

Teil VI

Grammatiken

29. Knobelaufgabe #15

Ist es möglich, ein *nicht-rekursives* Programm zu schreiben, das *nicht terminiert*?

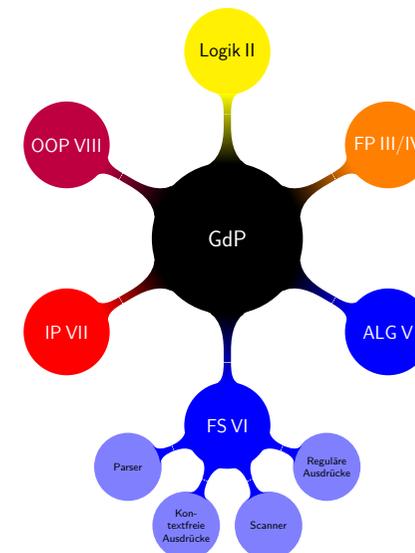
Nicht-rekursiv bedeutet, dass in dem Programm keine rekursiven Funktionsdefinitionen

```
let rec f (x1 : t1) : t2 = ... f ...
```

verwendet werden dürfen.

29. Gliederung

- 30 Reguläre Ausdrücke
- 31 Scanner
- 32 Kontextfreie Ausdrücke
- 33 Parser*



29. Lernziele

Nach Durcharbeitung dieses Kapitels sollten Sie

- ▶ den Unterschied zwischen konkreter und abstrakter Syntax verstanden haben,
- ▶ Syntax und Semantik regulärer Ausdrücke kennen,
- ▶ aus einem regulären Ausdruck einen Akzeptor konstruieren können,
- ▶ den Unterschied zwischen Interpreter und Übersetzer verstanden haben,
- ▶ Syntax und Semantik kontextfreier Ausdrücke kennen,
- ▶ (aus einem kontextfreien Ausdruck einen Parser konstruieren können,)
- ▶ Möglichkeiten und Grenzen der Sprachfamilien kennen.

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke

Scanner

Kontextfreie
Ausdrücke

Parser »

546

29. Überblick

Dieses Kapitel erzählt eine der großen Erfolgsgeschichten der Informatik: die automatische Umsetzung einer deskriptiven Beschreibung („was“) in ein ausführbares Programm („wie“).

Wir führen Formalismen ein, mit denen man die konkrete Syntax einer Programmiersprache beschreiben kann:

- ▶ die lexikalische Syntax und
- ▶ die kontextfreie Syntax.

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke

Scanner

Kontextfreie
Ausdrücke

Parser »

547

29. Wiederholung: Lexikalische Syntax

Betrachten wir ein einfaches Beispielprogramm:

```
4711* (a11 (* speed *) + 815 )
```

Mikroskopisch gesehen besteht das Programm aus einer Folge von Zeichen:

- ▶ der Ziffer 4,
- ▶ gefolgt von der Ziffer 7,
- ▶ gefolgt von der Ziffer 1,
- ▶ gefolgt von der Ziffer 1,
- ▶ gefolgt von einem Asteriskus *,
- ▶ gefolgt von einem Leerzeichen usw.

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke

Scanner

Kontextfreie
Ausdrücke

Parser »

548

29. Wiederholung: Lexeme

Als menschlicher Leser sind wir gewohnt — bzw. durch jahrelanges Training geschult — mehrere Zeichen zu einer Einheit zusammenzufassen.

```
4711 * ( a11 + 815 )
```

☞ Nicht alle Zeichen sind für den Rechner gedacht: (* speed *) ist ein Kommentar, der sich an den menschlichen Leser richtet.

In der *lexikalischen Syntax* einer Programmiersprache wird festgelegt, wie Zeichen zu größeren Einheiten, sogenannten *Lexemen* (engl. tokens), zusammengefasst werden.

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke

Scanner

Kontextfreie
Ausdrücke

Parser »

549

29. Wiederholung: Kontextfreie Syntax

Nicht alle Folgen von Lexemen stellen ein gültiges Programm dar:

```
) * 4711 815 + ( a11
```

umfasst die gleichen Lexeme, ist aber *kein* Mini-F# Programm.

In der *kontextfreien Syntax* einer Programmiersprache wird festgelegt, welche Folgen von Lexemen gültige Programme sind und welche nicht.

Die lexikalische und die kontextfreie Syntax bilden zusammen die *konkrete Syntax* einer Programmiersprache.

30. Lexeme in Mini-F#

- ▶ *Numerale*: ein Numeral in Mini-F# besteht aus einer nicht-leeren Folge von Dezimalziffern.
- ▶ *Bezeichner*: ein Bezeichner in Mini-F# beginnt mit einem Buchstaben, gefolgt von weiteren Buchstaben, Ziffern und Sonderzeichen, wie einem Unterstrich oder einem Apostroph.

🗨️ Wie können wir den Aufbau von Numeralen und Bezeichnern präzise beschreiben?

🗨️ Wir müssen uns eine Sprache ausdenken — eine Sprache, um Sprachen zu beschreiben.

30. Begriffe

- ▶ Ein *Alphabet* A ist eine Menge von Zeichen.

Beispiele:

- ▶ $\{a, b\}$,
- ▶ $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, (,), *, +\}$,
- ▶ ASCII-Alphabet,
- ▶ Menge aller Mini-F#-Lexeme.
- ▶ Eine *Sprache* ist eine Teilmenge von A^* , der Menge der Sequenzen über A .
- ▶ Statt von Sequenzen spricht man auch von *Wörtern*.
- ▶ Die Menge aller Sprachen ist $\mathbb{P}(A^*)$.

30. Motivation

Eine Sprache ist eine Menge von Wörtern; auf Wörtern haben wir bereits einige Operationen eingeführt.

- ▶ das *leere Wort*: ϵ ,
- ▶ die *Konkatenation* von Wörtern: $w_1 w_2$,
- ▶ die *n-fache Wiederholung*: w^n .

🗨️ Verallgemeinern wir w^n zu einer beliebigen Wiederholung w^* , so können wir Numerale beschreiben:

```
digit digit*
```

🗨️ *digit* steht für die Sprache der Ziffern. *Lies*: ein Numeral ist eine Ziffer gefolgt von einer beliebigen Folge von Ziffern.

30. Motivation

Wie können wir *digit* beschreiben? Eine Ziffer ist entweder 0, oder 1, oder ... Erfinden wir eine Notation für die Alternative, '|', dann ist

0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

die gesuchte Definition von *digit*.

Die lexikalische Syntax von Bezeichnern lässt sich auf ähnliche Art und Weise festlegen:

letter (*letter* | *digit* | _ | `)*

wobei *letter* durch a | ... | z | A | ... | Z gegeben ist.

30. Motivation

Jetzt können wir unser Routineprogramm abspulen!

Wir definieren die

- ▶ die abstrakte Syntax der Sprachbeschreibungssprache und
- ▶ die Semantik der Sprachbeschreibungssprache.



Mir schwindt der Kopf, eine Sprachbeschreibungssprache! Drehen wir uns da nicht im Kreis? Ich meine, wir wollen doch die konkrete Syntax von Mini-F# festlegen und jetzt müssen wir zuerst die Syntax der Sprachbeschreibungssprache festlegen.

Aber, welche Alternativen gibt es? Willst Du die konkrete Syntax umgangssprachlich festlegen?



Ist mir schon bewusst, dass das keine gute Idee ist.

Wenn Du einen Formalismus verwendest, muss doch klar sein, was erlaubt ist und was gemeint ist.



Die Gründe liegen aber noch tiefer: Im Endeffekt wollen wir die Feststellung, ob ein Stück Text ein gültiges Mini-F# Programm ist, dem Rechner übertragen. Dazu muss eben Syntax und Semantik präzise festgelegt werden.

Mir schwant Böses: ein Mini-F# Programm, das überprüft, ob Mini-F# Programme korrekt sind ...



30. Abstrakte Syntax

$a \in A$

$r \in \text{Reg} ::=$

- | ϵ
- | a
- | $r_1 r_2$
- | \emptyset
- | $r_1 | r_2$
- | r^*

Alphabet

reguläre Ausdrücke:

- das leere Wort
- einzelnes Zeichen \ Terminalsymbol
- Konkatenation \ Sequenz
- die leere Sprache
- Alternative
- Wiederholung

☞ Die Symbole ϵ und \emptyset sind *überladen*: je nach Kontext bedeuten sie etwas Verschiedenes — ϵ steht auch für das leere Wort und \emptyset für die leere Menge.

30. Konkrete Syntax

Zwei konkrete Beispiele für konkrete Syntaxen:

	Abstrakte Syntax	grep	egrep
das leere Wort	ϵ	-	-
einzelnes Zeichen	a	a	a
Konkatenation	$r_1 r_2$	$r_1 r_2$	$r_1 r_2$
die leere Sprache	\emptyset	-	-
Alternative	$r_1 \mid r_2$	$r_1 \setminus \mid r_2$	$r_1 \mid r_2$
Wiederholung	r^*	$r \setminus *$	r^*
Gruppierung	-	$\setminus (r \setminus)$	(r)
beliebiges Zeichen	$a_1 \mid \dots \mid a_n$.	.
optionales Vorkommen	$\epsilon \mid r$	$r \setminus ?$	$r ?$
positive Wiederholung	$r r^*$	$r \setminus +$	$r +$

☞ `grep` und `egrep` durchsuchen Sequenzen (typischerweise Textdateien) nach Teilsequenzen, die auf ein bestimmtes Muster passen.

☞ `\` ist ein Fluchtsymbol.

30. Semantik

☞ Eine Aufteilung in statische und dynamische Semantik ist nicht notwendig: reguläre Ausdrücke sind beliebig kombinierbar.

Wir definieren zwei verschiedene Semantiken:

- ▶ *Denotationelle Semantik*: „Welche Sprache bezeichnet der reguläre Ausdruck?“
- ▶ *Reduktionsemantik*: „Wie kann ich ein einzelnes Wort aus dem regulären Ausdruck ableiten?“

30. Denotationelle Semantik

- ▶ Die Bedeutung von ϵ ist $\{\epsilon\}$.
- ▶ Die Bedeutung von a ist $\{a\}$, wobei a das Wort $\{0 \mapsto a\}$ abkürzt.
- ▶ *Zur Erinnerung*: Die Bedeutung eines regulären Ausdrucks ist eine Sprache, kein Wort.
- ▶ Wir haben die Konkatenation von Wörtern definiert; jetzt müssen wir die Konkatenation auf Mengen von Wörtern fortsetzen.

$$L_1 \cdot L_2 = \{w_1 \cdot w_2 \mid w_1 \in L_1, w_2 \in L_2\}$$

☞ Damit ist auch ' \cdot ' überladen.

- ▶ Beliebige Wiederholung einer Sprache:

$$L^* = \bigcup \{L^n \mid n \in \mathbb{N}\}$$

mit $L^0 = \{\epsilon\}$ und $L^{n+1} = L \cdot L^n$.

30. Denotationelle Semantik

Die Semantik wird durch eine Bedeutungsfunktion angegeben:

$$\begin{aligned} \llbracket a \rrbracket &= \{a\} \\ \llbracket \epsilon \rrbracket &= \{\epsilon\} \\ \llbracket r_1 r_2 \rrbracket &= \llbracket r_1 \rrbracket \cdot \llbracket r_2 \rrbracket \\ \llbracket \emptyset \rrbracket &= \emptyset \\ \llbracket r_1 \mid r_2 \rrbracket &= \llbracket r_1 \rrbracket \cup \llbracket r_2 \rrbracket \\ \llbracket r^* \rrbracket &= \llbracket r \rrbracket^* \end{aligned}$$

☞ Die doppelten Klammern sind die sogenannten *Oxford* oder *Strachey Klammern* — sie umschließen die Syntax.

30. Beispiele

Zugrundeliegendes Alphabet $A = \{a, b\}$.

$$\begin{aligned} & \llbracket a \mid b \mid a \rrbracket \\ &= \llbracket a \mid b \rrbracket \cup \llbracket b \mid a \rrbracket \\ &= (\llbracket a \rrbracket \cdot \llbracket b \rrbracket) \cup (\llbracket b \rrbracket \cdot \llbracket a \rrbracket) \\ &= (\{a\} \cdot \{b\}) \cup (\{b\} \cdot \{a\}) \\ &= \{ab\} \cup \{ba\} \\ &= \{ab, ba\} \end{aligned}$$

Der reguläre Ausdruck $\llbracket a \mid b \mid a \rrbracket$ bezeichnet die Sprache $\{ab, ba\}$.

30. Beispiele

Vertauschen wir Konkatenation und Alternative erhalten wir eine andere Sprache:

$$\begin{aligned} & \llbracket (a \mid b) (b \mid a) \rrbracket \\ &= \llbracket a \mid b \rrbracket \cdot \llbracket b \mid a \rrbracket \\ &= (\llbracket a \rrbracket \cup \llbracket b \rrbracket) \cdot (\llbracket b \rrbracket \cup \llbracket a \rrbracket) \\ &= (\{a\} \cup \{b\}) \cdot (\{b\} \cup \{a\}) \\ &= \{a, b\} \cdot \{b, a\} \\ &= \{ab, aa, bb, ba\} \end{aligned}$$

Der reguläre Ausdruck $(a \mid b) (b \mid a)$ bezeichnet die Sprache aller Wörter der Länge 2.

30. Beispiele

$$\begin{aligned} & \llbracket ((a \mid b) (b \mid a))^* \rrbracket \\ &= \llbracket (a \mid b) (b \mid a) \rrbracket^* \\ &= \{ab, aa, bb, ba\}^* \end{aligned}$$

Die bezeichnete Sprache umfasst alle Wörter gerader Länge.

30. Reduktionsemantik

Zur Erinnerung:

- *Denotationelle Semantik*: „Welche Sprache bezeichnet der reguläre Ausdruck?“
 - Der Globale Sicht.
- *Reduktionsemantik*: „Wie kann ich ein einzelnes Wort aus dem regulären Ausdruck ableiten?“
 - Der Lokale Sicht.

30. Reduktionssemantik

Um die Semantik von Mini-F# festzulegen, haben wir Werte eingeführt, das sind besonders einfache Ausdrücke.

In Analogie dazu: für die Reduktionssemantik sind Wörter besonders einfache reguläre Ausdrücke.

$w ::= \epsilon$	das leere Wort
a	einzelnes Zeichen \ Terminalsymbol
$w_1 w_2$	Konkatenation \ Sequenz

Ein Wort zeichnet sich dadurch aus, dass seine Denotation eine einelementige Menge ist.

30. Reduktionssemantik

Die schrittweise Reduktion $r \rightarrow r'$ (lies: r kann in einem Schritt zu r' reduziert werden) wird durch ein Beweissystem formalisiert.

Rechenregeln:

$$\frac{}{r \epsilon \rightarrow r} \quad \frac{}{\epsilon r \rightarrow r}$$

$$\frac{}{r_1 \mid r_2 \rightarrow r_1} \quad \frac{}{r_1 \mid r_2 \rightarrow r_2}$$

$$\frac{}{r^* \rightarrow \epsilon} \quad \frac{}{r^* \rightarrow r r^*}$$

Die letzten vier Regeln formalisieren Wahlmöglichkeiten.

30. Beispiele

Für den regulären Ausdruck $a \mid b \mid a$ existieren zwei mögliche Reduktionsfolgen.

$$a \mid b \mid a \rightarrow a b$$

$$a \mid b \mid a \rightarrow b a$$

Die Sprache, die durch einen regulären Ausdruck beschrieben wird, ist die Menge aller *Worte*, die aus dem Ausdruck ableitbar sind. Für unser Beispiel: $\{ab, ba\}$.

30. Reduktionssemantik

Um den Ausdruck $(a \mid b) (b \mid a)$ zu vereinfachen, benötigen wir weitere Regeln. Die obigen Beweisregeln erlauben nur den *gesamten* Ausdruck umzuformen; wir brauchen also Regeln, die es uns ermöglichen, einen Ausdruck „mittendrin“ zu manipulieren.

Kongruenz- oder Kontextregeln:

$$\frac{r_1 \rightarrow r'_1}{r_1 r_2 \rightarrow r'_1 r_2} \quad \frac{r_2 \rightarrow r'_2}{r_1 r_2 \rightarrow r_1 r'_2}$$

30. Beispiele

Für $(a \mid b)(b \mid a)$ gibt es insgesamt vier mögliche Reduktionsfolgen.

$(a \mid b)(b \mid a) \rightarrow a(b \mid a) \rightarrow ab$
 $(a \mid b)(b \mid a) \rightarrow a(b \mid a) \rightarrow aa$
 $(a \mid b)(b \mid a) \rightarrow b(b \mid a) \rightarrow bb$
 $(a \mid b)(b \mid a) \rightarrow b(b \mid a) \rightarrow ba$

☞ Die bezeichnete Sprache ist somit $\{ab, aa, bb, ba\}$.

30. Beispiele

Wesentlich mehr Möglichkeiten gibt es für $((a \mid b)(b \mid a))^*$, nämlich unendlich viele.

$((a \mid b)(b \mid a))^* \rightarrow \epsilon$
 $((a \mid b)(b \mid a))^* \rightarrow (a \mid b)(b \mid a)((a \mid b)(b \mid a))^*$
 $\rightarrow a(b \mid a)((a \mid b)(b \mid a))^*$
 $\rightarrow a b ((a \mid b)(b \mid a))^*$
 $\rightarrow a b \epsilon$
 $\rightarrow a b$

...

30. Beispiele

Zur Erinnerung: Die Sprache, die durch einen regulären Ausdruck beschrieben wird, ist die Menge aller *Worte*, die aus dem Ausdruck ableitbar sind.

$\emptyset^* \rightarrow \emptyset \emptyset^* \rightarrow \emptyset \emptyset \emptyset^* \rightarrow \dots$

☞ Da es keine Regel gibt, \emptyset zu reduzieren, lässt sich mit dieser Reduktionsfolge kein Wort ableiten.

Die einzige mögliche Reduktionsfolge ist

$\emptyset^* \rightarrow \epsilon$

☞ Somit ist die Semantik von \emptyset^* die Sprache $\{\epsilon\}$. (Auch mit der denotationellen Semantik erhalten wir $\llbracket \emptyset^* \rrbracket = \{\epsilon\}$.)

30. Vertiefung

Eine Sprache kann in der Regel auf verschiedene Art und Weisen beschrieben werden.

Beispiel: Sprache aller Wörter mit einer geraden Anzahl von as und beliebig vielen bs.

$(a b^* a \mid b)^*$

Alternative Formulierung:

$(b \mid a b^* a)^*$

Alternative Formulierung ohne '|':

$b^* (a b^* a b^*)^*$

☞ Allgemein gilt: $(r_1 \mid r_2)^* = r_1^* (r_2 r_1^*)^*$.

30. Gleichheit regulärer Ausdrücke

☞ Wie zeigt man, dass zwei reguläre Ausdrücke r_1 und r_2 gleichwertig sind?

Mit Hilfe der Semantik:

- ▶ *Denotationelle Semantik*: $\llbracket r_1 \rrbracket = \llbracket r_2 \rrbracket$;
- ▶ also: Rückgriff auf die (naive) Mengenlehre;
- ▶ *Reduktionssemantik*: zu jeder Reduktionsfolge $r_1 \rightarrow \dots \rightarrow w$ gibt es eine korrespondierende Reduktionsfolge $r_2 \rightarrow \dots \rightarrow w$;
- ▶ also: Rückgriff auf Beweissysteme (Logik).

30. Vertiefung

Die Sprache aller Wörter, die eine gerade Anzahl von as (bs beliebig) *oder* eine ungerade Anzahl von bs (as beliebig) haben, lässt sich unter Rückgriff auf das vorherige Beispiel (Folie 573) leicht definieren.

$(a b^* a \mid b)^* \mid a^* b (a \mid b a^* b)^*$

☞ Aber wie beschreibt man die Sprache aller Wörter, die eine gerade Anzahl von as (bs beliebig) *und* eine ungerade Anzahl von bs (as beliebig) haben?

30. Vertiefung

Stände ein Pendant zum Mengendurchschnitt bereit, sagen wir $\&$, könnten wir formulieren:

$(a b^* a \mid b)^* \& a^* b (a \mid b a^* b)^*$

☞ Es spricht nichts dagegen, den Durchschnitt zu der Sprache der regulären Ausdrücke hinzuzunehmen.

☞ Interessanterweise ist die Erweiterung zwar bequem, aber nicht notwendig; sie erhöht nicht die Ausdruckskraft.

Versuchen wir also die Sprache mit den bisherigen Bordmitteln zu definieren ...

30. Vertiefung

Nützliche Abkürzungen: $g = (aa)^*$ und $u = a(aa)^*$.

1. Schritt: genau ein b und eine gerade Anzahl von as.

$g b g \mid u b u$

☞ Entweder kommen vor und hinter dem b eine gerade Anzahl von as vor oder an beiden Positionen eine ungerade Anzahl.

2. Schritt: wir erhöhen die Anzahl der bs auf zwei:

$g b g b g \mid g b u b u \mid u b g b u \mid u b u b g$

Jetzt werden die as über drei Positionen verteilt; entweder befinden sich an allen Stellen eine gerade Anzahl von as oder an exakt zwei Positionen eine ungerade Anzahl.

30. Vertiefung

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke
Motivation
Abstrakte Syntax
Denotationelle
Semantik
Reduktionsemantik
Vertiefung
Scanner
Kontextfreie
Ausdrücke
Parser*

3. Schritt: eine ungerade Zahl hat die Form $1 + 2n$, entsprechend definieren wir

$(g\ b\ g\ | \ u\ b\ u)\ (g\ b\ g\ b\ g\ | \ g\ b\ u\ b\ u\ | \ u\ b\ g\ b\ u\ | \ u\ b\ u\ b\ g)^*$

☞ Man sieht, ohne den Durchschnitt müssen wir bei der Formulierung von Sprachen sehr viel mehr Grips investieren.

578

30. Grenzen regulärer Ausdrücke

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke
Motivation
Abstrakte Syntax
Denotationelle
Semantik
Reduktionsemantik
Vertiefung
Scanner
Kontextfreie
Ausdrücke
Parser*

- ▶ Können wir mit regulären Ausdrücken alle Sprachen beschreiben?
- ▶ Nein, es gibt sehr viel mehr Sprachen als Ausdrücke, mit denen Sprachen beschrieben werden können (überabzählbar versus abzählbar).
- ▶ Können wir mit regulären Ausdrücken die Syntax von Programmiersprachen festlegen?
- ▶ Nein, einfache Hygienevorschriften lassen sich nicht fassen: etwa, dass es in Mini-F# Ausdrücken zu jeder „Klammer auf“ eine korrespondierende „Klammer zu“ geben muss. Die Sprache

$\{a^n b^n \mid n \in \mathbb{N}\}$

ist nicht regulär — lies a als (und b als) .

579

30. Lösung Knobelaufgabe #4

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke
Motivation
Abstrakte Syntax
Denotationelle
Semantik
Reduktionsemantik
Vertiefung
Scanner
Kontextfreie
Ausdrücke
Parser*



Endlich!!

Also: das sichselbstausgebende Programm muss mindestens einen String enthalten; ein String alleine reicht aber nicht, es muss noch etwas dazukommen. *Idee*: der String enthält den Programmcode des anderen Teils; damit das Programm sich selbst ausgibt, muss der String verdoppelt werden. Etwa so:

$s \hat{=} show\ s$



Warum einmal s und einmal $show\ s$?

s steht für den Programmcode und $show\ s$ für den Text des Programmcodes.

$(fun\ s \rightarrow s \hat{=} show\ s)\ (fun\ s \rightarrow s \hat{=} show\ s)''$



580

30. Lösung Knobelaufgabe #4 — Demo

VI Grammatiken

Ralf Hinze

Reguläre
Ausdrücke
Motivation
Abstrakte Syntax
Denotationelle
Semantik
Reduktionsemantik
Vertiefung
Scanner
Kontextfreie
Ausdrücke
Parser*

```
Mini> "Hello, world!"  
"Hello, world!"  
Mini> show "Hello, world!"  
"\\"Hello, world!\""  
Mini> length "Hello, world!"  
13  
Mini> length (show "Hello, world!")  
15  
Mini> (fun s -> s ^ show s) "(fun s -> s ^ show s)''"  
"(fun s -> s ^ show s)\"(fun s -> s ^ show s)\""  
Mini> putline it  
(fun s -> s ^ show s) "(fun s -> s ^ show s)''"
```

581