Basic Elements of Formal Languages

- What is Formal Language?

- What is Natural Language?

- How Formal Language is different from Natural Language?

Answers:

- A Natural Language or Ordinary Language is any language that has evolved naturally through use and repetition without conscious planning or premeditation.

- We discover the grammar of a natural language through empirical investigation.

- We don't discover the grammar of a formal language (artificial language), we stipulate it- we define it however we want.

BASIC ELEMENTS:

- **Symbol:** A *symbol* is an abstract entity that we shall not define formally, just as *point* and *line* are not defined in geometry.

- **Alphabet:** An *alphabet* is a finite, nonempty set of symbols. Conventionally, we use the symbol Σ (capital sigma) for an alphabet. Common alphabets are:

  1. Σ = {0, 1}, the binary alphabet.
  2. Σ = {a, b, ……, z}, the set of all lower-case letters.
  3. The set of all ASCII characters, or the set of all printable ASCII characters.

- **String/Word:** A *string* (or sometimes *word*) is a finite sequence of symbols chosen from some alphabet. Example:

  - 01101 is a string from the binary alphabet Σ = {0, 1}.
  - 111 is another string from the same alphabet.

  Empty String:

  - The *empty string* is the string with zero occurrences of symbols.
  - This string, denoted by ε (epsilon), is a string that may be chosen from any alphabet whatsoever.

  Length of a String:

  - The *length of a string* is the number of positions for symbols in the string. For example, 01101 has length 5.
  - The number of symbols as the length: accepted but not strictly correct. The string 01101 has only two symbols, 0 and 1.

- The standard notation for the length of a string w is |w|.
- Example: |011| = 3 and |ε| = 0.

Prefix of a String:

- A *prefix of a string* is any number of leading symbols of that string.
- Example: String *abc* has prefixes ε, a, ab, and abc.
- A prefix of a string, other than the string itself, is called a *proper prefix*.

Suffix of a String:

- A *suffix* of a string is any number of trailing symbols of that string.
- Example: String abc has suffixes ε, c, bc, and abc.
- A suffix of a string, other than the string itself, is called a *proper suffix*.

Concatenation of Strings:

- The *concatenation* of two strings is the string formed by writing the first, followed by the second, with no intervening space.
- Let *x* and *y* be strings of length *i* and *j* respectively. Then *xy* denotes the concatenation of x and y and the *length* of xy is *i+j*.
- Example: Let x = 1101 and y = 0011. Then xy = 11010011.
- The empty string is the *identity* for the concatenation operator. That is, εw = wε = w.

Powers of an Alphabet: Enigma (7c)

- If Σ is an alphabet, the set of all strings of a certain length from that alphabet can be expressed by using an exponential notation.
- Σ^k is defined as the set of strings of length k, each of whose symbols is in Σ.
- Σ^0 = ε, no matter what the alphabet Σ is. In other words, ε is the only string of length 0.
- Example: If Σ = {a, b, c} then Σ^1 = {a, b, c}, Σ^2 = {aa, ab, ac, ba, bb, bc, ca, cb, cc}, Σ^3 = {aaa, aab, aac, aba, abb, abc, aca, acb, acc, baa, bab, bac, bba, bbb, bbc, bca, bcb, bcc, caa, cab, cac, cba, cbb, cbc, cca, ccb, ccc}.
- The set of all possible strings of all possible lengths over an alphabet Σ is conventionally denoted by Σ^* (*Kleene Star*). For instance, $\{0, 1\}^*$ = {ε, 0, 1, 00, 01, 10, 11, 000, ...}.
- The set of nonempty strings from alphabet Σ is denoted by Σ^+ (Kleene Closure/Plus).

** Confusion between Σ and Σ^1?

- We shall use the same notation for the two sets. Instead, it may be inferred from the context whether we are speaking about an alphabet or a set of strings.

- **Languages:**
  - A set of strings all of which are chosen from some Σ^*, where Σ is a particular alphabet, is called a *(formal) language*.
  - If Σ is an alphabet, and $L \subseteq \Sigma^*$, then L is a language over Σ. A language over Σ need not include strings with all the symbols of Σ.
  - If L is a language over Σ, it is also a language over any alphabet that is a superset of Σ.
  - Complement of a formal language, Σ^* - L.
  - ∅, the empty language, is a language over any alphabet.
  - {ε}, the language consisting of only the empty string, is also a language over any alphabet. Notice that, ∅ ≠ {ε}; the former has no string but the latter has one string.

  Example:
  - The language of all strings consisting of *n* 0's followed by *n* 1's, for some n ≥ 0: {ε, 01, 0011, 000111, ...};
  - The set of strings of 0's and 1's with an equal number of each: {ε, 01, 10, 0011, 0101, ...}.
  - Set-Formers to define Language: {w| w consists of an equal number of 0's and 1's}.