Algorithmic Recursive Sequence Analysis (ARS 2.0): An Explanatory Bridge in Communication Research

Abstract

This essay describes the methodology of Algorithmic Recursive Sequence Analysis 2.0 (ARS 2.0), including its formal model, and critically compares it with established purely qualitative approaches, particularly Mayring's Qualitative Content Analysis, as well as the sole use of Large Language Models (LLMs) in data analysis. It is argued that ARS 2.0 provides an explanatory model that goes beyond the imitation of LLMs and the mere description of qualitative approaches.

1 Introduction

The analysis of natural language sequences is a central concern of many disciplines, from linguistics to communication studies and social research. While qualitative methods aim for in-depth interpretation and quantitative approaches focus on measuring frequencies and correlations, the question of explaining generative rules of social communication often remains in the background. Algorithmic Recursive Sequence Analysis 2.0 (ARS 2.0) offers an innovative approach that aims to decipher the hidden grammatical structures of dialogues. This essay describes the methodology of ARS 2.0, including its formal model, and critically compares it with established purely qualitative approaches, particularly Mayring's Qualitative Content Analysis, as well as the sole use of Large Language Models (LLMs) in data analysis. It will be argued that ARS 2.0 provides an explanatory model that goes beyond the imitation of LLMs and the mere description of qualitative approaches.

2 Methodology of Algorithmic Recursive Sequence Analysis 2.0

ARS 2.0 is a method for analyzing finite discrete sequences of characters and for inducing formal, probabilistic grammars from natural language sequences, such as those found in transcripts of sales conversations. Its overarching goal is the systematic extraction of rules that govern the sequence of interaction units

and the validation of these rules through simulation. The process is iterative and comprises several core steps:

1. **Hypothesis Generation:** Based on theoretical assumptions or initial exploratory analyses, hypotheses are formulated about the structure of interactions and potential terminal symbols (smallest meaningful units or interaction categories).

2. **Data Preparation and Symbol Assignment:** Empirical dialogue transcripts are translated into sequences of terminal symbols. This is a crucial qualitative step that requires careful content analysis and categorization of conversational contributions. For example, in sales conversations, symbols for "Buyer Greeting" (KBG) or "Seller Greeting" (VBG) could be defined.

3. **Grammar Induction:** At the core of ARS 2.0 is the algorithmic induction of a probabilistic grammar. This grammar, also referred to as a K-System, consists of production rules that describe with what probability a sequence of terminal or non-terminal symbols can be generated. This is often an iterative optimization process in which the grammar is adjusted to best represent the empirical sequences.

4. **Generation of Artificial Sequences and Simulation:** The induced grammar is used to generate a large number of artificial language sequences. This can be simulated in a multi-agent system where agents conduct dialogues based on the learned grammar.

5. **Validation and Statistical Comparison:** The generated artificial sequences are statistically compared with the original empirical sequences. This includes the analysis of frequency distributions of the terminal symbols and the calculation of correlation coefficients. The goal is to evaluate the congruence between the model and reality and to adjust the grammar if necessary to increase its explanatory power.

The **formal model of the grammar** is a K-System K, which comprises the following elements:

- An **Alphabet** $A = \{a_1, a_2, ..., a_n\}$, representing the set of all terminal symbols (e.g., KBG, VBG).

- All **words over the alphabet** A^*, which includes all possible sequences of terminal symbols.

- **Production rules** P, defined as a mapping $P := A \to A$. Each production rule $p_{a_i} \in P$ is a relation $p_{a_i} : A \times H \times A$. These rules describe how symbols follow each other in the sequence.

- An **occurrence measure** h, where $H = \{h \in \mathbb{N} | 0 \leq h \leq 100\}$ is the set of probabilities with which a particular production occurs. These probabilities reflect the empirical occurrence probabilities.

- An **axiomatic first string** $k_0 \in A^*$, which represents the starting point of a sequence.

A K-System K is formally defined as $K = (A, P, k_0)$. Starting from the axiom k_0, a K-System generates a string $k_0k_1k_2...$ by applying the production rule p to the symbol a_i of a string: $a_{i+1} := p_{a_i}(a_i)$. For a sequence $k_i := a_{i-2}a_{i-1}a_i$, the next sequence $k_{i+1} := a_{i-2}a_{i-1}a_ip_{a_i}(a_i)$ can be formed. These rules can be represented as a Context-Free Grammar. The grammar and the empirical occurrence probabilities allow for the simulation of protocols.

3 Comparison with Purely Qualitative Approaches (according to Mayring)

Qualitative Content Analysis according to Mayring is a widely used qualitative approach that also aims at systematizing the analysis of text material. It is typically theory-driven or inductive and works with category formation and coding units to identify meanings and structures in texts.

- **Similarities:**
  - Both approaches work with linguistic material and its reduction to analytical units (categories/symbols). The assignment of interactions to categories can be measured according to Mayring by the number of concordant assignments made by interpreters.
  - Both emphasize systematics and traceability of the analysis process.
  - The initial data collection and symbol assignment in ARS 2.0 show parallels to category formation and coding in qualitative content analysis.
- **Differences and Explanatory Claim:**
  - **Focus:** While Mayring's approach primarily aims at **description and interpretation** of content and structures ("What is said and how is it said?"), ARS 2.0 goes beyond this by providing a **generative explanatory model** ("By what rules can what is said be produced?").
  - **Formalization:** ARS 2.0 is significantly more formalized and mathematically grounded. The induced grammar is an explicit set of rules that enables the production of sequences. Mayring's categories are more flexible and interpretive, but do not lead to a formal, generative model.
  - **Validation:** ARS 2.0 uses statistical comparisons and correlations for model validation. Validation in qualitative content analysis is more concerned with criteria such as intersubjective comprehensibility and discussion processes.

- **Explanatory Character:** The grammar of ARS 2.0 is an **explanatory model**, as it maps the rules that generate the sequence of interaction events. Qualitative content analysis describes patterns but does not provide explicit generative explanations.

4 Comparison with the Pure Use of Large Language Models (LLMs)

LLMs have revolutionized text analysis and are increasingly used in qualitative social research. They are trained to recognize patterns in vast amounts of text and to generate coherent text.

- **Similarities:**
  - Both approaches (ARS and LLM use) deal with the analysis and potential generation of language sequences.
  - Both use computer-assisted methods for data processing.
- **Differences and Explanatory Claim:**
  - **Modeling Principle:** LLMs are at their core **imitation machines**. They learn statistical probabilities for the sequence of words and tokens, enabling them to generate convincingly human-like texts or identify patterns. However, they do not learn **explicit, interpretable grammars** or rules that could be understood as an explanation for language production. ARS 2.0, in contrast, precisely aims at the induction of such an explicit, explanatory grammar.
  - **Transparency (Opacity vs. Explainability):** LLMs are "black boxes." The reasons why an LLM generates a particular output or recognizes a pattern are often opaque to the user. The internal weights and neural connections are not directly interpretable as social or communicative rules. The grammar of ARS 2.0, on the other hand, is a **transparent and comprehensible explanatory model** whose rules can be directly interpreted.
  - **Understanding vs. Imitation:** LLMs do not "understand" dialogues in the human sense; they imitate them based on statistical correlations in their training data. The contingency and opacity of human behavior are reproduced but not causally or rule-based explained. ARS 2.0 attempts to reduce opacity by uncovering the underlying generative rules, thereby enabling a more causal understanding of communication dynamics.
  - **Quality Claim:** The uncritical use of LLMs in qualitative research carries the risk of "automated substandard work" if human, reflective interpretation is replaced by the rapid but superficial pattern recognition of AI. ARS 2.0, in contrast, demands a high degree of

methodical precision and critical reflection in symbol assignment and interpretation of the induced grammar.

5 Conclusion

Algorithmic Recursive Sequence Analysis 2.0 represents a valuable, yet underrepresented, approach in qualitative social research. It transcends the purely descriptive and interpretive level of many qualitative methods, such as Mayring's Qualitative Content Analysis, by providing a **formal, generative explanatory model in the form of a probabilistic grammar**. In contrast to the mere use of Large Language Models, which imitate dialogues but do not explain them transparently, ARS 2.0 offers insight into the underlying rules of communication.

The hesitant integration of such explanatory, formalized approaches into qualitative social research, while opaque LLMs are embraced with enthusiasm, may seem paradoxical. It could indicate that the convenience of automation and the immediate availability of tools are sometimes prioritized over methodological rigor and the pursuit of deep explanatory models. For a sustainable qualitative social research that claims both depth and relevance, a greater engagement with methods like ARS 2.0 would be desirable to move beyond mere imitation towards genuine, comprehensible explanations.