An early completion algorithm: Thue's 1914 paper on the transformation of symbol sequences

James F. Power
Department of Computer Science, National University of Ireland,
Maynooth, Co. Kildare, Ireland.
jpower@cs.nuim.ie

Preprint of a paper from *Computability in Europe 2014*
(LNCS 8493, pp. 343-346)

Abstract

References to Thue's 1914 paper on string transformation systems are based mainly on a small section of that work defining Thue systems. A closer study of the remaining parts of that paper highlight a number of important themes in the history of computing: the transition from algebra to formal language theory, the analysis of the "computational power" (in a pre-1936 sense) of rules, and the development of algorithms to generate rule-sets.

Of the many current models of computation, one of the oldest is the *Thue system*, first specified by Axel Thue 100 years ago [Thu14]. A Thue system is typically presentated as a sequence of string-pairs over some fixed alphabet:

$$\begin{array}{ccccc} A_1, & A_2, & A_3, & \ldots, & A_n \\ B_1, & B_2, & B_3, & \ldots, & B_n, \end{array}$$

Any other two strings P and Q over the same alphabet are said to be *similar* if it is possible to transform P into Q by replacing a substring matching some A_i with the corresponding string B_i (or vice versa, replacing some B_i with A_i). Two strings are said to be *equivalent* if we can form a finite sequence of strings, each similar to the former, taking us from P into Q.

Emil Post showed how this could be recast as a special form of one of his canonical systems and then to the decision problem for Turing machines

[Pos47]. At the time, it was important as one of the first undecidable problems outside of the original set from 1936, and Thue systems, with their close resemblance to unrestricted grammars, have since been established as one of the classical models of computation [RS97, BO93].

However, only the first two pages of Thue's paper are directly relevant to Post's proof, and the remainder of the paper seems to have been rarely explored. In what follows we review some of the remaining contributions of the paper, and to advocate its relevance for the history of computing.

Background to Thue's 1914 paper

Axel Thue (1863-1922) was a Norwegian mathematician who published a range of papers, 35 of which are collected in his *Selected Mathematical Papers* [NSST77]. Most of these relate to algebra and Diophantine approximations (he also worked in geometry and mechanics), and a recent conference was dedicated to his contributions in this area[1]. However, Axel Thue also published four papers directly relating to the theory of words and languages.

Two of these, published in 1906 and 1912, dealt with patterns in infinite strings [Thu06, Thu12] (Berstel provides a translation and discussion [Ber95]). They are known for being an early contribution to the field of combinatorics (though not the earliest [Mar04]) and, in particular, for the Thue-Morse sequence. This sequence can be specified by giving a morphism μ defining a mapping over strings (applied like the rules of an L-system) $\mu(a) = ab$, $\mu(b) = ba$. Thus, for example, starting with the string a we can produce the strings: a, ab, $abba$, $abbabaab$, $abbabaabbaababba$, ...

These strings have some interesting properties: in particular they are all *overlap-free*. Two strings have an overlap if they are of the form CU and UD, with the common substring U forming the overlap. A special case is where a string overlaps with itself, and a string is overlap-free if it does not contain any substring that overlaps with itself. Thue proves that the morphism μ preserves this property: it will always map overlap-free words to overlap-free words.

Thue's other two "language theory" papers from 1910 and 1914 discuss the more general problem of transformations [Thu10, Thu14]. Thue's 1910 paper deals with transformations between trees, and is thus a more direct predecessor of his 1914 paper. It been discussed by Steinby and Thomas [ST00].

[1] *Thue 150*, held in Bordeaux, France from Sept 30 - Oct 4, 2013

The importance of critical pairs

The 1914 paper, whose title translates roughly as *Problems concerning the transformation of symbol sequences according to given rules* specifically articulates the central problem in algorithmic terms:

> *Problem I:* For any arbitrary given sequences A and B, to find a method, where one can always calculate in a predictable number of operations, whether or not two arbitrary given symbol sequences are equivalent in respect of the sequences A and B.

Thue observes that this task of solving this problem is "extensive and of the utmost difficulty" and notes that he must settle for dealing with some special cases of the problem. Having posed the general problem in §II of his paper, Thue then presents an early example of a proof of (what we would now call) *termination* and *local confluence* for the special case where the rules are non-overlapping and non-increasing in size.

When reducing some string P, we must find some occurrence of A_i and replace it with B_i. A difficulty arises if there is an overlap: some substring CUD in P, such that A_i matches both CU and UD, and thus choosing one option will eliminate our ability to later choose the other. In the modern setting of term rewriting, CU and UD are known as a *critical pair*, and the problem has been well-studied in the literature [Buc87], starting at least from Newman [New42].

Thus, having studied overlap-free strings in his previous papers, Thue's focus in 1914 is the converse, and the overlap situation of strings CU and UD is the focus of study for most of the paper.

Completion in the context of a monoid presentation

Thue deals with the special case where a language is defined by specifying some identity string, R. This is not the usual case in language theory but is not an unusual approach when presenting an algebraic group. In Thue's case he is presenting a *monoid*: a set with an associative binary operator and an identity element (but no inverse function).

So, given a monoid, represented by specifying the identity string, the word problem here simply involves transforming some string P to some string Q by repeated insertions and deletions of R. Thue calls this relation "equivalence with respect to R", writes it as $P = Q$, and formulates:

> *Problem II:* Given an arbitrary sequence R, to find a method where one can always decide in a finite number of investigations whether or not two arbitrary given sequences are equivalent with respect to R.

As before, a difficulty arises when two overlapping instances of R occur as substrings of P. If we represent these as CU and UD as above, then we have $R \equiv CU \equiv UD$. But in this case $C = CR \equiv C(UD) \equiv (CU)D \equiv RD = D$. This tells us that $C = D$ (modulo R). The importance of this equation is that if we choose to delete either CU and UD from the string containing CUD we are left with either C or D, but adding the equation $C = D$ restores the confluence of our derivation.

Moreover, since both C and D are constructed from R by removing the common substring U they have the same length and contain the same symbols. In this case, as Thue notes, it is relatively easy to derive an algorithm for solving the word problem, and Thue describes one in §V of his paper.

Given this solution for the special case of Problem I, Thue now can outline his *completion algorithm* to solve Problem II:

1. Start with the given identity word R.

2. Form equations $C = D$ based on the remainder from the overlaps within R. For all these equations C and D will have the same length, and this will be less than the length of R.

3. Form a new set of identity strings $R', R'', \ldots$ by applying the equivalences from step 2 in R. These new identity strings will all have the same length as the original R.

4. Iterate steps 2 and 3 until we reach the fixed point. We know the process terminates, since each new identity string we create can only be a permutation of the original identity string (and there are only finitely many of these).

Thue's algorithm lacks typical features of modern completion algorithms such as the Knuth-Bendix algorithm: in particular there is no need for a complex unification process when we are dealing with concrete strings. However, it certainly contains many of the "basic features" of the algorithm as described by Buchberger [Buc87], and could be considered as an embryonic version of it.

An early computational flavour

Throughout Thue's paper he distinguishes between the case where two strings are equal (modulo some R), and when two strings are *provably* equal with respect to some given set of rules. He also investigates special cases where these two relations coincide, and where he can formulate (what we would now call) an algorithm to solve the word problem.

Thue's perspective is vital from a computational point of view and is neatly summarised by Matiyasevich and Sénizergues [MS05]:

> "put[ting] more attention to the process of transformation of words rather than to its result [...] is typical to computer science but has no counterpart in, say, algebra"

Thue was writing as a mathematician, and well before the identification of computer science as a discipline, but in his 1914 paper we can recognise much of the computational DNA that would allow algebra to evolve into language theory.

References

[Ber95] Jean Berstel. Axel Thue's papers on repetitions in words: a translation. Publications du LaCIM, Université du Québec à Montréal, 1995.

[BO93] Ronald V. Book and Friedrich Otto. *String-rewriting Systems.* Springer, 1993.

[Buc87] Bruno Buchberger. History and basic features of the critical-pair/completion procedure. *Journal of Symbolic Computation*, 3(1-2):3–38, 1987.

[Mar04] Solomon Marcus. Words and languages everywhere. In Carlos Martin-Vide, Victor Mitrana, and Gheorghe Paun, editors, *Formal languages and applications*, pages 11–53. Springer, 2004.

[MS05] Yuri Matiyasevicha and Géraud Sénizergue. Decision problems for semi-Thue systems with a few rules. *Theoretical Computer Science*, (330):145169, 2005.

[New42] M.H.A. Newman. On theories with a combinatorial definition of "equivalence". *Annals of Mathematics*, 43(2):223–243, 1942.

[NSST77] Trygve Nagell, Atle Selberg, Sigmund Selberg, and Knut Thalberg, editors. *Selected Mathematical Papers of Axel Thue.* Universitetsforlaget, Oslo, 1977.

[Pos47] Emil L. Post. Recursive unsolvability of a problem of Thue. *Journal of Symbolic Logic*, 12(1):1–11, March 1947.

[RS97] Grzegorz Rozenberg and Arto Salomaa, editors. *Handbook of Formal Languages.* Springer, 1997.

[ST00] M. Steinby and W. Thomas. Trees and term rewriting in 1910: On a paper by Axel Thue. *EATCS Bull.*, 72:256–269, 2000.

[Thu06] Axel Thue. Über unendliche Zeichenreihen. *Christiana Videnskabs-Selskabs Skrifter, I. Math.-naturv. Klasse*, 7, 1906.

[Thu10] Axel Thue. Die Lösung eines Spezialfalles eines generellen logischen Problems. *Christiana Videnskabs-Selskabs Skrifter, I. Math.-naturv. Klasse*, 8, 1910.

[Thu12] Axel Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Christiana Videnskabs-Selskabs Skrifter, I. Math.-naturv. Klasse*, 1, 1912.

[Thu14] Axel Thue. Probleme über Veränderungen von Zeichenreihen nach gegebenen Regeln. *Christiana Videnskabs-Selskabs Skrifter, I. Math.-naturv. Klasse*, 10, 1914.