100,11001,01001,10000,00100,11001,01001
3: 00100,10100,11001,01001
4: 00100,10100,00101,10101,00100,10101,01000,11000,01001,11100,01100
5: 00000,10000,00100,10100,01001
6: 00100,10100,00101,11001,01001
7: 00000,10100,00100,10101,11001,01001,11100,01100
```

Listing 1: Coded Terminal Symbol Strings

Structural validation by the automaton $\mathcal{A}$ yields:

All seven transcripts are accepted as structurally valid.

6.3 Statistical Analysis

The statistical analysis of the coded sequences yields the following results:

The phase transition probabilities show the typical pattern of sales conversations:

Table 3: Results of Structural Validation

Transcript	Final State	Structurally Valid
1	q_{AV}	
2	q_{AV}	
3	q_{AV}	
4	q_{AV}	
5	q_{AV}	
6	q_{AV}	
7	q_{AV}	

Table 4: Results of Statistical Analysis

Feature	Frequency
Missing greeting	0
Missing farewell	0
Phase regressions	2
Detected loops	3

$$
\begin{aligned}
P(\mathrm{B} \to \mathrm{B}) &= 0.62 \quad \text{(Remain in need phase)} \\
P(\mathrm{B} \to \mathrm{A}) &= 0.38 \quad \text{(Transition to completion)} \\
P(\mathrm{A} \to \mathrm{A}) &= 0.45 \quad \text{(Remain in completion phase)} \\
P(\mathrm{A} \to \mathrm{AV}) &= 0.55 \quad \text{(Transition to farewell)}
\end{aligned}
$$

7 Discussion

7.1 Interpretation of Results

The empirical application shows that all seven transcripts fulfill the structural requirements – they are well-formed in the sense of the automaton. At the same time, the statistical analyses show typical patterns of empirical variation:

- Repetitions in the need phase (KBBd, VBBd, KBA, VBA)
- Varying lengths of phases
- Occasional phase regressions

These deviations from the ideal structure are not errors but expressions of empirical

reality. The two-layer structure allows recognizing and documenting them as such without sacrificing structural clarity.

7.2 Comparison with Purely Statistical Methods

In contrast to purely statistical methods (such as HMM or PCFG), the approach presented here offers decisive advantages:

- The structural decision is **deterministic** and not probabilistic.
- The statistical analysis is **subsequent** and does not influence the structural decision.
- Deviations are **documented**, not smoothed.
- The results are **explainable** in the strict sense of the XAI criteria.

7.3 Limitations of the Procedure

The limitations of the procedure are identical to the limitations of the underlying grammar:

- The procedure captures only the intended phases and transitions.
- More complex interaction patterns (interruptions, parallelism) require an extension of the state space.
- The statistical analysis is descriptive and does not allow causal inferences.

8 Conclusion and Outlook

This paper has shown how a strict separation of structural decidability and statistical regularity can be implemented in sequence analysis. The two-layer model of a deterministic automaton and subsequent statistics fulfills the XAI criteria of transparency, meaningfulness, and reconstructibility while allowing a realistic representation of empirical data.

The methodological significance of this approach lies in the clear distinction between what is *principally* possible (structure) and what is *empirically* frequent (statistics). This distinction is fundamental for any science pursuing both nomothetic and idiographic interests.

Further research could:

1. Extend the procedure to more complex interaction types (multi-person interactions, interruptions).
2. Complement the statistical analysis with inferential statistical methods (confidence intervals, significance tests).
3. Systematically investigate the interaction with machine learning methods.

What remains crucial throughout is methodological control: the formal structure must respect the interpretive character of the analysis and must not lead to its automation.

References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

Flick, U. (2019). *Qualitative Social Research: An Introduction* (9th ed.). Rowohlt. [German original]

Oevermann, U., Allert, T., Konau, E., & Krambeck, J. (1979). The methodology of 'objective hermeneutics' and its general research-logical significance for the social sciences. In H.-G. Soeffner (Ed.), *Interpretive Procedures in the Social and Text Sciences* (pp. 352-434). Metzler. [German original]

Przyborski, A., & Wohlrab-Sahr, M. (2021). *Qualitative Social Research: A Workbook* (5th ed.). De Gruyter Oldenbourg. [German original]

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696-735.

Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 1-10). Springer.

A The Seven Transcripts in Coded Form

A.1 Transcript 1

Original: KBG, VBG, KBBd, VBBd, KBA, VBA, KBBd, VBBd, KBA, VAA, KAA, VAV, KAV

Coded: 00000, 10000, 00100, 10100, 00101, 10101, 00100, 10100, 00101, 11001, 01001, 11100, 01100

A.2 Transcript 2

Original: VBG, KBBd, VBBd, VAA, KAA, VBG, KBBd, VAA, KAA

Coded: 10000, 00100, 10100, 11001, 01001, 10000, 00100, 11001, 01001

A.3 Transcript 3

Original: KBBd, VBBd, VAA, KAA

Coded: 00100, 10100, 11001, 01001

A.4 Transcript 4

Original: KBBd, VBBd, KBA, VBA, KBBd, VBA, KAE, VAE, KAA, VAV, KAV

Coded: 00100, 10100, 00101, 10101, 00100, 10101, 01000, 11000, 01001, 11100, 01100

A.5 Transcript 5

Original: KBG, VBG, KBBd, VBBd, KAA

Coded: 00000, 10000, 00100, 10100, 01001

A.6 Transcript 6

Original: KBBd, VBBd, KBA, VAA, KAA

Coded: 00100, 10100, 00101, 11001, 01001

A.7 Transcript 7

Original: KBG, VBBd, KBBd, VBA, VAA, KAA, VAV, KAV

Coded: 00000, 10100, 00100, 10101, 11001, 01001, 11100, 01100