Die Mehrdeutigkeit von Homomorphismen in freien Monoiden und ihr Einfluß auf algorithmische Eigenschaften von Patternsprachen*

Daniel Reidenbach

Fachbereich Informatik, Technische Universität Kaiserslautern,
Postfach 3049, 67653 Kaiserslautern
reidenba@informatik.uni-kl.de

Abstract: Die vorliegende Arbeit untersucht eine fundamentale kombinatorische Eigenschaft von Homomorphismen in freien Monoiden, nämlich ihre *Mehrdeutigkeit*. Dieser Begriff bezeichnet den Umstand, daß zu einem gegebenen Wort α und einem Homomorphismus σ durchaus ein zweiter Homomorphismus τ existieren kann, der α auf dasselbe Wort abbildet wie σ – es gilt also $\sigma(\alpha) = \tau(\alpha)$, obwohl ein Symbol x in α existiert, für das sich $\sigma(x)$ von $\tau(x)$ unterscheidet.

Aufgrund ihres elementaren Charakters ist Mehrdeutigkeit von Homomorphismen eng verwoben mit einer Fülle von wichtigen Themen der Informatik. So stellt sie nicht nur die Grundlage des *Postschen Korrespondenzproblems* dar, sondern beeinflußt auch etliche Eigenschaften von *Patternsprachen*, welche insbesondere in der algorithmischen Lerntheorie von großer Bedeutung sind. Die kombinatorischen Hauptergebnisse der Arbeit – insbesondere zur *Existenz* von *eindeutigen* und sogenannten *moderat mehrdeutigen* Homomorphismen – erlauben daher diverse nichttriviale Rückschlüsse zu einigen klassischen Problemen für Patternsprachen.

1 Motivation und erste Beobachtungen

Im Rahmen der vorliegenden Arbeit werden spezielle Funktionen (nämlich Homomorphismen, s. u.) untersucht, die ein endliches Wort über einem unendlichen Alphabet Δ auf ein endliches Wort über einem wenigstens zweielementigen Alphabet Σ abbilden. Im Sinne einer möglichst klaren Terminologie werde im folgenden ein Wort über Δ als *Pattern* bezeichnet und ein Symbol in Δ als *Variable*. Außerdem diene die Menge $\mathbb{N}$ der natürlichen Zahlen als Alphabet Δ; es ist also beispielsweise $1 \cdot 2 \cdot 12 \cdot 1$ ein Pattern aus insgesamt vier Variablen (wobei das Zeichen $\cdot$ als Trennsymbol fungiert). Das Symbol ε stehe für das *leere Wort* und $|X|$ für die Mächtigkeit einer Menge X bzw. die Länge eines Pattern X.

Eine Abbildung $\sigma : \mathbb{N}^* \to \Sigma^*$ ist genau dann ein *Homomorphismus*, wenn für alle Pattern $\alpha, \beta \in \mathbb{N}^*$ gilt: $\sigma(\alpha \cdot \beta) = \sigma(\alpha) \cdot \sigma(\beta)$. Anschaulich impliziert dies, daß σ ein Pattern α „zeichenweise“ abbildet: das Bild von α unter σ setzt sich zusammen aus dem Bild

*Dieser Artikel präsentiert ausgewählte Resultate der Dissertation [Rei06a] des Autors, welche wiederum auf den Zeitschriftenpublikationen [Rei06b, FRS06, Reib, Reia] fußt.

der ersten Variablen in α unter σ, gefolgt vom Bild der zweiten Variablen in α usw. Ein Homomorphismus ist somit bereits dann für alle Pattern in $\mathbb{N}^*$ vollständig definiert, wenn er für alle Variablen in $\mathbb{N}$ definiert ist. Da Homomorphismen also die Struktur ihrer Eingabe berücksichtigen, kann ein Wort $w \in \Sigma^*$ nur dann das Bild eines Pattern $\alpha \in \mathbb{N}^*$ sein, wenn w und α gewisse strukturelle Gemeinsamkeiten aufweisen.

Aufgrund ihrer intuitiven Definition und ihres strukturerhaltenden Potentials stellen Homomorphismen das Fundament zahlreicher und wichtiger Gebiete der Informatik dar, wie z. B. Kodierungstheorie (s. [BP85]), Wortgleichungen (s. [Lot02]), *morphic sequences* (s. [AS03]), D0L-Systeme (s. [KRS97]), *equality sets* (und das daraus abgeleitete Postsche Korrespondenzproblem, s. [HK97]) und Patternsprachen (d. h. die Mengen aller Bilder von Pattern unter beliebigen Homomorphismen; s. Kapitel 2 und [MS97]). Insbesondere im Rahmen der beiden letztgenannten Bereiche – wo nämlich Bilder ein und desselben Pattern unter *verschiedenen* Homomorphismen betrachtet werden – kommt eine elementare Eigenschaft von Homomorphismen zum Tragen: ihre (mögliche) Mehrdeutigkeit. Ein Homomorphismus $\sigma : \mathbb{N}^* \to \Sigma^*$ heißt *mehrdeutig (für ein Pattern $\alpha \in \mathbb{N}^*$)*, wenn ein Homomorphismus $\tau : \mathbb{N}^* \to \Sigma^*$ existiert, der $\tau(\alpha) = \sigma(\alpha)$ erfüllt, obwohl es eine Variable x in α gibt, für die $\tau(x) \neq \sigma(x)$ gilt; anderenfalls heißt σ *eindeutig (für α)*.

Wenn man nun die Mehrdeutigkeit von Homomorphismen $\sigma : \mathbb{N}^* \to \Sigma^*$ untersucht (und zusätzlich voraussetzt, daß σ *nichtlöschend* sein soll, also keine Variable im Pattern auf das leere Wort abbildet), dann läßt sich leicht beobachten, daß gewisse Pattern selbst bei beliebiger Wahl von $\Sigma \subseteq \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \ldots\}$ überhaupt nicht eindeutig abgebildet werden können. So gilt beispielsweise für das Pattern $\alpha_0 := 1 \cdot 2$ und jeden nichtlöschenden Homomorphismus σ, daß z. B. der Homomorphismus $\tau : \mathbb{N}^* \to \Sigma^*$, definiert durch $\tau(1) := \sigma(\alpha)$ und $\tau(x) := \varepsilon$, $x \in \mathbb{N} \setminus \{1\}$, α_0 auf dasselbe Wort abbildet wie σ, obwohl er sich von σ unterscheidet. Für andere Pattern läßt sich hingegen trivialerweise eine Fülle von eindeutigen Homomorphismen finden. So ist z. B. jeder Homomorphismus eindeutig für solche Pattern α, für die $|\operatorname{var}(\alpha)| = 1$ gilt (wobei $\operatorname{var}(\alpha)$ die Menge der in α auftretenden Variablen bezeichne). Zusätzlich kann beobachtet werden, daß auch bestimmte komplexere Pattern eindeutig abgebildet werden können. Betrachtet man beispielsweise $\alpha_1 := 1 \cdot 2 \cdot 3 \cdot 4 \cdot 1 \cdot 4 \cdot 3 \cdot 2$ und das Wort $w := \mathtt{a\,a\,b\,b\,b\,a\,a\,b\,a\,b\,a\,b}$, so kann man durch Ausprobieren leicht verifizieren, daß α_1 auf genau eine Art durch einen Homomorphismus σ auf w abgebildet werden kann, nämlich dann, wenn $\sigma(1) = \mathtt{a}$, $\sigma(2) = \mathtt{a\,b}$, $\sigma(3) = \mathtt{b}$ und $\sigma(4) = \mathtt{b\,a}$ erfüllt ist. Andere Homomorphismen, wie beispielsweise der prominente Suffixcode $\sigma_{\mathrm{cod}} : \mathbb{N}^* \to \{\mathtt{a}, \mathtt{b}\}^*$, gegeben durch $\sigma_{\mathrm{cod}}(x) := \mathtt{ab}^x$, $x \in \mathbb{N}$, sind wiederum mehrdeutig für α_1, wie durch den im nachfolgenden Diagramm implizit definierten Homomorphismus τ (der die Variable 2 auf das leere Wort abbildet) belegt wird:

$$\overbrace{\underbrace{\mathtt{a\,b}}_{\sigma_{\mathrm{cod}}(1)}\ \underbrace{\mathtt{a\,b\,b}}_{\sigma_{\mathrm{cod}}(2)}}^{\tau(1)}\ \overbrace{\underbrace{\mathtt{a\,b\,b\,b}}_{\sigma_{\mathrm{cod}}(3)}\ \underbrace{\mathtt{a\,b\,b}}_{\sigma_{\mathrm{cod}}(4)}}^{\tau(3)}\ \overbrace{\underbrace{\mathtt{b\,b}}_{\sigma_{\mathrm{cod}}(4)}}^{\tau(4)}\ \overbrace{\underbrace{\mathtt{a\,b}}_{\sigma_{\mathrm{cod}}(1)}}^{\tau(1)}\ \overbrace{\underbrace{\mathtt{a\,b\,b}}_{\sigma_{\mathrm{cod}}(4)}}^{\tau(4)}\ \overbrace{\underbrace{\mathtt{b\,b}}_{\sigma_{\mathrm{cod}}(4)}}^{\tau(4)}\ \overbrace{\underbrace{\mathtt{a\,b\,b\,b}}_{\sigma_{\mathrm{cod}}(3)}\ \underbrace{\mathtt{a\,b\,b}}_{\sigma_{\mathrm{cod}}(2)}}^{\tau(3)}$$

Mit Bezug auf die Mehrdeutigkeit von Homomorphismen stellt sich also zunächst einmal die nichttriviale Frage nach der *Existenz* von *eindeutigen* Homomorphismen und es zeichnet sich ab, daß solche Homomorphismen – so sie denn existieren – mit Bedacht gewählt werden müssen.

Frage 1. *Sei Σ ein Alphabet, $|\Sigma| \geq 2$, und sei $\alpha \in \mathbb{N}^+$. Gibt es einen nichtlöschenden Homomorphismus $\sigma : \mathbb{N}^* \to \Sigma^*$, der eindeutig ist für α?*

Betrachtet man die gängige Definition eines Pattern in der Literatur zu Patternsprachen, so fällt auf, daß hierunter im allgemeinen eine endliche Zeichenkette verstanden wird, die aus Variablen *und* aus den sogenannten *Terminalsymbolen* in Σ besteht (weswegen Pattern in $\mathbb{N}^+$ in der Regel auch als *terminalfrei* bezeichnet werden). Für die betrachteten Homomorphismen, auf deren Grundlage die Patternsprache eines solchen Pattern definiert ist, bedeutet dies, daß sie *terminalerhaltend* sein müssen, also die Terminalsymbole im Pattern auf sich selbst abbilden. Im folgenden soll daher für beliebige Alphabete Σ ein *Pattern (über Σ)* eine Zeichenkette in $(\mathbb{N} \cup \Sigma)^+$ sein, und es werden anstelle der bisher untersuchten, allgemeinen Homomorphismen nun *Substitutionen*, d. h. terminalerhaltende Homomorphismen $\sigma : (\mathbb{N} \cup \Sigma)^* \to \Sigma^*$ betrachtet. Derartige Substitutionen sind im übrigen nicht nur aus dem Bereich der Patternsprachen bekannt, sondern stellen auch das definitorische Fundament von Wortgleichungen dar, es handelt sich hierbei also um eine verbreitete Abwandlung der gewöhnlichen Homomorphismen.

In Hinblick auf die Mehrdeutigkeit von Substitutionen läßt sich anhand von einfachen Beispielen zeigen, daß sie erhebliche Unterschiede zu der von allgemeinen Homomorphismen aufweisen kann. So gilt – wie oben erläutert – beispielsweise für das terminalfreie Pattern $\alpha_0 = 1 \cdot 2$, daß es überhaupt nicht eindeutig abgebildet werden kann, wohingegen sich für das daraus abgeleitete Pattern $\alpha_0' := 1 \cdot \mathtt{a} \cdot 2$ ohne größeren Aufwand eindeutige Substitutionen finden lassen, wie z. B. $\sigma : (\mathbb{N} \cup \{\mathtt{a}, \mathtt{b}\})^* \to \{\mathtt{a}, \mathtt{b}\}^*$, gegeben durch $\sigma(1) := \sigma(2) := \mathtt{b}$. Umgekehrt ist der allgemeine Homomorphismus $\sigma : \mathbb{N}^* \to \{\mathtt{a}, \mathtt{b}\}^*$, definiert durch $\sigma(1) := \mathtt{ab}$ und $\sigma(2) := \mathtt{ba}$, zwar eindeutig für das terminalfreie Pattern $\alpha_2 := 1 \cdot 1 \cdot 2 \cdot 2$, aber die entsprechende Substitution σ ist mehrdeutig für das Pattern $\alpha_2' := 1 \cdot 1 \cdot \mathtt{a} \cdot 2 \cdot 2$, weil es eine zweite Substitution $\tau : (\mathbb{N} \cup \{\mathtt{a}, \mathtt{b}\})^* \to \{\mathtt{a}, \mathtt{b}\}^*$ gibt, die α_2' auf dasselbe Wort abbildet:

$$\underbrace{\overbrace{\mathtt{a}\,\mathtt{b}}^{\sigma(1)}\;\overbrace{\mathtt{a}\,\mathtt{b}}^{\sigma(1)}}_{\tau(1)}\;\underbrace{\mathtt{a}\;\overbrace{\mathtt{b}\,\mathtt{a}}^{\sigma(2)}\;\overbrace{\mathtt{b}}^{\sigma(2)}\,\mathtt{a}}_{\tau(1)}$$

Es deutet sich somit an, daß die Mehrdeutigkeit von Substitutionen ein eigenständiges Problem darstellt, welches nicht ohne weiteres durch Erkenntnisse zur Mehrdeutigkeit von allgemeinen Homomorphismen gelöst werden kann:

Frage 2. *Sei Σ ein Alphabet, $|\Sigma| \geq 2$, und sei $\alpha \in (\mathbb{N} \cup \Sigma)^+ \setminus \mathbb{N}^+$. Gibt es eine nichtlöschende Substitution $\sigma : (\mathbb{N} \cup \Sigma)^* \to \Sigma^*$, die eindeutig ist für α?*

Die Mehrdeutigkeit von allgemeinen und terminalerhaltenden Homomorphismen ist also ein elementarer Themenkomplex der Wortkombinatorik, in dem etliche grundlegende Fragen nichttrivialer Natur sind. In den nachfolgenden Kapiteln soll vor allem die Existenz von eindeutigen Homomorphismen diskutiert werden, wobei sich Kapitel 3 hauptsächlich mit Frage 1 und Kapitel 4 mit Frage 2 befaßt.

Aufgrund ihrer einfachen Definition und der großen Bedeutung von Homomorphismen in freien Monoiden weist der Hauptgegenstand dieser Arbeit diverse Querbezüge zu anderen

Feldern der Informatik und der diskreten Mathematik auf. Die in den nachfolgenden Kapiteln präsentierten kombinatorischen Resultate implizieren daher sogar einen erheblichen Erkenntnisgewinn zu wohlbekannten Problemen in anderen Domänen. Dies soll insbesondere für die Frage nach der algorithmischen Erlernbarkeit von Patternsprachen explizit diskutiert werden. Zu diesem Zweck werden im folgenden Abschnitt einige weitere elementare Begriffe eingeführt.

2 Grundlegende Begriffe und Erkenntnisse

Aus Platzgründen sollen hier die bereits im vorigen Kapitel hinreichend präzise eingeführten Begriffe und Symbole nicht erneut definiert werden. In bezug auf grundlegende Notationen sei deshalb lediglich noch erwähnt, daß im folgenden die Anzahl der Vorkommen einer Variablen x in einem Pattern α mit $|\alpha|_x$ bezeichnet wird; es gilt also beispielsweise $|1 \cdot 2 \cdot 12 \cdot 1|_1 = 2$. Für ein beliebiges Alphabet Σ beschreibt außerdem der Begriff *Teilwort (eines Wortes $w \in \Sigma^*$)* – in der Literatur auch häufig *Faktor* genannt – ein Wort $v \in \Sigma^*$, das $w = v_1\, v\, v_2$ erfüllt, $v_1, v_2 \in \Sigma^*$. Für andere, nicht näher erläuterte Begriffe sei auf [RS97] verwiesen.

Hinsichtlich des Konzeptes einer *Patternsprache* (eingeführt in [Ang80a] und [Shi82]) erfordern die Ausführungen in den Kapiteln 3 und 4 ein etwas solideres Fundament als das in Kapitel 1 geschaffene. Sei also Σ ein Alphabet und $\alpha \in (\mathbb{N} \cup \Sigma)^+$ ein Pattern. Dann werden in der Literatur grundsätzlich zwei verschiedene Arten von Patternsprachen von α unterschieden: die *E-Patternsprache*

$$L_{\mathrm{E},\Sigma}(\alpha) = \{w \in \Sigma \mid w = \sigma(\alpha) \text{ für eine Substitution } \sigma : (\mathbb{N} \cup \Sigma)^* \to \Sigma^*\}$$

(wobei „E“ für *erweitert* bzw. im Englischen für *extended* oder *erasing* steht) und die *NE-Patternsprache* (kurz für *Nonerasing*-Patternsprache)

$$L_{\mathrm{NE},\Sigma}(\alpha) = \{w \in \Sigma \mid w = \sigma(\alpha) \text{ für eine Substitution } \sigma : (\mathbb{N} \cup \Sigma)^+ \to \Sigma^+\}.$$

Bei E-Patternsprachen sind also auch jene Substitutionen zugelassen, die beliebig viele Variablen in α auf das leere Wort abbilden, während sich die Definition der NE-Patternsprache auf nichtlöschende Substitutionen beschränkt. Im Rahmen der vorliegenden Arbeit werden primär die E-Patternsprachen betrachtet. Wird eine Patternsprache von einem terminalfreien Pattern erzeugt, so spricht man von einer *terminalfreien Patternsprache*.

Eine besondere Bedeutung kommt Patternsprachen im Rahmen der *Induktiven Inferenz* (s. [AS83]) – einem Ansatz der *Algorithmischen Lerntheorie* – zu, wo untersucht wird, ob und gegebenenfalls wie sich ein erzeugendes Pattern einer Patternsprache L aus den Wörtern in L algorithmisch rekonstruieren läßt. Die Überlegungen der vorliegenden Arbeit stützen sich in dieser Hinsicht auf das elementare Modell des *Lernen im Limes (anhand von positiven Beispielen)*, welches auf [Gol67] zurückgeht und – grob gesagt – eine Klasse $\mathrm{PAT}_\star$ von Patternsprachen dann als lernbar betrachtet, wenn eine berechenbare sogenannte *Lernstrategie* existiert, die für jede Sprache $L \in \mathrm{PAT}_\star$ und bei schrittweiser Eingabe jeder beliebigen unendlichen Sequenz $w_1, w_2, \ldots$ von Wörtern, die $L = \{w_i \mid i \in \mathbb{N}\}$ erfüllt, gegen ein Pattern α *konvergiert*, das genau L erzeugt.

Aus einem charakteristischen Kriterium in [Ang80b] zum Lernen im Limes läßt sich ablesen, daß eine Klasse $\mathrm{PAT}_\star$ von Patternsprachen genau dann erlernbar ist, wenn eine Prozedur existiert, die zu jeder Sprache L in $\mathrm{PAT}_\star$ einen *Telltale* T_L *(in bezug auf* $\mathrm{PAT}_\star$*)* aufzählt. Hierbei handelt es sich formal um eine endliche Teilsprache von L, durch die man L von all ihren echten Teilsprachen in $\mathrm{PAT}_\star$ unterscheiden kann. Anschaulich kann T_L als eine Menge von Wörtern interpretiert werden, die genügend *Information* über L enthalten, um die Struktur eines erzeugenden Pattern von L aus T_L eindeutig zu rekonstruieren.

Der Kenntnisstand zur Lernbarkeit von Patternsprachen weist in Abhängigkeit von der Frage, ob NE- oder E-Patternsprachen betrachtet werden, bemerkenswerte Unterschiede auf. Für die Klasse nePAT_Σ aller NE-Patternsprachen über einem beliebigen Alphabet Σ ist wohlbekannt, daß sie erlernbar ist:

Satz 1 ([Ang80a])**.** *Sei* Σ *ein Alphabet. Dann ist* nePAT_Σ *im Limes anhand von positiven Beispielen erlernbar.*

Aufbauend auf dieser Einsicht existieren eine Fülle von weiteren Arbeiten, so z. B. [LW91, WZ94, RZ00, RZ01], die sich insbesondere mit effizienten Lernstrategien für nePAT_Σ (und ausgewählte Teilklassen) befassen. Trotz intensiver Bemühungen in den vergangenen Jahrzehnten ist hingegen die Frage nach Lernbarkeit der Klasse ePAT_Σ aller E-Patternsprachen – die als „one of the outstanding open problems in inductive inference“ gilt (s. [Mit98]) – immer noch nicht abschließend beantwortet. Der diesbezüglich im Rahmen von [Rei06a] erzielte Erkenntnisgewinn wird in den Abschnitten 3.1 und 4.1 beschrieben.

3 Allgemeine Homomorphismen

Das vorliegende Kapitel beschäftigt sich mit der Mehrdeutigkeit von allgemeinen Homomorphismen $\sigma : \mathbb{N}^* \to \Sigma^*$ und insbesondere mit Frage 1. Da in Kapitel 1 bereits zwei Beispielpattern in $\mathbb{N}^+$ eingeführt worden sind, von denen das eine (nämlich $\alpha_0 = 1 \cdot 2$) überhaupt nicht und das andere (nämlich $\alpha_1 = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 1 \cdot 4 \cdot 3 \cdot 2$) sehr wohl eindeutig abgebildet werden kann, ist es notwendig, diese Beispiele zielgerichtet zu generalisieren. Diese Aufgabe übernimmt – wie noch zu belegen ist – die folgende Definition:

Definition 1. Sei $\alpha \in \mathbb{N}^+$. Dann heißt α *m-imprimitiv* (für: *morphically imprimitive*), falls es ein $\beta \in \mathbb{N}^+$ und Homomorphismen $\phi, \psi : \mathbb{N}^* \to \mathbb{N}^*$ gibt, so daß $|\beta| < |\alpha|$, $\phi(\alpha) = \beta$ und $\psi(\beta) = \alpha$. Entsprechend ist α *m-primitiv*, wenn es nicht m-imprimitiv ist.

Mit Hinweis auf das Pattern $\beta := 1$ läßt sich leicht sehen, daß das oben eingeführte Pattern α_0 m-imprimitiv ist. Tatsächlich stellt sich heraus, daß die m-imprimitiven Pattern im vorliegenden Kontext eine Verallgemeinerung von α_0 sind – sie können also von *keinem* nichtlöschenden Homomorphismus eindeutig abgebildet werden:

Satz 2. *Sei* Σ *ein Alphabet. Dann ist jeder nichtlöschende Homomorphismus* $\sigma : \mathbb{N}^* \to \Sigma^*$ *mehrdeutig für jedes m-imprimitive* $\alpha \in \mathbb{N}^+$.

Die Suche nach eindeutig abbildbaren Pattern muß sich also zwangsläufig auf die m-primitiven beschränken (welche im übrigen das oben definierte Pattern α_1 umfassen). In

Hinblick auf diese stellt sich zunächst jedoch heraus, daß es keinen *einzelnen* nichtlöschenden Homomorphismus geben kann, der *alle* m-primitiven Pattern eindeutig abbildet:

Satz 3. *Sei Σ ein Alphabet. Dann gibt es keinen nichtlöschenden Homomorphismus $\sigma : \mathbb{N}^* \to \Sigma^*$, der für jedes m-primitive $\alpha \in \mathbb{N}^+$ eindeutig ist.*

Falls also für ein beliebiges m-primitives Pattern α ein eindeutiger Homomorphismus existiert, so muß dieser für die Struktur von α maßgeschneidert werden. Zu diesem Zweck ist es notwendig, ein ausgefeiltes technisches Instrumentarium einzuführen, mit dessen Hilfe die Nachbarschaftsbeziehungen zwischen den Variablen in α analysiert werden können und das letztlich zu einer speziellen Partition von $\mathrm{var}(\alpha)$ führt. In Abhängigkeit von dieser Partition kann dann ein komplexer (und hier aus Platzgründen nicht angegebener) Homomorphismus $\sigma_{\mathrm{un},\alpha} : \mathbb{N}^* \to \{\mathtt{a}, \mathtt{b}\}^*$ definiert werden, der sich dadurch auszeichnet, daß er *heterogen (für α)* ist, was bedeutet, daß er gewisse Variablen in α auf ein Wort abbildet, das mit dem Buchstaben $\mathtt{a}$ beginnt (bzw. endet), während das Bild von anderen Variablen in α mit $\mathtt{b}$ beginnt (bzw. endet). Unter anderem aufgrund eben dieser Eigenschaft läßt sich nachweisen, daß $\sigma_{\mathrm{un},\alpha}$ tatsächlich eindeutig ist für α. Unter Berücksichtigung von Satz 2 lassen sich somit diejenigen Pattern charakterisieren, die von nichtlöschenden Homomorphismen eindeutig abgebildet werden können. Da $\sigma_{\mathrm{un},\alpha}$ zusätzlich noch injektiv ist, lautet die abschließende Antwort auf Frage 1 wie folgt:

Satz 4. *Sei Σ ein Alphabet und $\alpha \in \mathbb{N}^+$. Es existiert genau dann ein injektiver Homomorphismus $\sigma : \mathbb{N}^* \to \Sigma^*$, der eindeutig ist für α, wenn α m-primitiv ist.*

Neben der oben angesprochenen Heterogenität weist $\sigma_{\mathrm{un},\alpha}$ ein zweites wichtiges Merkmal auf, das überdies auch seine Injektivität garantiert: Er bildet jede Variable in α auf ein Wort ab, das aus *drei* eindeutigen Teilwörtern (sogenannten *Segmenten*) $\mathtt{ab}^+\mathtt{a}$ oder (im Dienste der Heterogenität) $\mathtt{ba}^+\mathtt{b}$ besteht. Für jedes m-primitive Pattern α ist $\sigma_{\mathrm{un},\alpha}$ also eine heterogene Abwandlung des *homogenen* (d. h. nicht-heterogenen) Homomorphismus $\sigma_{3\text{-seg}} : \mathbb{N}^* \to \{\mathtt{a}, \mathtt{b}\}^*$ mit

$$\sigma_{3\text{-seg}}(x) := \mathtt{ab}^{3x-2}\mathtt{a}\,\mathtt{ab}^{3x-1}\mathtt{a}\,\mathtt{ab}^{3x}\mathtt{a},$$

$x \in \mathbb{N}$. Diesem Homomorphismus kommt deshalb eine besondere Bedeutung zu, weil er trotz seiner Homogenität bereits eine stark *eingeschränkte Mehrdeutigkeit* an den Tag legt. Es gilt nämlich für jedes m-primitive Pattern α, für jeden Homomorphismus τ mit $\tau(\alpha) = \sigma_{3\text{-seg}}(\alpha)$ und für jede Variable $x \in \mathrm{var}(\alpha)$, daß $\tau(x)$ wenigstens das mittlere Segment von $\sigma_{3\text{-seg}}(x)$ enthalten muß. Der Homomorphismus $\sigma_{3\text{-seg}}$ erfüllt daher ein (technisch komplexes und deshalb hier nicht weiter ausgeführtes) Kriterium, das in [Rei06a] als *moderate Mehrdeutigkeit* bezeichnet wird. Für m-imprimitive Pattern gilt der geschilderte Umstand im übrigen nicht, weswegen eine zweite Charakterisierung der m-primitiven Pattern folgendermaßen formuliert werden kann:

Satz 5. *Sei $\alpha \in \mathbb{N}^+$. Der Homomorphismus $\sigma_{3\text{-seg}} : \mathbb{N}^* \to \{\mathtt{a}, \mathtt{b}\}^*$ ist genau dann moderat mehrdeutig für α, wenn α m-primitiv ist.*

Durch überaus komplizierte Beispiele kann außerdem belegt werden, daß ein Homomorphismus, der jede Variable auf *zwei* eindeutige Segmente abbildet, diese Eigenschaft nicht

mehr aufweist. Die Untersuchung von $\sigma_{3\text{-seg}}$ liefert also nicht nur einen verheißungsvollen Ansatz, neben der leicht zu definierenden Eindeutigkeit auch die Mehrdeutigkeit von Homomorphismen begrifflich zu fassen, sondern sie impliziert zusätzlich auch einen deutlichen Hinweis darauf, daß $\sigma_{3\text{-seg}}$ und $\sigma_{\mathrm{un},\alpha}$ im allgemeinen „optimal"gewählt sein könnten.

3.1 Applikation: Induktive Inferenz von terminalfreien E-Patternsprachen

Die Mehrdeutigkeit von *allgemeinen* Homomorphismen ist ein Phänomen, das naturgemäß im Rahmen der Klasse $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$ aller *terminalfreien E*-Patternsprachen seine Wirkung entfaltet. Der Zusammenhang zwischen der Existenz eindeutiger bzw. moderat mehrdeutiger Homomorphismen und der *Lernbarkeit* von Patternsprachen ergibt sich aus dem folgenden Charakterisierungssatz für die Telltales der terminalfreien E-Patternsprachen:

Satz 6. *Sei Σ ein Alphabet, $|\Sigma| \geq 2$, und sei $\alpha \in \mathbb{N}^+$ ein m-primitives Pattern. Sei außerdem $T_\alpha := \{w_1, w_2, \ldots, w_n\} \subseteq L_\Sigma(\alpha)$. T_α ist genau dann ein Telltale für $L_\Sigma(\alpha)$ in bezug auf $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$, wenn es zu jedem $x \in \mathrm{var}(\alpha)$ ein Wort $w \in T_\alpha$ gibt, so daß **für jeden Homomorphismus** $\sigma : \mathbb{N}^* \to \Sigma^*$ mit $\sigma(\alpha) = w$ ein Buchstabe $\mathtt{A} \in \Sigma$ existiert, der $|\sigma(x)|_{\mathtt{A}} = 1$ und $|\sigma(y)|_{\mathtt{A}} = 0$, $y \in \mathrm{var}(\alpha) \setminus \{x\}$, erfüllt.*

Es ist also *unabdingbare Voraussetzung* für die Wörter im Telltale einer terminalfreien E-Patternsprache, daß *alle* ihre erzeugenden Homomorphismen gewissen Anforderungen genügen. Das algorithmische Problem der Lernbarkeit von $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$ kann deshalb – und weil zu jedem m-imprimitiven Pattern ein m-primitives existiert, das dieselbe Sprache erzeugt – äquivalent als ein wortkombinatorisches Problem zur Existenz von Homomorphismen mit eingeschränkter Mehrdeutigkeit aufgefaßt werden. Unter Verweis auf die in Kapitel 2 beschriebene intuitive Interpretation von Telltales läßt sich außerdem aus Satz 6 die unerwartete Einsicht ablesen, daß im Kontext von Patternsprachen die eindeutigen oder zumindest „halbwegs" eindeutigen Homomorphismen ein stärker strukturerhaltendes Potential in sich tragen als die injektiven.

Es kann nun gezeigt werden, daß es für manche m-primitiven Pattern bei *binärem* Alphabet Σ keine Homomorphismen gibt, welche den sich aus Satz 6 ergebenden Anforderungen genügen. Für *dreielementige* Alphabete existieren hingegen solche Homomorphismen sehr wohl, da sie nämlich aufgrund der moderaten Mehrdeutigkeit von $\sigma_{3\text{-seg}}$ aus diesem unter Hinzunahme eines dritten Buchstaben konstruiert werden können. Es gilt deshalb das folgende, kuriose Resultat:

Satz 7. *Sei Σ ein Alphabet. $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$ ist genau dann im Limes anhand von positiven Beispielen erlernbar, wenn $|\Sigma| \neq 2$.*

Satz 7 liefert somit nicht nur einen durch wortkombinatorische Mittel gewonnenen, ganz erheblichen Wissenszuwachs zu einem der bekanntesten offenen Probleme der Induktiven Inferenz, sondern steht auch in einem reizvollen und verblüffenden Kontrast zur Kodierungstheorie, wo nämlich Wörter über einem zweielementigen Alphabet grundsätzlich dieselbe „Ausdruckskraft" haben wie solche über einem dreielementigen Alphabet.

4 Terminalerhaltende Homomorphismen

Aus den Erläuterungen in Kapitel 1 läßt sich die begründete Vermutung ableiten, daß die Mehrdeutigkeit von Substitutionen in bezug auf Frage 2 andere Eigenschaften aufweisen könnte als die von allgemeinen Homomorphismen hinsichtlich Frage 1. Diese Vermutung kann zumindest für „kleine“ Alphabete Σ mit höchstens vier Buchstaben bestätigt werden. Erweitert man nämlich die Definition der m-primitiven Pattern (s. Definition 1) kanonisch auf $(\mathbb{N} \cup \Sigma)^+$, indem man terminalerhaltende Homomorphismen $\phi, \psi : (\mathbb{N} \cup \Sigma)^* \to (\mathbb{N} \cup \Sigma)^*$ verwendet, so gilt der folgende Sachverhalt:

Satz 8. *Sei Σ ein Alphabet, $2 \leq |\Sigma| \leq 4$. Dann gibt es ein m-primitives $\alpha \in (\mathbb{N} \cup \Sigma)^+$, so daß jede nichtlöschende Substitution $\sigma : (\mathbb{N} \cup \Sigma)^* \to \Sigma^*$ mehrdeutig für α ist.*

Zum Beweis dieses Satzes kann für $|\Sigma| = 2$ auf $\alpha_{\mathtt{ab}} := 1 \cdot \mathtt{a} \cdot 2 \cdot \mathtt{b} \cdot 3$ verwiesen werden. Im Fall $|\Sigma| = 3$ bzw. $|\Sigma| = 4$ sind erheblich komplexere Pattern vonnöten, wie z. B.

$$\begin{aligned}
\alpha_{\mathtt{abc}} &:= 1 \cdot \mathtt{a} \cdot 2 \cdot 3^2 \cdot 4^2 \cdot 5^2 \cdot 6^2 \cdot 7 \cdot \mathtt{b} \cdot 8 \cdot \mathtt{a} \cdot 2 \cdot 9^2 \cdot 4^2 \cdot 5^2 \cdot 10^2 \cdot 7 \cdot \mathtt{b} \cdot 11\,, \\
\alpha_{\mathtt{abcd}} &:= 1 \cdot \mathtt{a} \cdot 2 \cdot 3^2 \cdot 4^2 \cdot 5^2 \cdot 6^2 \cdot 7 \cdot \mathtt{b} \cdot 8 \cdot \mathtt{a} \cdot 2 \cdot 9^2 \cdot 4^2 \cdot 5^2 \cdot 10^2 \cdot 7 \cdot \mathtt{b} \cdot 11 \cdot \mathtt{c} \\
&\quad \cdot 12 \cdot 13^2 \cdot 14^2 \cdot 15^2 \cdot 16^2 \cdot 17 \cdot \mathtt{d} \cdot 18 \cdot \mathtt{c} \cdot 12 \cdot 19^2 \cdot 14^2 \cdot 15^2 \cdot 20^2 \cdot 17 \\
&\quad \cdot \mathtt{d} \cdot 21 \cdot 14^2 \cdot 15^2 \cdot 14^2 \cdot 15^2 \cdot 14^2 \cdot 15^2 \cdot 22 \cdot 4^2 \cdot 5^2 \cdot 4^2 \cdot 5^2 \cdot 4^2 \cdot 5^2
\end{aligned}$$

(wobei sich die Exponenten hier stets auf die Konkatenation beziehen, es steht also beispielsweise 3^2 für das Teilpattern $3 \cdot 3$). Da sich außerdem Satz 2 auf Substitutionen verallgemeinern läßt, muß folglich die Menge der eindeutig abbildbaren Pattern in $(\mathbb{N} \cup \Sigma)^+$ eine echte Teilmenge der m-primitiven sein.

Die Ursache für das in Satz 8 beschriebene Phänomen und die einhergehenden Unterschiede zu Satz 4 stellt ein Typus von Mehrdeutigkeit dar, welcher bei terminalfreien Pattern per definitionem nicht auftreten kann: die *terminalumfassende* Mehrdeutigkeit einer Substitution σ für ein Pattern α, d. h. die Existenz einer Substitution τ mit $\tau(\alpha) = \sigma(\alpha) =: w$, so daß ein Vorkommen eines Buchstaben $\mathtt{A}$ in w bei einer der beiden Substitutionen durch das Bild einer in α auftretenden Variablen entsteht und bei der anderen das Bild eines Vorkommens des Terminalsymbols $\mathtt{A}$ in α ist. Illustriert wird dies durch das Beispielpattern α_2' und die zugehörigen Substitutionen σ und τ in Kapitel 1, da der letzte Buchstabe des Wortes $\sigma(\alpha_2')[= \tau(\alpha_2')]$ einerseits aus $\sigma(2)$ und andererseits aus $\tau(\mathtt{a})$ hervorgeht. Das Hauptproblem bei der Übertragung von Satz 4 ist deshalb die Frage nach der *Vermeidbarkeit* von terminalumfassender Mehrdeutigkeit, d. h. die Existenz von Substitutionen, die aus $\sigma_{\mathrm{un},\alpha}$ abgeleitet werden können und nicht terminalumfassend mehrdeutig sind. Satz 8 besagt also, daß zu Alphabeten mit höchstens vier Buchstaben gewisse Pattern existieren, für welche die terminalumfassende Mehrdeutigkeit von Substitutionen unvermeidbar ist. In Hinblick auf größere Alphabet ist unbekannt, ob solche Beispiele ebenfalls existieren:

Problem 1. *Sei Σ ein Alphabet, $|\Sigma| \geq 5$. Ist die terminalumfassende Mehrdeutigkeit für alle m-primitiven Pattern über Σ vermeidbar?*

Die beträchtliche Komplexität der obigen Beispielpattern $\alpha_{\mathtt{abc}}$ und $\alpha_{\mathtt{abcd}}$ legt die Vermutung nahe, daß Problem 1 einen außerordentlich hohen Schwierigkeitsgrad hat.

4.1 Applikation: Eigenschaften von generellen E-Patternsprachen

Trotz der bislang spärlichen Einsichten in die terminalumfassende Mehrdeutigkeit von Substitutionen lassen sich aus den oben definierten Pattern und ähnlichen Beispielen bereits diverse nichttriviale Rückschlüsse auf vorher unverstandene Eigenschaften von E-Patternsprachen ziehen. Insbesondere können auf Grundlage von $\alpha_{\mathtt{abc}}$, $\alpha_{\mathtt{abcd}}$ und ihrer in Satz 8 beschriebenen Eigenschaft geeignete Beispielpattern konstruiert werden, welche für Alphabetgrößen 3 und 4 die Hauptvermutung aus [OU97] zum Äquivalenzproblem für E-Patternsprachen widerlegen. Darüber hinaus gilt, daß aufgrund der terminalumfassenden Mehrdeutigkeit spezieller Substitutionen für ähnliche Pattern das in Satz 7 enthaltene positive Lernbarkeitsresultat für $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$ im Falle von $|\Sigma| \in \{3,4\}$ nicht auf die Klasse aller E-Patternsprachen übertragen werden kann:

Satz 9. *Sei Σ ein Alphabet, $|\Sigma| \in \{3,4\}$. Dann ist ePAT_{Σ} nicht im Limes anhand von positiven Beispielen erlernbar.*

Es existieren deutliche Hinweise darauf, daß jeder substantielle Fortschritt zu Problem 1 etliche, vergleichbar tiefe Einsichten zu bislang ungeklärten Eigenschaften von E-Patternsprachen erlaubt. Problem 1 ist also nicht nur in wortkombinatorischer, sondern auch in sprach- und lerntheoretischer Hinsicht von herausragender Bedeutung.

Literatur

[Ang80a] D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62, 1980.

[Ang80b] D. Angluin. Inductive Inference of Formal Languages from Positive Data. *Information and Control*, 45:117–135, 1980.

[AS83] D. Angluin und C. Smith. Inductive Inference: Theory and Methods. *Computing Surveys*, 15:237–269, 1983.

[AS03] J.-P. Allouche und J. Shallit. *Automatic Sequences*. Cambridge University Press, Cambridge, New York, 2003.

[BP85] J. Berstel und D. Perrin. *Theory of Codes*. Academic Press, Orlando, 1985.

[FRS06] D.D. Freydenberger, D. Reidenbach und J.C. Schneider. Unambiguous Morphic Images of Strings. *International Journal of Foundations of Computer Science*, 17:601–628, 2006.

[Gol67] E.M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[HK97] T. Harju und J. Karhumäki. Morphisms. In [RS97], Kapitel 7, Seiten 439–510. 1997.

[KRS97] L. Kari, G. Rozenberg und A. Salomaa. L Systems. In [RS97], Kapitel 5, Seiten 253–328. 1997.

[Lot02] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge, New York, 2002.

[LW91] S. Lange und R. Wiehagen. Polynomial-time inference of arbitrary pattern languages. *New Generation Computing*, 8:361–370, 1991.

[Mit98] A.R. Mitchell. Learnability of a subclass of extended pattern languages. In *Proc. 11th Annual Conference on Computational Learning Theory, COLT 1998*, Seiten 64–71, 1998.

[MS97] A. Mateescu und A. Salomaa. Patterns. In [RS97], Kapitel 4.6, Seiten 230–242. 1997.

[OU97] E. Ohlebusch und E. Ukkonen. On the equivalence problem for E-pattern languages. *Theoretical Computer Science*, 186:231–248, 1997.

[Reia] D. Reidenbach. Discontinuities in pattern inference. *Theoretical Computer Science*. Zur Publikation angenommen.

[Reib] D. Reidenbach. An examination of Ohlebusch and Ukkonen's Conjecture on the equivalence problem for E-pattern languages. *Journal of Automata, Languages and Combinatorics*. Zur Publikation angenommen.

[Rei06a] D. Reidenbach. *The Ambiguity of Morphisms in Free Monoids and its Impact on Algorithmic Properties of Pattern Languages*. Dissertation, Fachbereich Informatik, Technische Universität Kaiserslautern, 2006. Logos Verlag, Berlin.

[Rei06b] D. Reidenbach. A non-learnable class of E-pattern languages. *Theoretical Computer Science*, 350:91–102, 2006.

[RS97] G. Rozenberg und A. Salomaa. *Handbook of Formal Languages, Bd. 1*. Springer, Berlin, 1997.

[RZ00] R. Reischuk und T. Zeugmann. An Average-Case Optimal One-Variable Pattern Language Learner. *Journal of Computer and System Sciences*, 60:302–335, 2000.

[RZ01] P. Rossmanith und T. Zeugmann. Stochastic Finite Learning of the Pattern Languages. *Machine Learning*, 44:67–91, 2001.

[Shi82] T. Shinohara. Polynomial Time Inference of Extended Regular Pattern Languages. In *Proc. RIMS Symposia on Software Science and Engineering, Kyoto*, Lecture Notes in Computer Science, Bd. 147, Seiten 115–127, 1982.

[WZ94] R. Wiehagen und T. Zeugmann. Ignoring data may be the only way to learn efficiently. *Journal of Experimental and Theoretical Artificial Intelligence*, 6:131–144, 1994.

Daniel Reidenbach wurde am 7. November 1973 in Trier geboren. Er besuchte dort das Friedrich-Wilhelm-Gymnasium, welches er 1993 mit dem Abitur abschloß. Er beendete sein Studium an der Universität Kaiserslautern im Jahre 2003 als Diplom-Informatiker, und arbeitet seitdem an eben dieser (mittlerweile in Technische Universität Kaiserslautern umbenannten) Hochschule als Wissenschaftlicher Mitarbeiter in der Arbeitsgruppe „Algorithmisches Lernen“ von Prof. Dr. R. Wiehagen. Er promovierte am Fachbereich Informatik der TU Kaiserslautern im Dezember 2006. Seine Arbeiten zur Lern- und Sprachtheorie sind mit dem *E. Mark Gold Award 2002*, dem *Mark Fulk Award 2004* und dem *Best Student Paper Award* der *DLT 2004* ausgezeichnet worden.

Daniel Reidenbach ist verheiratet und Vater zweier Kinder.