# Languages

ian.mcloughlin@gmit.ie

# Alphabets and strings

Start with a set, call it an alphabet.

$$A = \{0, 1\}$$

Call tuples over the alphabet strings:

$$s = (1, 0, 1, 0, 1, 1) = 101011$$

Don't forget the empty tuple/string:

$$\epsilon = ()$$

Strings can be concatenated, denoted by . or not:

$$s.t = st = 101011.001 = 101011001$$

$$t.s = ts = 001.101011 = 001101011$$

# Set of all strings

Start with an alphabet.

$$A = \{0, 1\}$$

Write $A^i$ for the set of all tuples over $A$ of length $i$ where $i \in \mathbb{N}_0$:

$$A^2 = \{00, 01, 10, 11\}$$

Don't forget the empty tuple/string:

$$A^0 = \{\epsilon\}$$

The set of all strings over $A$ is called the Kleene star $(*)$ of $A$:

$$A^* = \bigcup_{i \in \mathbb{N}_0} A^i$$

# Languages

Each subset of $A^*$ is called a language over $A$:

$$L \subseteq A^*$$

Note the empty set is a language:

$$\{\} \subseteq A^*$$

And so is $A^*$ itself:

$$A^* \subseteq A^*$$

We are typically interested in the proper subsets of $A^*$:

$$L \subseteq A^* \qquad \text{where} \qquad L \neq \{\}, A^*$$

# Union of Languages

Suppose we've two languages over the same alphabet:

$$L_1 = \{000, 111\} \subseteq \{0, 1\}^*$$

$$L_2 = \{101, 010, 111\} \subseteq \{0, 1\}^*$$

The union of $L_1$ and $L_2$ is the set containing the elements of both:

$$L_1 \cup L_2 = \{000, 010, 101, 111\}$$

Remember that sets don't keep count of elements.

# Intersection of Languages

Suppose we've two languages over the same alphabet:

$$L_1 = \{000, 111\} \subseteq \{0, 1\}^*$$

$$L_2 = \{101, 010, 111\} \subseteq \{0, 1\}^*$$

The intersection of $L_1$ and $L_2$ is the set containing the elements in both:

$$L_1 \cap L_2 = \{111\}$$

Two languages can have an empty intersection.

# Concatenation of Languages

Suppose we've two languages over the same alphabet:

$$L_1 = \{000, 111\} \subseteq \{0, 1\}^*$$

$$L_2 = \{101, 010, 111\} \subseteq \{0, 1\}^*$$

The concatenation of $L_1$ and $L_2$ is the set containing the concatenation of each of the elements of the first language with each of the elements of the second:

$$L_1 L_2 = \{000101, 000010, 000111, 111101, 111010, 111111\}$$

If $\epsilon$ is in either language then the elements of the other are in the concatenation.

# Kleene star of a language

Write $L^2$ for the concatenation of $L$ with $L$.

$$L^2 = \{000, 111\}^2 \subseteq \{000000, 000111, 111000, 111111\}^*$$

Set $L^0$ and $L^1$ as follows:

$$L^0 = \{\}, \qquad L^1 = L$$

Then set:

$$L^i = L^{i-1}L \qquad \forall i \in \mathbb{N},\ i > 2$$

Then the Kleene star (*) of the language $L$ is:

$$L^* = \bigcup_{i \in \mathbb{N}_0} L^i = \{\epsilon, 000, 111, 000000, 000111, \ldots\}$$

# Files types as languages

- Computer files are stored as 0's and 1's.
- A file is a string over $\{0, 1\}$
- File types are languages over $\{0, 1\}$.
- Set of all valid pdf files is a language over $\{0, 1\}$.
- As is the set of valid docx files.
- A computer program that converts pdf's to docx's maps one subset of $A^*$ to another.
- Executable files are also strings over $\{0, 1\}$.