

Die Algorithmisch Rekursive Sequenzanalyse (ARS 2.0): Ein erklärender Brückenschlag in der Kommunikationsforschung

Zusammenfassung

Dieser Essay beschreibt die Methodologie der Algorithmisch Rekursiven Sequenzanalyse 2.0 (ARS 2.0) einschließlich ihres formalen Modells und vergleicht sie kritisch mit etablierten rein qualitativen Ansätzen, insbesondere der Qualitativen Inhaltsanalyse nach Mayring, sowie der reinen Nutzung von Large Language Models (LLM) in der Datenanalyse. Es wird argumentiert, dass ARS 2.0 ein Erklärungsmodell liefert, das über die Imitation von LLM und die bloße Beschreibung qualitativer Ansätze hinausgeht.

1 Einleitung

Die Analyse natürlicher Sprachsequenzen ist ein zentrales Anliegen vieler Disziplinen, von der Linguistik über die Kommunikationswissenschaft bis zur Sozialforschung. Während qualitative Methoden auf tiefgehende Interpretation abzielen und quantitative Ansätze auf die Messung von Häufigkeiten und Korrelationen fokussieren, steht die Frage nach der Erklärung generativer Regeln sozialer Kommunikation oft im Hintergrund. Die Algorithmisch Rekursive Sequenzanalyse 2.0 (ARS 2.0) bietet einen innovativen Ansatz, der darauf abzielt, die verborgenen grammatikalischen Strukturen von Dialogen zu entschlüsseln. Dieser Essay beschreibt die Methodologie der ARS 2.0 einschließlich ihres formalen Modells und vergleicht sie kritisch mit etablierten rein qualitativen Ansätzen, insbesondere der Qualitativen Inhaltsanalyse nach Mayring, sowie der reinen Nutzung von Large Language Models (LLM) in der Datenanalyse. Es wird argumentiert, dass ARS 2.0 ein Erklärungsmodell liefert, das über die Imitation von LLM und die bloße Beschreibung qualitativer Ansätze hinausgeht.

2 Methodologie der Algorithmisch Rekursiven Sequenzanalyse 2.0

Die ARS 2.0 ist eine Methode zur Analyse endlicher diskreter Zeichenketten und zur Induktion formaler, probabilistischer Grammatiken aus natürlichen Sprach-

sequenzen, wie sie beispielsweise in Transkripten von Verkaufsgesprächen vorliegen. Ihr übergeordnetes Ziel ist die systematische Extraktion von Regeln, die die Abfolge von Interaktionseinheiten steuern, und die Validierung dieser Regeln durch Simulation. Der Prozess ist iterativ und umfasst mehrere Kernschritte:

1. **Hypothesenbildung:** Ausgehend von theoretischen Vorannahmen oder ersten explorativen Analysen werden Hypothesen über die Struktur der Interaktionen und potenzielle Terminalsymbole (kleinste bedeutungstragende Einheiten oder Interaktionskategorien) aufgestellt.
2. **Datenaufbereitung und Symbolzuweisung:** Empirische Dialogtranskripte werden in Sequenzen von Terminalsymbolen übersetzt. Dies ist ein entscheidender qualitativer Schritt, der eine sorgfältige Inhaltsanalyse und Kategorisierung der Gesprächsbeiträge erfordert. Beispielsweise könnten in Verkaufsgesprächen Symbole für "Käufer Begrüßung"(KBG) oder "Verkäufer Begrüßung"(VBG) definiert werden.
3. **Grammatikinduktion:** Im Zentrum der ARS 2.0 steht die algorithmische Induktion einer probabilistischen Grammatik. Diese Grammatik, auch als K-System bezeichnet, besteht aus Produktionsregeln, die beschreiben, mit welcher Wahrscheinlichkeit eine Abfolge von Terminal- oder nicht-terminalen Symbolen erzeugt werden kann. Dies ist oft ein iterativer Optimierungsprozess, bei dem die Grammatik so angepasst wird, dass sie die empirischen Sequenzen möglichst gut abbildet.
4. **Generierung künstlicher Sequenzen und Simulation:** Die induzierte Grammatik wird verwendet, um eine große Anzahl künstlicher Sprachsequenzen zu generieren. Dies kann in einem Multi-Agenten-System simuliert werden, in dem Agenten Dialoge auf Basis der gelernten Grammatik führen.
5. **Validierung und statistischer Vergleich:** Die generierten künstlichen Sequenzen werden statistisch mit den ursprünglichen empirischen Sequenzen verglichen. Dies umfasst die Analyse von Häufigkeitsverteilungen der Terminalsymbole und die Berechnung von Korrelationskoeffizienten. Ziel ist es, die Übereinstimmung zwischen dem Modell und der Realität zu bewerten und die Grammatik bei Bedarf anzupassen, um die Erklärungskraft zu erhöhen.

Das **formale Modell der Grammatik** ist ein K-System K , das folgende Elemente umfasst:

- Ein **Alphabet** $A = \{a_1, a_2, \dots, a_n\}$, das die Menge aller Terminalsymbole (z.B. KBG, VBG) repräsentiert.
- Alle **Worte über dem Alphabet** A^* , was alle möglichen Sequenzen der Terminalsymbole umfasst.

- **Produktionsregeln** P , definiert als eine Abbildung $P := A \rightarrow A$. Jede Produktionsregel $p_{a_i} \in P$ ist dabei eine Relation $p_{a_i} : A \times H \times A$. Diese Regeln beschreiben, wie Symbole in der Sequenz aufeinanderfolgen.
- Ein **Auftrittsmaß** h , wobei $H = \{h \in \mathbb{N} | 0 \leq h \leq 100\}$ die Menge der Wahrscheinlichkeiten ist, mit der eine bestimmte Produktion auftritt. Diese Wahrscheinlichkeiten spiegeln die empirischen Auftrittswahrscheinlichkeiten wider.
- Eine **axiomatische erste Zeichenkette** $k_0 \in A^*$, die den Startpunkt einer Sequenz darstellt.

Ein K-System K wird formal definiert als $K = (A, P, k_0)$. Ausgehend vom Axiom k_0 erzeugt ein K-System eine Zeichenkette $k_0 k_1 k_2 \dots$, indem die Produktionsregel p auf das Zeichen a_i einer Kette angewendet wird: $a_{i+1} := p_{a_i}(a_i)$. Für eine Sequenz $k_i := a_{i-2} a_{i-1} a_i$ kann die nächste Sequenz $k_{i+1} := a_{i-2} a_{i-1} a_i p_{a_i}(a_i)$ gebildet werden. Diese Regeln können als Kontextfreie Grammatik dargestellt werden. Die Grammatik und die empirischen Auftrittswahrscheinlichkeiten ermöglichen die Simulation von Protokollen.

3 Vergleich mit rein qualitativen Ansätzen (nach Mayring)

Die **Qualitative Inhaltsanalyse nach Mayring** ist ein weit verbreiteter qualitativer Ansatz, der ebenfalls auf die Systematisierung der Analyse von Textmaterial abzielt. Sie ist typischerweise theoriegeleitet oder induktiv und arbeitet mit Kategorienbildung und Kodiereinheiten, um Bedeutungen und Strukturen in Texten zu identifizieren.

- **Ähnlichkeiten:**
 - Beide Ansätze arbeiten mit Sprachmaterial und dessen Reduktion auf analytische Einheiten (Kategorien/Symbole). Die Zuordnung der Interakte zu Kategorien kann nach Mayring durch die Anzahl der übereinstimmend vorgenommenen Zuordnungen von Interpreten gemessen werden.
 - Beide legen Wert auf Systematik und Nachvollziehbarkeit des Analyseprozesses.
 - Die initiale Datenerfassung und Symbolzuweisung in ARS 2.0 hat Parallelen zur Kategorienbildung und Kodierung in der qualitativen Inhaltsanalyse.
- **Unterschiede und Erklärungsanspruch:**
 - **Fokus:** Während Mayrings Ansatz primär auf **Beschreibung und Interpretation** der Inhalte und Strukturen abzielt ("Was wird gesagt und wie wird es gesagt?"), geht ARS 2.0 darüber hinaus, indem

es ein **generatives Erklärungsmodell** liefert (Nach welchen Regeln kann das Gesagte produziert werden?).

- **Formalisierung:** ARS 2.0 ist deutlich stärker formalisiert und mathematisch fundiert. Die induzierte Grammatik ist ein explizites Regelwerk, das eine Produktion von Sequenzen ermöglicht. Mayrings Kategorien sind flexibler und interpretativer, führen aber nicht zu einem formalen, generativen Modell.
- **Validierung:** ARS 2.0 nutzt statistische Vergleiche und Korrelationen zur Validierung des Modells. Die Validierung in der qualitativen Inhaltsanalyse erfolgt eher über Kriterien wie intersubjektive Nachvollziehbarkeit und Diskussionsprozesse.
- **Erklärungscharakter:** Die Grammatik der ARS 2.0 ist ein **erklärendes Modell**, da sie die Regeln abbildet, die die Abfolge von Interaktionsereignissen erzeugen. Die Qualitative Inhaltsanalyse beschreibt Muster, bietet aber keine expliziten generativen Erklärungen.

4 Vergleich mit der reinen Nutzung von Large Language Models (LLM)

LLM haben die Textanalyse revolutioniert und werden zunehmend in der qualitativen Sozialforschung eingesetzt. Sie sind darauf trainiert, Muster in riesigen Textmengen zu erkennen und kohärenten Text zu generieren.

- **Ähnlichkeiten:**

- Beide Ansätze (ARS und LLM-Nutzung) befassen sich mit der Analyse und potenziellen Generierung von Sprachsequenzen.
- Beide nutzen computergestützte Verfahren zur Datenverarbeitung.

- **Unterschiede und Erklärungsanspruch:**

- **Modellierungsprinzip:** LLM sind im Kern **Imitationsmaschinen**. Sie lernen statistische Wahrscheinlichkeiten für die Abfolge von Wörtern und Token und können dadurch überzeugend menschlich klingende Texte generieren oder Muster identifizieren. Sie lernen jedoch keine **expliziten, interpretierbaren Grammatiken** oder Regeln, die man als Erklärung für die Sprachproduktion verstehen könnte. ARS 2.0 hingegen zielt genau auf die Induktion einer solchen expliziten, erklärenden Grammatik ab.
- **Transparenz (Opazität vs. Erklärbarkeit):** LLM sind "Black Boxes". Die Gründe, warum ein LLM eine bestimmte Ausgabe generiert oder ein Muster erkennt, sind für den Nutzer oft undurchsichtig. Die internen Gewichte und neuronalen Verbindungen sind nicht direkt als

soziale oder kommunikative Regeln interpretierbar. Die Grammatik der ARS 2.0 ist hingegen ein **transparentes und nachvollziehbares Erklärungsmodell**, dessen Regeln direkt interpretiert werden können.

- **Verständnis vs. Imitation:** LLM "verstehen" Dialoge nicht im menschlichen Sinne; sie imitieren sie basierend auf statistischen Korrelationen in ihren Trainingsdaten. Die Kontingenz und Opazität des menschlichen Verhaltens werden reproduziert, aber nicht kausal oder regelbasiert erklärt. ARS 2.0 versucht, die Opazität durch die Aufdeckung der zugrunde liegenden generativen Regeln zu reduzieren und dadurch ein kausaleres Verständnis der Kommunikationsdynamik zu ermöglichen.
- **Qualitätsanspruch:** Der unkritische Einsatz von LLM in der qualitativen Forschung birgt die Gefahr der "Automatisierten schlechteren Arbeit", wenn die menschliche, reflexive Interpretation durch die schnelle, aber oberflächliche Mustererkennung der KI ersetzt wird. ARS 2.0 fordert hingegen einen hohen Grad an methodischer Präzision und kritischer Reflexion bei der Symbolzuweisung und der Interpretation der induzierten Grammatik.

5 Fazit

Die Algorithmisch Rekursive Sequenzanalyse 2.0 stellt einen wertvollen, aber bisher in der qualitativen Sozialforschung unterrepräsentierten Ansatz dar. Sie überwindet die rein beschreibende und interpretierende Ebene vieler qualitativer Methoden, wie der Qualitativen Inhaltsanalyse nach Mayring, indem sie ein **formales, generatives Erklärungsmodell in Form einer probabilistischen Grammatik** liefert. Im Gegensatz zur bloßen Nutzung von Large Language Models, die Dialoge imitieren, aber nicht transparent erklären, bietet ARS 2.0 einen Einblick in die zugrunde liegenden Regeln der Kommunikation.

Das zögerliche Einbetten solcher erklärenden, formalisierten Ansätze in der qualitativen Sozialforschung, während gleichzeitig opake LLM mit Begeisterung aufgenommen werden, mag paradox erscheinen. Es könnte ein Hinweis darauf sein, dass der Komfort der Automatisierung und die unmittelbare Verfügbarkeit von Tools manchmal über die methodische Stringenz und den Anspruch an tiefgehende Erklärungsmodelle gestellt werden. Für eine zukunftsfähige qualitative Sozialforschung, die sowohl Tiefe als auch Relevanz beansprucht, wäre eine stärkere Auseinandersetzung mit Methoden wie der ARS 2.0 wünschenswert, um über die reine Imitation hinaus zu echten, nachvollziehbaren Erklärungen zu gelangen.