

Theorem:

If A is a finite alphabet then the set of all words over A $A^* = \bigcup_{j=0}^{\infty} A^j$ is countably infinite.

Proof:

We showed A^j is a finite set for each j . In fact, $\#(A^j) = n^j$. $\bigcup_{j=0}^{\infty} A^j$ is therefore a countably infinite union of disjoint finite sets.

Note that $A^j \cap A^k = \emptyset$ if $j \neq k$ as no words of length j can be of length k if $j \neq k$.

By Corollary 1 to the theorems that a countably infinite union of countably infinite sets is countably infinite,

$A^* = \bigcup_{j=0}^{\infty} A^j$ is countably infinite. (q.e.d)

Corollary 1: If A is a finite alphabet, the set of all languages over A is uncountably infinite.

Proof: Recall that a language L is any subset of words in the alphabet A , hence L is any subset of A^* . Therefore the set of all languages over A is precisely $P(A^*)$, the power set of A^* . We showed in the previous theorem that A^* is countably infinite, i.e. $A^* \sim \mathbb{N} \Rightarrow P(A^*) \sim P(\mathbb{N})$, but we previously proved $P(\mathbb{N})$ is uncountably infinite by putting it in one-to-one correspondence with the set of all sequences of 0s and 1s $\Rightarrow P(A^*)$ is uncountably infinite. (q.e.d)

Corollary 2: The set of all programs in any programming language is countably infinite.

Proof: For any programming language, a program is a finite string over a finite alphabet, the set of characters allowable in that programming language. Let us call this finite alphabet A . Then the set of all programs in the given programming language is A^* . Since $A^* \sim \mathbb{N}$ as proven in the theorem, the set of all programs is countably infinite. (*q.e.d*)

Recall:

Theorem: A language over a finite alphabet is regular \Leftrightarrow it is given by a regular expression.

Recall the definition of a regular expression:

Definition: Let A be an alphabet.

1. \emptyset , ϵ , and all elements of A are regular expressions.
2. If w and w' are regular expressions then $w w'$, $w \cup w'$ and w^* are regular expressions.

Note that regular expressions sometimes have parentheses in order to change the priority of operations $*$, \circ (concatenation) and \cup (union). Therefore, any regular expression over the alphabet A is a string over the enlarged alphabet \tilde{A} .

Theorem: The set of all regular languages over a finite alphabet A is countably infinite.

Proof:

Since the alphabet A is finite, the enlarged alphabet $\tilde{A} = A \cup \{\emptyset, \epsilon, *, \circ, \cup, (,)\}$ is also finite. By the theorem proven earlier, \tilde{A}^* is therefore countably infinite. A regular language then is given by a regular expression, which is a string over the enlarged alphabet \tilde{A} , hence an element of \tilde{A}^* . Therefore, the set of all regular languages over the alphabet A is countably infinite. (*q.e.d*)

Moral of the Story:

Given a finite alphabet A , the set of regular languages (which is countably infinite) is much smaller than the set of all languages over A (which is uncountably infinite). Therefore, regular languages constitute a special category within the set of all languages over a given alphabet.