Unambiguity of Extended Regular Expressions in SGML Document Grammars

Anne Brüggemann-Klein*

Abstract

In the Standard Generalized Markup Language (SGML), document types are defined by context-free grammars in an extended Backus-Naur form. The right-hand side of a production is called a *content model*. Content models are extended regular expressions that have to be unambiguous in the sense that "an element ... that occurs in the document instance must be able to satisfy only one primitive content token without looking ahead in the document instance." In this paper, we present a linear-time algorithm that decides whether a given content model is unambiguous.

A similar result has previously been obtained not for content models but for the smaller class of standard regular expressions. It relies on the fact that the languages of marked regular expressions are local—a property that does not hold any more for content models that contain the new & operator. Therefore, it is necessary to develop new techniques for content models.

Besides solving an interesting problem in formal language theory, our results are relevant for developers of SGML systems. In fact, our definitions are causing changes to the revised edition of the SGML standard, and the algorithm to test content models for unambiguity has been implemented in an SGML parser.

*Institut für Informatik, Universität Freiburg, Rheinstr. 10–12, 79104 Freiburg, Germany. E-mail: brueggem@informatik.uni-freiburg.de.

1 Introduction

The Standard Generalized Markup Language (SGML) is an ISO standard that provides a syntactic meta-language for the definition of textual markup systems, which are used to indicate the structure of documents so that they can be electronically typeset, searched, and communicated [ISO86, Bar89, Gol90].

In SGML, document types are defined by context-free grammars in an extended Backus-Naur form. The right-hand side of a production is called a *content model*. Content models are similar to standard regular expressions; they are built from symbols in an alphabet Σ with unary operators ?, *, and $^+$ and binary operators $+$, $\cdot$, and &. The language $\mathrm{L}(E)$ represented by a content model E is defined inductively:

$$
\begin{aligned}
&\mathrm{L}(a) = \{a\} \text{ for } a \in \Sigma\\
&\mathrm{L}(F+G) = \mathrm{L}(F) \cup \mathrm{L}(G)\\
&\mathrm{L}(F \,\&\, G) = \mathrm{L}(FG + GF)\\
&\mathrm{L}(F?) = \mathrm{L}(F) \cup \{\epsilon\}\\
&\mathrm{L}(F^*) = \{v_1 \dots v_n \mid n \geq 0,\ v_1, \dots, v_n \in \mathrm{L}(F)\}\\
&\mathrm{L}(F^+) = \{v_1 \dots v_n \mid n \geq 1,\ v_1, \dots, v_n \in \mathrm{L}(F)\}
\end{aligned}
$$

Note that neither $\emptyset$ nor ϵ are syntactic constituents of content models. It is not hard to see that the languages denoted by content models are exactly the regular languages that are different from $\emptyset$ and $\{\epsilon\}$.

In an SGML document grammar, only those content models are allowed that are *unambiguous* in the sense of Clause 11.2.4.3 of the standard, as cited in the abstract. In other words, only such content models are valid that enable us to uniquely determine which appearance of a symbol in a content model matches the next symbol in an input word without looking beyond that symbol in the input word. For example, $((a+b)^*a)?$ is ambiguous, whereas $(b^*a)^*$ is unambiguous.

This definition gives rise to the following question, both of theoretical interest and of relevance for systems supporting SGML: Given a content model E, how can we decide whether E is unambiguous? This question has already been answered not for content models, but for the smaller class of standard regular expressions [BW92, Bru92a, Bru92b]. (As usual, a *standard regular expression* is built from ϵ, $\emptyset$, and symbols in Σ with the unary operator * and the binary operators $\cdot$ and $+$.) In this paper, we give a rigorous definition of unambiguity for content models, and we present an optimal-time algorithm to test content models for unambiguity.

The definition of unambiguity is based on the concept of marking a content model; that is, assigning different subscripts to different occurrences of the same symbol. For example, $(a? \,\&\, b)a^+$ can be marked as $(a_1? \,\&\, b_1)a_2^+$. Now, each word denoted by the content model corresponds to at least one sequence of subscripted symbols. In our example, $baaa$ corresponds to $b_1a_1a_2a_2$ and to $b_1a_2a_2a_2$. If we mark a standard regular expression, the language L of the marked expression is local [Pin92]: If $u_1xu_2, v_1xv_2 \in L$ for some words u_1, u_2, v_1, and v_2 of subscripted symbols and some subscripted symbol x, then $u_1xv_2, v_1xu_2 \in L$. This observation leads to a decision algorithm for unambiguity of standard regular expressions [Bru92a, Bru92b]. However, this algorithm is not capable of dealing with content models containing & operators because they do in general not have the property of locality. In our example above, the possible continuations of the word b_1a_1 are a_2^n, $n \geq 1$, but the possible continuations of the word a_1 are $b_1a_2^n$, $n \geq 1$. Clark [Cla92] has proposed a modification of our algorithm [Bru92a, Bru92b] that also works for content models with & operators. Yet he has not given a proof *why* this algorithm works. We describe this approach in Section 3, give a formal proof of correctness, and optimize the running time. Our results imply also that the unambiguity test that is implemented in the SGML parser *sgmls* [Cla92] is correct.

2 Definitions

In this paper, we consider *extended regular expressions* built from ϵ and symbols in Σ with unary operators ?, *, and $^+$ and binary operators $+$, $\cdot$, and &. Since $\emptyset$ can be eliminated as a syntactic constituent without disturbing unambiguity, we do not allow it in extended regular expressions. From now on, we use the term *expression* for extended regular expressions.

A *marked* expression E is an expression over Σ', the alphabet of subscripted symbols, such that each subscripted symbol occurs at most once in E. For a subscripted symbol x, let $\chi(x)$ denote the underlying symbol in Σ. We use uppercase letters from E through J as variables for expressions and for marked expressions, a, b, and c for symbols in Σ, x, y, and z for subscripted symbols in Σ', also called positions, and u, v, and w for words over Σ or over Σ'.

Note that, given a *marking* E' of expression E, the words of $\mathrm{L}(E)$ can be obtained from the words of $\mathrm{L}(E')$ by dropping the subscripts. In general, several words in $\mathrm{L}(E')$ may correspond to a single word in $\mathrm{L}(E)$. For example, let $E = a?(a+b)^*$ and $E' = a_1?(a_2+b_3)^*$. Then, aaa in $\mathrm{L}(E)$ corresponds to $a_1a_2a_2$ and to $a_2a_2a_2$ in $\mathrm{L}(E')$. We can now give a concise definition of what the SGML standard calls unambiguous.

Definition 2.1 Let E' be a marking of the expression E. Then, E' is *unambiguous* if and only if for all words u, v, and w over Σ' and all symbols x, y in Σ' holds: If uxv and uyw are in $\mathrm{L}(E')$ and if $\chi(x) = \chi(y)$, then $x = y$. The expression E is *unambiguous* if and only if E' is unambiguous.

It is not hard to see that this definition is independent of the marking chosen for E.

3 The decision algorithm for unambiguity

In this section, we compute, given a marked expression E, all pairs of positions in E that *compete* in the sense of Definition 3.1 below. Then, E is unambiguous if and only if any two competing positions of E have different underlying symbols in Σ.

Definition 3.1 Let E be a marked expression. Two positions x and y *compete* in E if and only if there are words u, v, and w such that uxv and uyw are in $\mathrm{L}(E)$.

Definition 3.2 For a marked expression E, we define

$$sym(E) = \{y \mid vyw \in \mathrm{L}(E) \text{ for some words } v \text{ and } w \text{ over } \Sigma', \}$$

$$first(E) = \{y \mid yw \in \mathrm{L}(E) \text{ for some word } w \text{ over } \Sigma'\},$$

$last(E) = \{y \mid wy \in \mathrm{L}(E)$ for some word w over $\Sigma'\}$, and

$follow(E, x) = \{y \mid vxyw \in \mathrm{L}(E)$ for some words v and w over $\Pi\}$,

for each x in $sym(E)$.

It is not hard to see that, for a marked standard regular expression E, positions x and y of E compete if and only if $x, y \in first(E)$ or if there is a position z of E such that $x, y \in follow(E, z)$. This fact is due to the principle of locality. However, if E contains an & operator, there may be positions x, y, and z such that $x, y \in follow(E, z)$ yet x and y do not compete. Consider $E = (I \,\&\, J)H$, $z \in last(J)$, $x \in first(I)$, $y \in first(H)$. Then we have $x, y \in follow(E, z)$. Yet for any prefix uz of a word in $\mathrm{L}(E)$, either u has not yet satisfied I and, thus, uz may be continued with x but (assuming $\epsilon \notin \mathrm{L}(I)$) not with y, or u has already satisfied I and thus, uz may be continued with y but not with x. Therefore, x and y do not compete. Clark [Cla92] has proposed to consider a subset $follow^{-}(E, z)$ of $follow(E, z)$ that in this case no longer contains x. In Theorem A, we characterize competing positions using Clark's definition.

Definition 3.3 For a marked expression E, we define $follow^{-}(E, x)$ for x in $sym(E)$ by induction on E as follows.

$E = x$:

$$follow^{-}(E, x) = \emptyset.$$

$E = F + G$:

$$follow^{-}(E, x) = \begin{cases} follow^{-}(F, x) & \text{if } x \in sym(F), \\ follow^{-}(G, x) & \text{if } x \in sym(G). \end{cases}$$

$E = FG$:

$$follow^{-}(E, x) = \begin{cases} follow^{-}(F, x) & \text{if } x \in sym(F),\ x \notin last(F), \\ follow^{-}(F, x) & \\ \quad \cup\, first(G) & \text{if } x \in last(F), \\ follow^{-}(G, x) & \text{if } x \in sym(G). \end{cases}$$

$E = F \mathbin{\&} G$:

$$follow^-(E,x) = \begin{cases} follow^-(F,x) & \text{if } x \in sym(F),\ x \notin last(F) \\ & \text{or if } x \in last(F),\ \epsilon \notin \mathrm{L}(G), \\ follow^-(F,x) & \\ \quad \cup\, first(G) & \text{if } x \in last(F),\ \epsilon \in \mathrm{L}(G), \\ follow^-(G,x) & \text{if } x \in sym(G),\ x \notin last(G) \\ & \text{or if } x \in last(G),\ \epsilon \notin \mathrm{L}(F), \\ follow^-(G,x) & \\ \quad \cup\, first(F) & \text{if } x \in last(G),\ \epsilon \in \mathrm{L}(F). \end{cases}$$

$E = F?$:

$$follow^-(E,x) = follow^-(F,x).$$

$E = F^*, F^+$:

$$follow^-(E,x) = \begin{cases} follow^-(F,x) & \text{if } x \in sym(F),\ x \notin last(F), \\ follow^-(F,x) & \\ \quad \cup\, first(F) & \text{if } x \in last(F). \end{cases}$$

Theorem A *Let E be a marked expression. Then, positions x and y of E compete if and only if one of the following three conditions holds:*

1. *$x, y \in first(E)$,*

2. *there is a position z of E such that $x, y \in follow^-(E,z)$, or*

3. *there is a subexpression $I \mathbin{\&} J$ or $J \mathbin{\&} I$ of E and a position $z \in last(I)$ such that $x \in follow^-(I,z)$ and $y \in first(J)$ or $y \in follow^-(I,z)$ and $x \in first(J)$.*

Theorem B *A marked expression E is unambiguous if and only if it satisfies the following three conditions:*

1. *If $x, y \in first(E)$ and $\chi(x) = \chi(y)$, then $x = y$.*

2. *If $x, y \in follow^-(E,z)$ and $\chi(x) = \chi(y)$, then $x = y$.*

3. *If $I \mathbin{\&} J$ or $J \mathbin{\&} I$ is a subexpression of E and $z \in last(I)$ such that $x \in follow^-(I,z)$ and $y \in first(J)$, then $\chi(x) \neq \chi(y)$.*

Theorem A immediately implies Theorem B. We prove now Theorem A. As an auxiliary result we use Lemma 3.2, which establishes a weak form of locality for extended expressions. Four types of arguments that are valid only in the context of marked expressions repeat themselves throughout the proofs. We present these arguments first in Lemma 3.1 below. For a set S of symbols, the *S-prefix* (*S-suffix*) of a word is its longest prefix (suffix) that consists only of symbols in S.

Lemma 3.1 *Let E be a marked expression.*

1. *Let $E = F + G$ or $E = G + F$. If w in $\mathrm{L}(E)$ contains a symbol in $sym(F)$, then $w \in \mathrm{L}(F)$ and $w \notin \mathrm{L}(G)$.*

2. *Let $E = FG$. If $uxv \in \mathrm{L}(E)$ and $x \in sym(F)$, we factorize v as $\dot{v}\ddot{v}$ where $\dot{v}$ is the $sym(F)$-prefix of v; then $ux\dot{v} \in \mathrm{L}(F)$ and $\ddot{v} \in \mathrm{L}(G)$. Analogously, if $x \in sym(G)$, we factorize u as $\dot{u}\ddot{u}$ where $\ddot{u}$ is the $sym(G)$-suffix of u; then $\dot{u} \in \mathrm{L}(F)$ and $\ddot{u}xv \in \mathrm{L}(G)$.*

3. *Let $E = F \,\&\, G$ or $E = G \,\&\, F$. If $uxv \in \mathrm{L}(E)$ and $x \in sym(F)$, we factorize u as $\dot{u}\ddot{u}$ and v as $\dot{v}\ddot{v}$ where $\ddot{u}$ is the $sym(F)$-suffix of u and $\dot{v}$ is the $sym(F)$-prefix of v; then $\ddot{u}x\dot{v} \in \mathrm{L}(F)$, one of $\dot{u}$ and $\ddot{v}$ is in $\mathrm{L}(G)$ and the other one is the empty word.*

4. *Let $E = H^*$ or $E = H^+$ and $H = F + G$, $H = G + F$, $H = FG$, $H = GF$, $H = F\&G$, or $H = G\&F$. If $uzv \in \mathrm{L}(E)$ and $z \in sym(F)$, then $\dot{u}z\dot{v} \in \mathrm{L}(F^*)$ where $\dot{u}$ is the $sym(F)$-suffix of u and $\dot{v}$ is the $sym(F)$-prefix of v. If H is a concatenation and $\epsilon \notin \mathrm{L}(G)$, then even $\dot{u}z\dot{v} \in \mathrm{L}(F)$.*

Lemma 3.2 *Let E be a marked expression and $z \in last(E)$. Then $x \in follow^-(E, z)$ if and only if $uz, uzxv \in \mathrm{L}(E)$ for some u and v.*

Proof The proof is by induction on the size of E.

$E = \epsilon$: The precondition that $z \in last(E)$ cannot be satisfied.

$E = z$: Neither $x \in follow^-(E, z)$ nor $uzxv \in \mathrm{L}(E)$.

$E = F + G$: Without loss of generality, $z \in sym(F)$; that is, $z \in last(F)$ by Lemma 3.1. By definition, $follow^-(E, z) = follow^-(F, z)$. Furthermore, $uz, uzxv \in \mathrm{L}(E)$ if and only if $uz, uzxv \in \mathrm{L}(F)$, by Lemma 3.1. An application of the induction hypothesis to F completes the proof.

$E = FG$: Since $z \in last(E)$, the language of E is not empty, and neither are the languages of F nor of G. We divide the proof into three cases. First, let $z \in sym(G)$; that is, $z \in last(G)$, by Lemma 3.1. Then $follow^{-}(E, z) = follow^{-}(G, z)$. Finally, by Lemma 3.1, $uz, uzxv \in \mathrm{L}(E)$ for some u and v if and only if $\dot{u}z, \dot{u}zxv \in \mathrm{L}(G)$ for some $\dot{u}$ and v, because $\mathrm{L}(F) \neq \emptyset$. An application of the induction hypothesis to G completes the proof for this case.

Next, let $z \in sym(F)$ and $x \in sym(G)$. Then $x \in follow^{-}(E, z)$ if and only if $z \in last(F)$ and $x \in first(G)$; but this means there are u and v such that $uz \in \mathrm{L}(F)$ and $xv \in \mathrm{L}(G)$, or, equivalently, $uz, uzxv \in \mathrm{L}(E)$, by Lemma 3.1. Hence, in this case we have a direct proof that does not resort to the induction hypothesis.

Finally, let $z, x \in sym(F)$; in particular, $z \in last(F)$ by Lemma 3.1. Then $x \in follow^{-}(E, z)$ if and only if $x \in follow^{-}(F, z)$. Furthermore, $uz, uzxv \in \mathrm{L}(E)$ for some u and v if and only if $uz, uzx\dot{v} \in \mathrm{L}(F)$ for some u and $\dot{v}$, because $\mathrm{L}(G) \neq \emptyset$. An application of the induction hypothesis to F completes the proof.

$E = F \,\&\, G$: As in the previous case, neither $\mathrm{L}(F)$ nor $\mathrm{L}(G)$ are empty. Without loss of generality, $z \in sym(F)$; thus, $z \in last(F)$, by Lemma 3.1. If $x \in sym(F)$, then $x \in follow^{-}(E, z)$ if and only if $x \in follow^{-}(F, z)$; furthermore, $uz, uzxv \in \mathrm{L}(E)$ for some u and v if and only if $\dot{u}z, \dot{u}zx\dot{v} \in \mathrm{L}(F)$ for some $\dot{u}$ and $\dot{v}$, by Lemma 3.1; hence, the claim follows by the induction hypothesis for F.

On the other hand, if $x \in sym(G)$, then $x \in follow^{-}(E, z)$ if and only if $z \in last(F)$ $x \in first(G)$, and $\epsilon \in \mathrm{L}(G)$; but this means there are u and v such that $uz \in \mathrm{L}(F)$, $xw \in \mathrm{L}(G)$, and still $\epsilon \in \mathrm{L}(G)$, or, equivalently, $uz, uzxw \in \mathrm{L}(E)$, by Lemma 3.1.

$E = F?$: This case is proved by a simple application of the induction hypothesis to F.

$E = H^*$ or $E = H^+$: To make sure that the induction hypothesis can be applied, we observe that $last(E) = last(H)$ and, hence, $z \in last(H)$. Now we prove the two implications of the lemma separately.

First, let $x \in follow^-(E, z)$. We wish to prove that $uz, uzxv \in \mathrm{L}(E)$ for some u and v. If $x \in follow^-(H, z)$, we only have to apply the induction hypothesis to H. On the other hand, if $x \notin follow^-(H, z)$, then $x \in first(H)$; therefore, $uz, xv \in \mathrm{L}(H)$ for some u and v; that is, $uz, uzxv \in \mathrm{L}(E)$.

Second, let $uz, uzxv \in \mathrm{L}(E)$. To prove that $x \in follow^-(E, z)$, we carry out a case analysis that depends on the structure of H. Since $uz \in \mathrm{L}(E)$, it cannot happen that $H = \epsilon$. If $H = y$, then $z = x = y$ and, hence, $x \in follow^-(E, z)$. If $H = F?$, $H = F^*$, or $H = F^+$, the proof is an easy application of the induction hypothesis to F^*, which is smaller than E. We demonstrate now the case when $H = F + G$, $H = FG$, or $H = F \,\&\, G$. Without loss of generality, $z \in sym(F)$; that is, $z \in last(F)$. We consider two cases. If $x \in sym(G)$, then $uzxv \in \mathrm{L}(E)$ implies that $z \in last(F)$ and $x \in first(G)$; hence, $x \in follow^-(E, z)$. On the other hand, if $x \in sym(F)$, we consider the $sym(F)$-suffix $\dot{u}$ of u and the $sym(F)$-prefix $\dot{v}$ of v; then $\dot{u}z, \dot{u}zx\dot{v} \in \mathrm{L}(F^*)$, by Lemma 3.2; the induction hypothesis for F^* implies that $x \in follow^-(F^*, z)$; then either $x \in follow^-(E^*, z)$, and we are done, or H is a concatenation and $\epsilon \notin \mathrm{L}(G)$; in the latter case, however, $\dot{u}z, \dot{u}zx\dot{v} \in \mathrm{L}(F)$ even, and an application of the induction hypothesis to F completes the proof.

$\square$

We are now ready to prove the left-to-right direction of Theorem A. If the two positions x and y compete, then $x, y \in first(E)$ or there are u, v, w, and z such that $uzxv, uzyw \in \mathrm{L}(E)$. In the latter case, there are three possibilities: First, $x, y \in follow^-(E, z)$; second, exactly one of x and y is in $follow^-(E, z)$; or, third, $x, y \notin follow^-(E, z)$. In the first case we are done; the second case is dealt with in Lemma 3.3 and the third case in Lemmas 3.4 and 3.5 below.

Lemma 3.3 *Let E be a marked expression. If $uzxv, uzyw \in \mathrm{L}(E)$, $x \in follow^-(E, z)$, and $y \notin follow^-(E, z)$, then $x \in follow^-(F, z)$ and $y \in$*

$first(G)$ for some subexpression $F \,\&\, G$ or $G \,\&\, F$ of E where $z \in last(F)$.

PROOF The proof is by induction on the size of E.

$E = \epsilon$ or $E = x$: In these cases, the preconditions of the lemma are not satisfied.

$E = F + G$: By Lemma 3.1 and the definition of $follow^{-}(E, z)$, all preconditions for an application of the induction hypothesis to F respectively to G are satisfied, depending on whether $z \in sym(F)$ or $z \in sym(G)$.

$E = FG$: If $z \in sym(G)$, we can, by Lemma 3.1, reduce the proof to an application of the induction hypothesis to G. If $z \in sym(F)$, an application of Lemma 3.1 to $uzyw$ yields $y \in sym(F)$, since $y \notin follow^{-}(E, z)$. By Lemma 3.2, $x \in sym(F)$ as well, since otherwise $uz, uzy\dot{w} \in \mathrm{L}(F)$, yet $y \notin follow^{-}(F, z)$. we are now prepared for an application of the induction hypothesis to F, since, by Lemma 3.1, $uzx\dot{v}, uzy\dot{w} \in \mathrm{L}(F)$ for some $\dot{v}$ and $\dot{w}$ and $x \in follow^{-}(F, z)$, $y \notin follow^{-}(F, z)$.

$E = F \,\&\, G$: Without loss of generality, $z \in sym(F)$. First we consider the case when $\epsilon \notin \mathrm{L}(G)$; in particular, $x \in follow^{-}(F, z)$. If $y \in sym(G)$, then $z \in last(F)$ and $y \in first(G)$ and the claim is proved. On the other hand, if $y \in sym(F)$, then $y \notin follow^{-}(F, z)$ and $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F)$ for some $\dot{u}$, $\dot{v}$, and $\dot{w}$, by Lemma 3.1 applied to $uzxv$ and $uzyw$; hence, we can now apply the induction hypothesis to F.

Next we consider the case when $\epsilon \in \mathrm{L}(G)$. Then $y \in sym(F)$, because $y \notin follow^{-}(E, z)$. As in the case when $E = FG$, Lemma 3.2 implies that $x \in sym(F)$; that is, $x \in follow^{-}(F, z)$; by Lemma 3.1, $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F)$ for some $\dot{u}$, $\dot{v}$, and $\dot{w}$. We can now apply the induction hypothesis to F again.

$E = H^{*}$ or $E = H^{+}$: As in the proof of Lemma 3.2, we carry out a case analysis that depends on the structure of H. In the cases when $H = \epsilon$ or $H = x$, the preconditions of the lemma are not satisfied. If $H = F?$, $H = F^{*}$, or $H = F^{+}$, the proof is an easy application of the induction hypothesis to F^{*}, which is smaller than E. We demonstrate now the cases when $H = F + G$, $H = FG$, or $H = F \,\&\, G$. Then $y \in sym(F)$, since $y \notin follow^{-}(E, z)$. If $H = FG$ and $\epsilon \notin \mathrm{L}(G)$, then $x \in sym(G)$ implies that $\dot{u}z, \dot{u}zy\dot{w} \in \mathrm{L}(F)$ for some $\dot{u}$ and $\dot{w}$, whereas $y \notin \mathit{follow}^-(F, z)$, in contradiction to Lemma 3.2; thus, $x \in \mathit{sym}(F)$; that is, $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F)$ for some $\dot{u}$, $\dot{v}$ and $\dot{w}$, $x \in \mathit{follow}^-(F, z)$, and $y \notin \mathit{follow}^-(F, z)$; hence, in this case, an application of the induction hypothesis to F completes the proof. In all other cases, $x \in \mathit{sym}(G)$ implies that $\dot{u}z$, $\dot{u}zy\dot{w} \in \mathrm{L}(F^*)$ for some $\dot{u}$ and $\dot{w}$. Yet it contradicts Lemma 3.2 that $y \notin \mathit{follow}^-(F^*, z)$. Thus, $x \in \mathit{sym}(F)$; that is, $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F^*)$ for some $\dot{u}$, $\dot{v}$ and $\dot{w}$, $x \in \mathit{follow}^-(F^*, z)$, and $y \notin \mathit{follow}^-(F^*, z)$; now an application of the induction hypothesis to F^* completes the proof.

$\square$

Lemma 3.4 *Let E be a marked expression. If $uzxv, uzyw \in \mathrm{L}(E)$, $x, y \notin \mathit{follow}^-(E, z)$, then $x, y \in \mathit{first}(H)$ for some subexpression H of E.*

PROOF The proof is by induction on the size of E.

$E = \epsilon$ or $E = x$: The precondition that $uzxv \in \mathrm{L}(E)$ is not satisfied.

$E = F + G$: By Lemma 3.1, the proof is reduced to the induction hypothesis for F or G, depending on whether $z \in \mathit{sym}(F)$ or $z \in \mathit{sym}(G)$.

$E = FG$: If $z \in \mathit{sym}(G)$, then $\dot{u}zxv, \dot{u}zyw \in \mathrm{L}(G)$ for some $\dot{u}$ and $x, y \notin \mathit{follow}^-(G, z)$. On the other hand, if $z \in \mathit{sym}(F)$, then $x, y \in \mathit{sym}(F)$, since $x, y \notin \mathit{follow}^-(E, z)$; therefore, $uzx\dot{v}, uzy\dot{w} \in \mathrm{L}(F)$ for some $\dot{v}$ and $\dot{w}$ and $x, y \notin \mathit{follow}^-(F, z)$. We can now apply the induction hypothesis to F and G, respectively.

$E = F \mathbin{\&} G$: Without loss of generality, let $z \in \mathit{sym}(F)$. First, if $x, y \in \mathit{sym}(F)$ as well, then $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F)$ for some $\dot{u}$, $\dot{v}$ and $\dot{w}$ and $x, y \notin \mathit{follow}^-(F, z)$; we can then apply the induction hypothesis to F. Second, if $x, y \in \mathit{sym}(G)$, then $x, y \in \mathit{first}(G)$ and the claim is proved. Finally, we show by contradiction that either both or none of x and y must belong to $\mathit{sym}(F)$; without loss of generality, we assume $x \in \mathit{sym}(F)$ and $y \in \mathit{sym}(G)$. Then, by Lemma 3.2, $uzx\dot{v}, uz \in \mathrm{L}(F)$ for some $\dot{v}$; by Lemma 3.2, $x \in \mathit{follow}^-(F, z)$, in contradiction to the assumption that $x \notin \mathit{follow}^-(E, z)$.

$E = H^*$ or $E = H^+$: Again we carry out a case analysis that depends on the structure of H. If $H = \epsilon$ or $H = x$, the preconditions are not satisfied. If $H = F?$, $H = F^*$, or $H = F^+$, the proof is an easy application of the induction hypothesis to F^*, which is smaller than E. We demonstrate now the cases when $H = F + G$, $H = FG$, or $H = F \,\&\, G$. Without loss of generality, $z \in sym(F)$. Then $x, y \in sym(F)$, because $x, y \notin follow^-(E, z)$. In the subcase when $H = F \& G$ and $\epsilon \notin \mathrm{L}(G)$, we have $x, y \notin follow^-(F, z)$ and $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F)$ for some $\dot{u}$, $\dot{v}$ and $\dot{w}$. In all other subcases, $x, y \notin follow^-(F^*, z)$ and $\dot{u}zx\dot{v}, \dot{u}zy\dot{w} \in \mathrm{L}(F^*)$. We can complete the proof by applying the induction hypothesis to F in the former case and to F^* in the latter case.

$\square$

Lemma 3.5 *Let H be a subexpression of a marked expression E. Then $first(H) \subseteq first(E)$ or $first(H) \subseteq follow^-(E, z)$ for some z in $sym(E)$.*

The proof is a straightforward induction on E that we omit.

Finally, we prove the right-to-left direction of Theorem A. We wish to show that each of the three conditions in the lemma implies that x and y compete in E. Obviously, this is correct for the first condition. For the second condition, we carry out a structural induction on E. So we assume that $z \in sym(E)$ and $x, y \in follow^-(E, z)$. We have to demonstrate that x and y compete in E.

$E = \epsilon$ or $E = z$: The preconditions are not satisfied.

$E = F + G$: This case is an easy application of the induction hypothesis to F respectively to G, depending on whether $z \in sym(F)$ or $z \in sym(G)$.

$E = FG$ or $E = F \,\&\, G$: Since $z \in sym(E)$, neither $\mathrm{L}(F)$ nor $\mathrm{L}(G)$ are empty. Without loss of generality, $z \in sym(F)$. If $x, y \in sym(F)$ as well, then $x, y \in follow^-(F, z)$ and we can apply the induction hypothesis to F; since $\mathrm{L}(G) \neq \emptyset$, this completes the proof. If $x, y \in sym(G)$, then $x, y \in first(G)$; that is, $xv, yw \in \mathrm{L}(G)$ for some v and w; since $\mathrm{L}(F) \neq \emptyset$, x and y compete in E. Finally, we assume without loss of generality that $x \in sym(F)$ and $y \in sym(G)$; then $x, y \in follow^-(E, z)$ implies that $z \in last(F)$, $x \in follow^-(F, z)$, and $y \in first(G)$; that is, by

Lemma 3.2, $uz, uzxv \in \mathrm{L}(F)$ and $yw \in \mathrm{L}(G)$ for some u, v and w; hence, $uzxvyw, uzyw \in \mathrm{L}(E)$.

$E = F^*$ or $E = F^+$: If $x, y \in follow^-(F, z)$, then we have only to apply the induction hypothesis to F. If $x, y \notin follow^-(F, z)$, then $x, y \in follow^-(E, z)$ implies that $x, y \in first(F)$ and we are done. Finally, we assume without loss of generality that $x \in follow^-(F, z)$ and $y \notin follow^-(F, z)$. The fact that $y \in follow^-(E, z)$ implies that $z \in last(F)$ and $y \in first(F)$; since $x \in follow^-(F, z)$, we can apply Lemma 3.2 and get $uz, uzxv \in \mathrm{L}(F)$ for some u and v; furthermore, $yw \in \mathrm{L}(F)$ for some w; hence, $uzxv, uzyw \in \mathrm{L}(E)$.

Finally, we turn to the third condition. If $E = F \,\&\, G$ or $E = G \,\&\, F$ and $z \in last(F)$, $x \in follow^-(F, z)$, and $y \in first(G)$, then, as in the induction above, $uz, uzxv \in \mathrm{L}(F)$ and $yw \in \mathrm{L}(G)$ for some u, v, and w, and, thus, $uzxvyw, uzyw \in \mathrm{L}(E)$; hence, x and y compete in E. We complete the proof with the following observation:

Lemma 3.6 *If x and y compete in a subexpression H of a marked expression E, then they compete in E as well.*

The proof of this observation is a straightforward induction on E that we omit.

Theorem C *For a fixed-size alphabet, it can be decided, for an expression E, in time linear in the size of E whether E is unambiguous.*

PROOF We only sketch the proof for the case when E is marked. To test E for unambiguity, we compute $first(E)$, $last(E)$, and $follow^-(E, x)$ for x in $sym(E)$ bottom up from the subexpressions of E. Definition 3.3 gives the equations necessary to compute $follow^-$. We are especially interested in the case when all the set unions in Definition 3.3 are disjoint; that is, E is in *star normal form* as defined below. In this case, we can partition the computation of $follow^-(E, x)$ into constant-time steps such that each step computes a new element in $follow^-(E, x)$ for some x in $sym(E)$. During this computation, we also monitor conditions 2 and 3 of Theorem B. Therefore, we can detect any unambiguity in E before more than time linear in the size of E has been spent. The next theorem justifies the assumption that E be in star normal form. It is a generalization of an earlier result for standard regular expressions [Bru92a, Bru92b]. □

Definition 3.4 Let E' be a marking of the expression E. Then, E' is in *star normal form* if and only if for each subexpression H^* or H^+ of E' the condition

$$follow^-(H, last(H)) \cap first(H) = \emptyset$$

holds. E is in *star normal form* if and only if E' is in star normal form.

Theorem D *Given a marked expression E, we can compute in time linear in the size of E an expression $E^\bullet$ in star normal form such that E is unambiguous if and only if $E^\bullet$ is unambiguous. If E contains no & operator, then $E^\bullet$ does not contain one either; in this case, even* $\mathrm{L}(E) = \mathrm{L}(E^\bullet)$.

4 Conclusions

We have presented an optimal-time algorithm to test SGML content models for unambiguity. This paper clarifies the SGML standard and entails a proof of correctness for the unambiguity test in Clark's parser *sgmls*. Our work is currently reviewed by DIN and ISO for the revised edition of the SGML standard.

The "non-local" & operator, that can occur in content models but not in standard regular expressions, has required a new approach to unambiguity testing. It is also quite powerful from the languages point of view: Applying our characterization of unambiguous regular languages [BW92], it is not hard to see that the language L_0 of the unambiguous content model $(a\&b?\&c?)^*$ cannot be denoted by any unambiguous standard regular expression. Incidentally, L_0 can neither be denoted by any unambiguous content model in star normal form [Bru93]. While we have solved the semantic problem for unambiguous standard regular expressions in an earlier paper [BW92], it is an open problem to characterize the regular languages that can be denoted by unambiguous content models.

5 Acknowledgement

In January 1992, Derick Wood's and mine work on unambiguous standard regular expressions and its implementation by James Clark in his SGML parser *sgmls* kindled a discussion on unambiguous content models in the usenet newsgroup comp.text.sgml. The contributions of James Clark and Erik Naggum stimulated the work in this paper and are gratefully acknowledged.

References

[Bar89] D. Barron. Why use SGML? *Electronic Publishing—Origination, Dissemination and Design*, 2(1):3–24, April 1989.

[Bru92a] A. Brüggemann-Klein. Regular expressions into finite automata. In I. Simon, editor, *Latin '92*, pages 87–98, Springer-Verlag, Berlin, 1992. Lecture Notes in Computer Science 583.

[Bru92b] A. Brüggemann-Klein. Regular expressions into finite automata. 1992. To appear in Theoretical Computer Science.

[Bru93] A. Brüggemann-Klein. Formal models in document processing. Habilitationsschrift. Submitted to the Faculty of Mathematics at the University of Freiburg, 1993.

[BW92] A. Brüggemann-Klein and D. Wood. Deterministic regular languages. In A. Finkel and M. Jantzen, editors, *STACS 92*, pages 173–184, Springer-Verlag, Berlin, 1992. Lecture Notes in Computer Science 577.

[Cla92] J. Clark. 1992. Source code for SGMLS. Available by anonymous ftp from ftp.uu.net and sgml1.ex.ac.uk.

[Gol90] C. F. Goldfarb. *The SGML Handbook.* Clarendon Press, Oxford, 1990.

[ISO86] ISO 8879: Information processing—Text and office systems—Standard Generalized Markup Language (SGML). October 1986. International Organization for Standardization.

[Pin92] J.-E. Pin. Local languages and the Berry-Sethi algorithm. Unpublished Manuscript, 1992.