## Jak opisać ładny język?

Przyjmijmy, że słowem będzie dla nas dowolny ciąg liter. Nie tylko taki, który da się wymówić albo którego znaczenie jest objaśnione w słowniku. Zupełnie dowolny ciąg, złożony z jednej litery, kilku lub nawet długości całej encyklopedii. Możemy wziąć także bardzo wyjątkowy ciąg o długości zero, nazywany słowem pustym. Niektórzy zajmują się też nieskończonymi słowami, ale to inna historia. Dla nas każde słowo ma długość będącą liczbą całkowitą nieujemną.

A właściwie jakie litery dopuszczamy? Czy pozwalamy tylko na alfabet angielski, czy akceptujemy wszelkie ogonki, kropki, kreślenia i akcenty? Otóż możemy ustalić, co nam się podoba, ale trzeba to zrobić na początku zabawy. Czyli zaczynamy od wybrania *alfabetu* – zestawu znaków, z których będziemy składać słowa. Może to być alfabet angielski, może być polski, może być zbiór egipskich hieroglifów albo nawet zestaw $*$, $\bigcirc$, $\square$, $\bullet$. Na przykład komputery, wczytując program, pracują zwykle na słowach nad alfabetem złożonym z małych i dużych liter, cyfr i kilku znaków specjalnych (np. kropka, podkreślenie). A zapisując dane, używają dwuliterowego alfabetu: 0 i 1. Wszystkie znaki w alfabecie, jakkolwiek by one nie wyglądały, będziemy nazywać *literami*.

Jeśli mamy wybrany alfabet, to umiemy budować słowa, więc możemy też tworzyć języki. Językiem będzie dla nas po prostu zbiór słów zbudowanych z liter wybranego alfabetu. Może być skończony, tak jak zbiór wszystkich słów języka polskiego albo język nad alfabetem $\{a, b, c\}$ złożony z czterech słów $a$, $ab$, $abc$ i $abbccc$, albo też język złożony tylko ze słowa pustego. Ale może być nieskończony. Dla dowolnego alfabetu możemy mianowicie wziąć zbiór wszystkich słów, które da się poskładać z jego liter. Poza jednym wyjątkiem – kiedy alfabet ma zero liter i jedyne słowo nad nim to słowo puste – tak utworzony język jest nieskończony.

Przyjrzyjmy się najprostszemu przypadkowi, który ma szanse być ciekawy: alfabetowi złożonemu z jednej litery $a$. Możemy napisać słowo jednoliterowe $a$, dwuliterowe $aa$, trzyliterowe $aaa, \ldots$ Wszystkie słowa, czyli pełny język nad tym alfabetem, to ciągi powtórzeń litery $a$. Słowo puste, oznaczane tradycyjnie $\varepsilon$, możemy rozumieć jako „zero powtórzeń litery $a$”. Ale są też inne języki nieskończone nad tym alfabetem, mniejsze od pełnego, czyli będące jego podzbiorami. Na przykład język słów o parzystej długości $\{\varepsilon, aa, aaaa, aaaaaa, \ldots\}$, język słów o nieparzystej długości $\{a, aaa, aaaaa, aaaaaaa, \ldots\}$, język słów o długościach będących liczbami pierwszymi, język słów o długościach podzielnych przez 157, język słów o długościach będących sześcianami liczb nieparzystych… Mamy mnóstwo możliwości. A jeśli weźmiemy większy alfabet, chociażby dwuliterowy $a$ i $b$, to pojawią się takie przykłady, jak język słów mających na początku $n$ liter $a$, a dalej tyle samo liter $b$:

$$\{\varepsilon, ab, aabb, aaabbb, aaaabbbb, \ldots\},$$

język słów zawierających dokładnie jedną literę $b$:

$$\{b, ab, ba, aab, aba, baa, aaab, aaba, abaa, baaa, aaaab, aaaba, \ldots\},$$

czy język słów zawierających przynajmniej dwa razy więcej liter $b$ niż $a$:

$$\{\varepsilon, b, bb, bba, bab, abb, bbb, bbba, bbab, babb, abbb, bbbb, bbbba, babbb, bbbbaa, bbabab, \ldots\}.$$

Podaliśmy tylko po kilka krótkich słów, ale na pewno, Czytelniku, umiesz te listy przedłużyć, a pewnie też znaleźć algorytm wypisywania słów z każdego z podanych języków, tak żeby żadnego nie pominąć.

Trudno się spodziewać, aby wszystkie języki były tak samo interesujące. Pewnie, na przykład, częściej umiemy wykazać jakieś ciekawe własności języka określonego, jak wyżej, jakąś regułą niż losowego zbioru słów. Dobrym sposobem opisywania porządnych języków są wzory składające się z liter wybranego alfabetu, trzech symboli $+$, $\cdot$ i $*$, o których za chwilę opowiemy, i, jak zwykle, nawiasów dla zaznaczenia kolejności działań. Takie wzory na konstrukcję języka nazywają się *wyrażeniami regularnymi*, a języki, które za ich pomocą można opisać, to *języki regularne*. Są to najporządniejsze z interesujących języków – zwykle, żeby język wyrażał coś ciekawego, musi mieć bardziej skomplikowaną strukturę. Nie zawsze: budowa lekserów (części kompilatorów) jest oparta na językach regularnych, ale już parsery, wykonujące kolejny po analizie leksykalnej krok kompilacji, używają języków opisanych bardziej złożonymi „równaniami" niż wyrażenia regularne.

Zajmijmy się naszymi regułami. Można się domyślić, że $+$ oznacza sumowanie dwóch języków, czyli wrzucenie do jednego języka wszystkich słów z obu danych zbiorów. Dalej, $\cdot$ to też operacja na dwóch językach – nowy język to sklejenia wszystkich słów z pierwszego języka ze wszystkimi z drugiego. Kropkę w zapisie najczęściej się pomija, tak jak znak mnożenia w arytmetyce. Na przykład,

$$\{a, ab, ba\} + \{b, aa\} = \{a, ab, ba, b, aa\},$$
$$\{a, ab, ba\} \cdot \{b, aa\} = \{ab, aaa, abb, abaa, bab, baaa\}.$$

Natomiast $*$ to operacja na jednym języku. Nowy język składa się ze wszystkich sklejeń dowolnej długości słów z podanego języka. Czyli, na przykład, zbiór wszystkich słów złożonych z litery $a$ zapisuje się jako $\{a\}^*$, a zbiór wszystkich słów z liter $a$, $b$ i $c$ to $\{a, b, c\}^*$.

Teraz jeszcze trzeba się przyzwyczaić do tego, że symboli $+$, $\cdot$ i $*$ będziemy używać nie tylko do języków, ale też do wyrażeń opisujących języki. Czyli przez $(\{a\} + \{b\} + \{c\})^*$ rozumiemy zastosowanie operacji $*$ do języka $\{a\} + \{b\} + \{c\} = \{a, b, c\}$. Nawiasy klamrowe wokół pojedynczych słów będziemy dla wygody opuszczać, więc ostatecznie wyrażenie opisujące pełny język nad alfabetem $a$ to po prostu $a^*$, a wyrażenie dla pełnego języka nad alfabetem $a$, $b$, $c$ to $(a + b + c)^*$.

A wyrażenia dla innych języków? Zacznijmy od alfabetu jednoliterowego i słów o parzystej długości. Każde takie słowo można podzielić na pary $aa$, czyli jest ono sklejeniem pewnej liczby słów $aa$. Wobec tego odpowiednim wyrażeniem regularnym jest $(aa)^*$. Ponieważ $*$ obejmuje też możliwość sklejenia zera słów z danego języka, to nie musimy się osobno martwić o słowo puste.

Teraz słowa o długości nieparzystej. Jeśli obetniemy pierwszą literę $a$ z takiego słowa, to zostanie słowo o parzystej długości, czyli ten język powstaje przez sklejanie litery $a$ ze słowami z języka z poprzedniego przykładu. Wystarczy więc dokleić literę $a$ do opisanego wyżej wyrażenia: $a \cdot (aa)^*$. A gdybyśmy napisali $(aa)^* \cdot a$? Jak widać, jeden język może być opisywany przez wiele wyrażeń regularnych.

Z pewnością już wiesz, Czytelniku, że język słów o długości podzielnej przez 157 jest opisywany przez wyrażenie $(a^{157})^*$, gdzie $a^{157}$ zastępuje, oczywiście, odpowiednio długi ciąg, i potrafisz podać wyrażenie dla języka słów o długościach dających resztę 19 z dzielenia przez 91. Co z wyrażeniem dla słów o długościach będących liczbami pierwszymi? Lepiej będzie poszukać dowodu, że takie wyrażenie nie istnieje...

Język nad alfabetem $a$, $b$ złożony ze słów z dokładnie jedną literą $b$ konstruujemy tak: najpierw piszemy pewną liczbę liter $a$ (może zero), później wstawiamy $b$ i kończymy ciągiem $a$ (znów może pustym). Czyli początek i koniec zapisujemy jako $a^*$, a całe wyrażenie to $a^*ba^*$. Możemy też opisać zbiór słów (nad dwuliterowym alfabetem), w których żadna litera nie występuje trzy razy pod rząd. Ten warunek oznacza, że litery $a$ i $b$ występują blokami co najwyżej dwuelementowymi. Weźmy słowo zaczynające się i kończące literą $a$. Wtedy początek to $a$ lub $aa$. Jeśli słowo się na tym nie kończy, to dalej mamy $b$ lub $bb$, potem znów $a$ lub $aa$ i ta sytuacja może się powtarzać. Wobec tego słowa z naszego języka zaczynające się i kończące $a$ są opisywane wyrażeniem $(a + aa)[(b + bb)(a + aa)]^*$. Pozostałe trzy przypadki: słowa zaczynające się $b$ i kończące $a$, zaczynające i kończące $b$ oraz zaczynające $a$ i kończące $b$ możemy opisać podobnie i dodać wyrażenia definiujące cztery części języka:

$$(a + aa)[(b + bb)(a + aa)]^* + [(b + bb)(a + aa)]^* + \\ + (b + bb)[(a + aa)(b + bb)]^* + [(a + aa)(b + bb)]^*.$$

Jak opisać język zawierający trzy kolejne wystąpienia pewnej litery? A język słów o nieparzystej długości? Albo język słów, w których litera $a$ pojawia się nieparzystą liczbę razy? Można rozważać alfabet dwuliterowy, a można większy. Podpowiemy jeszcze, że wyrażeń dla pozostałych przykładów z poprzedniej strony nie ma sensu szukać nawet w długie zimowe wieczory. Ale warto pomyśleć nad dowodem, że ich nie ma.

*Małą Deltę przygotowała Maria DONTEN-BURY*