Lexicographic decomposition of k-valued transducers*

Jacques Sakarovitch† Rodrigo de Souza‡§

Abstract

We give a new, and hopefully more easily understandable, structural proof of the decomposition of a k-valued transducer into k unambiguous functional ones, a result established by A. Weber in 1996. Our construction is based on a lexicographic ordering of computations of automata and on two coverings that can be build by means of this ordering. The complexity of the construction, measured as the number of states of the transducers involved in the decomposition, improves the original one by one exponential.

1 Introduction

Transducers are automata with outputs which realise relations between words, the so-called *rational relations*. Besides being a fundamental concept in the field of automata theory, finite transducers are a convenient model for computation in real world applications. For example, algorithms such as determinisation and minimisation of transducers have been used extensively in natural language processing [**?**].

In particular, the family of partial functions which can be realised by transducers, or *rational functions*, has received much attention due to its remarkable properties. To cite two of them, the equivalence of functional transducers is decidable, whereas in general the equivalence of transducers reduces to the Post Correspondence Problem, and every rational function is unambiguous. The former is indeed a corollary of the even more interesting result, proved by Schützenberger in 1975, that the functionality is a decidable property for transducers [**?**]. The unambiguity result has been obtained by Eilenberg [**?**], and independently by Schützenberger [**?**]; it can also be derived from the fundamental work of Elgot and Mezei [**?**]. In other words, it says that every rational function can be realised by a

*A preliminary version of this paper has been presented at the STACS 2008 conference under the title *On the decomposition of k-valued rational relations* [**?**].

†LTCI, ENST/CNRS, Paris (France), sakarovitch@enst.fr

‡DEINFO, Universidade Federal Rural de Pernambuco, Campus Universitário da UFRPE, Dois Irmãos, CEP 52171-900, Recife PE (Brazil), rsouza@deinfo.ufrpe.br

§A financial support of CAPES Foundation (Brazilian government) for doctoral studies is gratefully acknowledged by this author.

transducer where distinct successful computations have distinct labels. Moreover, Eilenberg and Schützenberger's arguments are effective and yield two procedures which turn a functional transducer into an unambiguous one.

Later, these properties have been generalised to the setting where the restriction "at most one output for every input word" is relaxed by putting that the cardinalities of the outputs are uniformly bounded. Such an upper bound is called the *valuedness* of the relation (or the transducer), and leads to the definition of the *k-valued* rational relations (for every positive integer k) which send every input word to a set containing at most k words. Of course, the rational functions are the 1-valued rational relations.

The earliest (to our knowledge) decidability result for the bounded valued relations is due to Gurari and Ibarra, who proved in 1983 that it is decidable in polynomial time whether a transducer is k-valued (for a given constant k) [**?**]. Next, Weber proved in 1989 that one can decide in polynomial time whether a transducer is bounded valued, that is, whether there exists such a finite upper bound for the valuedness [**?**]. The equivalence of k-valued transducers has been shown to be decidable by Culik and Karhumäki in 1986 [**?**] and also by Weber [**?**], the latter with a procedure of double exponential time complexity.

The unambiguity property does not hold verbatim for the bounded valued rational relations. Indeed, even some simple 2-valued relations are *inherently ambiguous* (see [**?**] for examples). But Weber proved in 1996 that every k-valued transducer can be decomposed into a sum of k unambiguous ones [**?**], as stated by Schützenberger twenty years earlier but without a complete proof [**?**].

It is noteworthy that the aforementioned results for functional transducers are now (if not in the original papers) established by means of constructions conducted on the transducers themselves [**?**, **?**]. On the other hand, the corresponding ones for k-valued transducers come, in some sense, "from outside" and, what is worse, from a different world for each of them. Gurari and Ibarra's proof for the decidability of the k-valuedness relies on a reduction to the emptiness problem for a class of multi-counter automata and Culik and Karhumäki's one for the decidability of the equivalence appears in the context of the solution of Ehrenfeucht's conjecture on HDT0L languages. Weber's proof of the decomposition (which we shall discuss more in detail below), in spite of being more oriented towards the structure of the involved transducers, is highly combinatorial and still somewhat detached from the transducers.

The present paper is part of a complete reworking and rewriting of the theory of k-valued rational relations and transducers which puts it in line with the theory of rational functions and makes it appear as a natural generalisation of the latter not only at the level of the results, but also *at the level of proofs*. Our approach for those results is based on constructions which depend directly on the structure of the automata. They give back the subject a full coherence and yield systematically better complexity bounds. For the decidability of the bounded valuedness and of

the equivalence of k-valued transducers, the corresponding new proofs we have designed have been presented in [?] and [?], respectively. Here we present a new proof for the decomposition theorem which we restate below as Theorem 1:

Theorem 1 (Weber [?]). *Every k-valued transducer $\mathcal{T}$ can be effectively decomposed into a sum of k (unambiguous) functional transducers.*

By "decomposed" we mean that the relation realised by $\mathcal{T}$ and the union of the functions realised by the k unambiguous transducers are the same.

The main improvement brought by our proof is an exponential gain in the size of the decomposition (measured as the number of states of the obtained transducers): in [?], this number is of double exponential order, whereas we obtain transducers of single exponential size. Another benefit gained from the deeper use we make of the structure of the automata is an enhancement of readability, that is, our decomposition seems to be more easily understandable.

Before explaining our proof, let us outline the one given in [?]. It rests heavily on an older decomposition presented by the same author in [?] and which yields an exponential number of functional transducers regardless of the valuedness of $\mathcal{T}$ (and as a by-product an upper bound for the valuedness). The construction of the latter is based on the definition of an intricate set S of tuples which describe some information on the successful computations of $\mathcal{T}$. Then, for each element of S, a functional transducer containing exactly the successful computations which fit the corresponding information is defined.

Next, this older decomposition is carried a bit further in order to obtain a decomposition of $\mathcal{T}$ into exponentially many *unambiguous* transducers, each one of double exponential size. The decomposition into k unambiguous transducers is then extracted from some "neighbourhood graphs" built on the preliminary decomposition. For every input word u, the vertices of these graphs are the successful computations reading u, the edges link computations whose *lag*[1] is bounded by some fixed value depending on the size of $\mathcal{T}$. Combinatorial arguments show that these graphs have at most k strongly connected components, and k unambiguous transducers can be extracted from them.

Our proof makes use two times of a common scheme: starting from an automaton $\mathcal{A}$, an "expansion" of $\mathcal{A}$ is built and yields a larger automaton $\mathcal{B}$, which "covers" $\mathcal{A}$; this means that the computations of $\mathcal{B}$ and those of $\mathcal{A}$ are in a $1-1$ correspondence and that they have, roughly speaking, the same "structure". In the larger $\mathcal{B}$, it is so to say easier to distinguish between the computations and the proof of the property aimed at by the construction amounts to an adequate choice within these computations.

Such expansions are what we call *covering of automata*. Coverings have been introduced by the first author in [?] and used to present a construction due to

[1] To be defined in the body of the paper.

Schützenberger — coined as the *Schützenberger covering* — to prove that rational functions are unambiguous. But coverings appear, explicitly or not, in many papers in the field of automata theory. This is the case of the construction due to H. Johnson to prove that the lexicographic selection of a deterministic rational relation is a deterministic rational function [**?**, **?**].

We build in our proof two coverings, one for transducers and another for automata with multiplicities in $\mathbb{N}$. Both are based on the same idea of putting an ordering on the transitions of the automata, and from it a lexicographic ordering on the computations (which are viewed as words on the alphabet of the transitions). This method can be seen as a conceptual generalisation of the aforementioned one of H. Johnson.

The first construction (Section **??**) is what we call the *lead or delay covering* of a (real time) transducer $\mathcal{T}$. It is parametrised by an integer N, and as such defines a family of coverings $\mathcal{U}_1, \mathcal{U}_2, \ldots$. In $\mathcal{U}_N$, the end state of every computation stores the "lead or delay" between the projected one in $\mathcal{T}$ and the smaller computations (according to the lexicographic ordering of the computations), provided that their "lag" is at most N. Then, by erasing the condition of being final of some states of $\mathcal{U}_N$, one can find a subtransducer $\mathcal{V}_N$ where distinct successful computations cannot have the same label and lag less than N. Based on these properties, we show that if $\mathcal{T}$ is k-valued, then, for some adequate value of N, the resulting subtransducer $\mathcal{V}_N$ is *input-k-ambiguous* (that is, every input word can be read in it by at most k successful computations). This is of course interesting in itself:

Theorem 2. *For every k-valued transducer, one can effectively construct an equivalent and input-k-ambiguous one.*

The second construction, explained in Section **??**, is what we call the *multi-skimming covering* of an $\mathbb{N}$-automaton. As before, it depends on a lexicographic ordering fixed on the computations of the $\mathbb{N}$-automaton (or rather on a split form of the automaton we define in Section **??**), and "counts", for every successful computation, the number of the smaller ones with the same input. It is noteworthy that this covering yields a proof for the following *multi-skimming theorem* for $\mathbb{N}$-rational series:

Theorem 3. *Let $\mathcal{A}$ be a finite $\mathbb{N}$-automaton with n states realising the series s. There exists an infinite $\mathbb{N}$-covering $\mathcal{B}$ of $\mathcal{A}$ such that for every integer $k > 0$, there exists a finite $\mathbb{N}$-quotient $\mathcal{B}^{(k)}$ of $\mathcal{B}$ which satisfies:*

i) $\mathcal{B}^{(k)}$ is an $\mathbb{N}$-covering of $\mathcal{A}$ with at most $n\,(k+1)^n$ states;

ii) for every i, $0 \leq i < k$, there exists an unambiguous subautomaton $\mathcal{B}^{(k,i)}$ of $\mathcal{B}^{(k)}$ which recognises the support of $s \dot{-} i$;

iii) there exists a subautomaton $\mathcal{D}^{(k)}$ of $\mathcal{B}^{(k)}$ whose behaviour is $s \dot{-} k$.