Regular Expressions Example: You can follow this "naturally" without consciously applying the formal defn pp63-66 ①

I How to represent Natural Whole Numbers? $\mathbb{N}$

Alphabet $DIG = \{0, 1, \ldots, 9\}$

Idea 1: DIG^* As in lecture, the $*$ means "zero or more"

Problem: allows the empty string ε. "Invisible zero".

Idea 2: DIG^+ : Superscript $+$ means "one or more."

Second Issue: Allows redundant codes for integers with leading 0s:
eg. $007 = 7 = 000000007$.

Idea 3: Require the first digit to be 1 thru 9:

$$\underbrace{(1 \cup 2 \cup 3 \cup 4 \cup 5 \cup 6 \cup 7 \cup 8 \cup 9)}_{\text{UNIX: shorten to } [1-9]} \cdot DIG^*.$$

or

followed by

This simple example already uses all three regular operations $\cup, \cdot, *$.

Problem: 0 has disappeared.

Idea 4: Union it as a special case:

$$Natural = [1-9] \cdot DIG^* \cup 0.$$

II Now suppose we allow negative numbers. We want to use an (optional) $-$ sign.

"Notation for "Optional": "BNF Grammar" [—] ②

UNIX/Linux —?

Regexp/Text ("—" ∪ ε)

$(\text{'}-\text{'} \cup \varepsilon)([1-9]\,DIG^* \cup 0)$ allows "−0"

$(- \cup \varepsilon)[1-9]\,DIG^* \cup 0$ does not allow −0.

$(- \cup + \cup \varepsilon)[1-9]\,DIG^* \cup 0$ allows optional + sign too.

III: Now how about floating point numbers?

$\underbrace{INT \,.\, DIG^+}$ now leading 0s are needed in the floating part, but trailing 0s cause redundancy. OK.

Issue: can't write .5
must write 0.5. Bug or feature?

If you write $(- \cup + \cup \varepsilon)\,DIG^* . DIG^*$ you can do .5, but then also gets "5." or just "." Of course we can't allow the latter, so let's do

$INT . DIG^*$ or $(- \cup + \cup \varepsilon)\,DIG^* . DIG^+$

Formal defn of concatenation:

$$A \circ B = \{x \cdot y : x \in A \wedge y \in B\}$$

Contrast with Cartesian Product:

$$A \times B = \{(x,y) : x \in A \wedge y \in B\}$$

$A = \{a, ab\}$ $A \circ B = \{a \cdot b, a \cdot bb, ab \cdot b, ab \cdot bb\}$

$B = \{b, bb\}$ $= \{ab, abb, abbb\}$ size 3

$A \times B = \{(a,b), (a,bb), (ab,b), (ab,bb)\}$

always $|A \times B| = |A| \cdot |B|$. here, $= 4$

$$A \circ A = \{x \cdot y : x \in A \wedge y \in A\}$$

Note: $\{\varepsilon\} \cdot A = \{\varepsilon \cdot y : y \in A\} = \{y : y \in A\} = A$.

$\neq \{x \cdot x : x \in A\}$ which would be $\subseteq$ the Doubleword language (from lecture)

$A^1 = A$, $A^2 = A \circ A$, $A^3 = A \circ A \circ A \ldots$ (happily, $\circ$ is associative)

Rule of Exponents: $A^q \circ A^r = A^{q+r}$ Like $0^0 = 1$.

Needs $A^0 \equiv \{\varepsilon\}$ not $\emptyset$. Even $\emptyset^0 = \{\varepsilon\}$, not $\emptyset$!