Formal Languages and the Generalized Star-Height Problem

Tom Bourne

University of St. Andrews

3rd June 2014
An alphabet $A$ is a non-empty set; its elements are called letters. A word (over $A$) is a finite sequence of letters from $A$. The length of a word is the number of letters it is composed of. The empty word, denoted by $\varepsilon$, is the unique word of length 0.

The set of all words over $A$ is denoted by $A^*$, where $*$ is the Kleene star operation. (For the semigroup theorists: $A^*$ is the free monoid on $A$ with operation word concatenation and identity $\varepsilon$.)

A (formal) language $L$ (over $A$) is a subset of $A^*$. 
Regular Expressions

Given an alphabet $A$, we define $\emptyset$ (the empty set), $\varepsilon$ (the empty word), and $a$ in $A$ to be regular expressions. If $E$ and $F$ are regular expressions then we recursively define new regular expressions by using the following operations:

- $EF$ (concatenation)
- $E \cup F$ (set union)
- $E^*$ (Kleene star)

We use regular expressions to represent languages. For example, if $A = \{a, b\}$ then $A^*a = \{a \cup b\}^*a$ represents the language in which all words end with the letter $a$. 
The **star-height** $h(E)$ of a regular expression $E$ is defined recursively by:

- $h(\emptyset) = h(\varepsilon) = h(a) = 0$, where $a \in A$;
- $h(EF) = h(E \cup F) = \max\{h(E), h(F)\}$;
- $h(E^*) = h(E) + 1$.

For example, over $A = \{a, b\}$, $h(a^*b) = 1$ and $h((a^*b)^*) = 2$.

Then, for $L \subseteq A^*$, we define the **star-height** of $L$ by

$$h(L) = \min\{h(E) : E \text{ is a regular expression for } L\}.$$
Now suppose that in addition to the aforementioned operations for defining regular expressions, we also allow complementation; that is, if $E$ is a regular expression then so is $E^c$.

Including the complement operation leads us to refer to $E$ as a generalized regular expression.

We then define $h(E^c) = h(E)$, and define the generalized star-height of $L \subseteq A^*$ as before.

It is useful to think of the (generalized) star-height of $L$ as the nesting depth of Kleene stars in the regular expression representing $L$ that features minimal Kleene stars.
A language which has (generalized) star-height 0 is said to be star-free. We have the following result:

**Schützenberger’s Theorem:** A language is star-free if and only if its syntactic monoid is finite and aperiodic.

This theorem gives us an algorithm for deciding whether a language has (generalized) star-height 0. However, the following is still an open question:

**The Generalized Star-Height Problem:** Does there exist a language of generalized star-height greater than 1?
Tackling the Problem

Most of the work that considers this problem was done by Pin, Straubing and Thérien in the late 80s/early 90s. It followed in the footsteps of Schützenberger by studying the syntactic monoids of languages.

Rather than trying to classify individual languages, they considered families of languages, coined \textit{varieties} in order to match up with the notion from universal algebra.

Research on the topic dwindled as researchers started to explore complexity theory and there have been no major results regarding the problem in the last 20 years.