Alphabets, strings and formal languages

An introduction to information representation
**Symbols**

**Definition:** A **symbol** is an object endowed with a denotation (*i.e.* literal meaning)

**Examples:**

- Variables are symbols that have a meaning (*i.e.* denotation) and also take values.
- There are symbols that do not take values. Example: the letter A (as a letter not as a variable, of course).
- A rock (on its own) is not a symbol.
Definition: Let $A$ be a set of symbols. A **concatenation of symbols** of $A$ is the operation of **joining** together an $n$-tuple of symbols of $A$, preserving the order of the symbols in the $n$-tuple.

Definition: The object **generated** by the concatenation of an $n$-tuple of symbols in $A$ is called **string (of length $n$) over $A$**.
Example of string generation

Thus, in order to generate a string $s$ of length $n$ over a given set $A$, select first an $n$-tuple

$$(a_1, a_2, ..., a_n) \in A \times A \times \cdots \times A$$

and then, concatenate its elements preserving the order:

$$s = a_1 a_2 \cdots a_n$$

- Example: Given $A = \{1, a, b\}$

$$(b, a, 1, b)$$

Select an $n$-tuple

$n=4$ in this case

$$s = ba 1b$$

generate the string
Example of string generation

Thus, in order to generate a string $s$ of length $n$ over a given set $A$, select first an $n$-tuple

$$(a_1, a_2, ..., a_n) \in A \times A \times \cdots \times A$$

and then, concatenate its elements preserving the order:

This is a good model for the operation of writing

- Example: Given $n=4$,

Select an $n$-tuple

$$(b, a, 1, b)$$

generate the string

$s = ba1b$
Properties of strings

Strings inherit all the properties of the tuples. In particular,

- The order of the symbols in the string is important
- The number of symbols in a string is important

**Definition:** Two **strings are equal** if and only if they consists of the same **symbols written in the same order**
A special string

**Definition:** The null – string (also called empty string), denoted $\lambda$, is the string written from the tuple ( ).

**Observations:**

1. $\lambda \neq \Phi$
2. For any string $s$, $s\lambda = \lambda s = s$
3. $\lambda\lambda = \lambda$
All **knowledge** is represented in strings

**Strings** are fundamental to **data storage and representation**

- A **book** stores its content as a rather long string of alphanumeric and punctuation symbols, and a special blank space character
- **DNA strings** are biochemical strings for storing the phenotype and other characteristics of living beings
- A **computer program** stores data and instructions as a string of 0 and 1’s
- A **picture** stores an image as an array of strings of pixels
Reading a string

**Definition:** Reading a string is the process of identifying the symbols that compose it, one by one, from left to right.

Since strings are written from $n$-tuples, reading may be interpreted as the reconstruction of the $n$-tuple from the string. Thus,

As an operation reading is (somewhat) the inverse of writing (or concatenation)

- Reading is analysis (identification of components) while concatenation is synthesis (object composition)
Illustration

Consider the string over \( \{x, 1, \Omega\} \) \( s = 1x11x \).
The reading of \( s \) is, step by step:

1. \( 1x11x \rightarrow (1) \)
2. \( 1x11x \rightarrow (1, x) \)
3. \( 1x11x \rightarrow (1, x, 1) \)
4. \( 1x11x \rightarrow (1, x, 1, 1) \)
5. \( 1x11x \rightarrow (1, x, 1, 1, x) \)
6. Halt: string is read!

This is a good model for the operation of reading!
Are write and read functions?

Let’s put our models in mathematical context:

$$Write \subseteq \bigcup_{j=1}^{\infty} \left( \bigcup_{i=1}^{A} X^j \right) \times \bigcup_{j=1}^{\infty} A^j$$

$$Write(a_1, a_2, \ldots, a_j) = \{a_1 a_2 \ldots a_j\}$$

**Write is a function** since a j-tuple over A never generate more than one string!
What about read?

Here is a tentative model:

\[
\text{Read} \subseteq \bigcup_{j=1}^{\infty} A^j \times \bigcup_{j=1}^{\infty} (X \times A)
\]

\[
\text{Read}(a_1a_2\cdots a_j) = \{ (a_1, a_2, \cdots, a_j) \}?
\]

Indeed, this relation is not, in general, a function
Read is not a function

Counterexample:
Let $A = \{1, a, b, b1\}$. Consider the string over $A$:

$s = 1ab1a$

Then,

$\text{Read}(1ab1a) = \{(1, a, b, 1, a), (1, a, b1, a)\}$

Not a singleton: not a function!!
Alphabets

The **reading** of a string depends on the **set of all symbols available for writing**. Some sets render more than one reading.

**Definition**: A finite, nonempty set is an **alphabet** if each string over it renders a unique reading.
Write is one – to – one

Theorem: Over an alphabet $A$, Write is one – to – one

Proof:

Assume that Write is *not* one – to – one. Then, there is a pair of tuples $t, r$ with $t \neq r$, such that $\text{Write}(t) = \text{Write}(r) = s$. But then, $\{t, r\}$ is a subset of $\text{Read}(s)$. This is a contradiction with the assumption that $A$ is an alphabet.
An important corollary

**Corollary**: Over an alphabet $A$, Read is a function and it is in fact, the inverse of Write

**Proof**: By definition of alphabet, for each string $s$ over $A$ as a unique tuple $t$ associated with $\text{Read}(s)$. Therefore, $\text{Read}(s) = \{ t \}$, is a singleton. Also, $\text{Read}(\text{Write}(t)) = t$ and $\text{Write}(\text{Read}(s)) = s$. Thus, Read is the inverse of Write.
Examples

Alphabets:

\[ A=\{a, b, c\}, \quad B=\{0, 1\} \]
\[ C=\{ab, ac, bc\}, \quad D=\{00, 01\} \]

Non-alphabet:

\[ E=\{a, b, ab\} \]
\[ A \cup C = \{a, b, c, ab, ac, bc\} \]

Remark: As the last example shows: the union of two alphabets is not always an alphabet!!!
(Despite what people may say!!!)
Are these artificial examples?

The definition of alphabets containing strings of symbols may look like as an unnatural way of getting a non-alphabet.

This is definitely not the case! Convincing examples are right ahead!

It is worth pointing out that, in general, proving that a set is indeed an alphabet may be difficult: not only sets that contain strings produce more than one reading!
Some biochemical alphabets

- **DNA alphabet:**
  - $D = \{A, T, G, C\}$

- **RNA alphabet:**
  - $R = \{A, U, G, C\}$

- **Protein transcription alphabet:** symbols are "codons". Each codon is formed with three RNA symbols.
Some biochemical alphabets

- **DNA alphabet:**
  - \( D = \{ A, T, G, C \} \)

- **RNA alphabet:**
  - \( R = \{ A, U, G, C \} \)

- **Protein transcription alphabet:** symbols are “codons”. Each codon is formed with three RNA symbols.
Languages

**Definition**: A language or, more specifically, a formal language over an alphabet $A$ is a set of strings over $A$ or the empty set.

**Examples**:

1. Every alphabet is a formal language over itself.
2. The set of natural numbers is a language over the alphabet $A=\{0, 1, 2, \ldots , 9\}$.
More examples of languages

– Let $A=\{0, 1\}$ then $A$ is a language over $A$
– Languages over the binary alphabet
  $\{0, 1\}$ are pervasive in computation
– $A=\{\lambda\}$ is a language (and an alphabet)
– Natural languages are formal languages that in addition to denotation have connotation (intended meaning)
Basic facts about languages

Theorem 5.3:
(a) The union of two languages over the same alphabet is a language
(b) The intersection of two languages over the same alphabet is a language
(c) The difference of two languages over the same alphabet is a language
Proof of (a)
Let \( L \) and \( M \) be languages over the same alphabet \( A \).
Then, \((\forall s)(s \in L \cup M) \Rightarrow s \in L \lor s \in M\)
\[\Rightarrow s \text{ is a string over } A\]
Thus, \( L \cup M \) is a language over \( A \).

Proof of (b)
Let \( L \) and \( M \) be languages over the same alphabet \( A \).
Then, \((\forall s)(s \in L \cap M) \Rightarrow s \in L \land s \in M\)
\[\Rightarrow s \text{ is a string over } A\]
Thus, \( L \cap M \) is a language over \( A \).
Proof of (c)
Let $L$ and $M$ be languages, $L \subseteq M$.
Then, $(\forall s)(s \in M) \ s$ is a string over $A$
Since $M - L \subseteq M$
$(\forall s)(s \in M - L \implies s \in M)$
Therefore, $(\forall s)(s \in M - L \text{ is a string over } A)$
String concatenation

**Definition:** Let $L$ and $S$ be nonempty languages. The **string concatenation over $L$** is defined to be the operation

$$Cat(s, t) = st$$

We define $Cat(L) = \{ st : s \in L \land t \in L \}$

**Example:** Let $L$ be the language of all strings over the alphabet $A=\{a, b, 1\}$. Let’s take $s=abaab$ and $t=1a1ab1$ (two strings in $L$). Their concatenation is

$$Cat(s, t) = st = abaab1a1ab1$$
Properties of concatenation

**Associativity** $\forall s, t, r (s, t, r \text{ strings over } A)$

$$Cat(s, Cat(t, r)) = Cat(Cat(s, t), r)$$

**Non-commutativity**

$(\exists s, t) (s, t \text{ strings over } A) \land Cat(s, t) \neq Cat(t, s)$

**Existence of an identity (null string)**

$(\exists \lambda) (\lambda \text{ string}) \land ((\forall s) (s \text{ string over } A))$

$$Cat(\lambda, s) = Cat(s, \lambda) = s$$

**Observation:** the null string **is not** the empty set
Closure under string concatenation

**Definition:** A language $L$ is said to be **closed under string concatenation** if $L = \text{Cat}(L)$

**Examples:**
- The language $L = \{ \lambda \}$ is closed under string concatenation
- The language $L = \emptyset$ is also closed under string concatenation (Why?)
- The language $L$ of all strings over an alphabet is closed under string concatenation
Closure under string concatenation

**Definition:** A language $L$ is said to be **closed under string concatenation** if $L = \text{Cat}(L)$

**Examples:**
- The language $L = \{\lambda\}$ is closed under string concatenation
- The language $L = \emptyset$ is also closed under string concatenation (Why?)
- The language $L$ of all strings over an alphabet is closed under string concatenation
More examples

The language \( L = \{0, 00, 000, \ldots \} = \{0^n : n \geq 1\} \) is closed under concatenation. In order to see this, take any two strings in \( L \), this is \( s = 0^n \) and \( t = 0^m \), \( s, t \geq 1 \).

Then, \( st = 0^n0^m = 0^{n+m} \in L \)

The language \( L = \{01^n : n \geq 1\} \) is not closed under string concatenation since, for instance: \( Cat (01,011) = 01011 \notin L \)
Why does closure matters?

String concatenation is pervasive as a language operation. If a language is closed under string concatenation, the operation of concatenation is a **language invariant** (i.e. does not alter the language). In particular, if the language is closed under string concatenation, any concatenation satisfies the predicate that defines membership in the language.

**Example:**

\[ Q(w): \text{"w string over \{0, 1\} \land w ends with 0"} \]

\[ L = \{ w : Q(w) = 1 \} \]

Now, since the concatenation of any finite collection of strings over \{0,1\} ending with 0 is a string that ends with 0, \( L \) is closed under string concatenation. Therefore,

\[ (\forall s, w \in L) Q(sw) = 1 \]
Theorem on string concatenation

**Theorem 5.4:** If $L$ is a nonempty language that is closed under string concatenation, then $L$ is finite if and only if $L = \{ \lambda \}$.

**Proof:** If $L = \{ \lambda \}$, since $\text{Cat}(\lambda, \lambda) = \lambda \lambda = \lambda$ we have that $\text{Cat}(L) = \{ \lambda \} = L$. Therefore, $L$ is closed under string concatenation, and finite.

If $L \neq \{ \lambda \}$, then there exists $s \in L \land s \neq \lambda$. 
Proof (cont.)

Since $L$ is closed under string concatenation the following inductive construction produces a infinite subset of $L$

Base case:

$Cat(s, s) = ss = s^2 \in L$ (true because $L$ is closed)

Induction:

$(\forall n)Cat(s, s^{n-1}) = s^n \in L \Rightarrow Cat(s, s^n) = s^{n+1} \in L$

(true because $L$ is closed)

Thus, the set $\{s, s^2, \ldots, s^n, \ldots\} = \{s^n : n \geq 1\} \subseteq L$ is an infinite subset of $L$. And therefore, $L$ is infinite.
Closure of a language

**Definition:** Let $A$ be an alphabet and $L$ be a language over $A$. Then, the *star – closure* (or Klein – closure) of $L$ is the smallest language over $A$ that contains $L$ and is closed under string concatenation. The star – closure of $L$ is denoted $L^*$.

**Remark:** In general, $L$ is a subset of $L^*$. In fact, $L = L^*$ if and only if $L$ is star – closed.
Star – closure of an alphabet

Theorem: Let $A$ be an alphabet. Then, any language $L$ over $A$ is a subset of $A^*$

Proof: Let $s$ be a string in $L$. Since $L$ is a language over $A$, $s$ is the result of a concatenation of symbols in $A$. But then, $s$ is an element of $A^*$
Encodings

Let $A, B$ be two alphabets. An **encoding of $A$ in $B$** is a **one-to-one** mapping

$$e : A \rightarrow B^*$$

satisfying the **additional condition** that the set

$$e(A) = \{ w : w \in B^* \land (\exists x \in A) \land e(x) = w \}$$

is also an **alphabet**.

**Example:** Let $A = \{0, 1\}; B = \{a, b\}$. Then, $e(0) = a; e(1) = b$ is an encoding, but $e(0) = ab; e(1) = abab$, is NOT
Popular examples of encodings

**ASCII**: The ASCII encoding specifies a one-to-one correspondence between an alphabet of letters and punctuation symbols of a written language and a set of hexadecimal numbers. The encoding specifies a one-to-one correspondence between the alphabet $A = \{0, 1, 2, \ldots, 9, A, B, C, D, E, F\}$ and strings over the alphabet $B = \{0, 1, 2, \ldots, 9\}$. Some values are:

- $e(A) = 1000001$
- $e(B) = 1000010$
- $e(C) = 1000011$, etc.

**Hexadecimal numbers**: The encoding specifies a one-to-one correspondence between the alphabet $A = \{0, 1, 2, \ldots, 9, A, B, C, D, E, F\}$ and strings over the alphabet $B = \{0, 1, 2, \ldots, 9\}$. The encoding is:

- $e(0) = 00$
- $e(1) = 01$, $e(9) = 09$
- $e(A) = 10$, $e(B) = 11$, $e(C) = 12$
- $e(D) = 13$, $e(E) = 14$, $e(F) = 15$
Knowledge representation is independent of alphabets.

Encodings are **alphabet translators**. Their existence ensures that the ability to represent objects does not depend on the choice of an alphabet.

Indeed, if \( A, B \) are two different alphabets, and \( w = a_1 a_2 \ldots a_n \) is a string over \( A \). Then, its “translation” as a string over \( B \) is

\[
e(a_1) e(a_2) \ldots e(a_n)
\]
A simple illustration

The word in Spanish language:

CABA

is encoded by ASCII over \{0, 1\}^* as

\[ e(C)e(A)e(B)e(A) = 1000011100000110000101000001 \]