Theoretical Computer Science 531 (2014) 1–25

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Graph transformation for incremental natural language analysis ☆

Suna Bensch [a], Frank Drewes [a,*], Helmut Jürgensen [b], Brink van der Merwe [c]

[a] Department of Computing Science, Umeå University, Sweden
[b] Department of Computer Science, Western University, London, Canada
[c] Department of Computer Science, Stellenbosch University, South Africa

ARTICLE INFO

Article history:
Received 9 April 2013
Received in revised form 12 January 2014
Accepted 2 February 2014
Communicated by G. Rozenberg

Keywords:
Graph transformation
Hyperedge replacement
Natural language analysis
Reader
Millstream system

ABSTRACT

Millstream systems have been proposed as a non-hierarchical method for modelling natural language. Millstream configurations represent and connect multiple structural aspects of sentences. We present a method by which the Millstream configurations corresponding to a sentence are constructed. The *construction* is incremental, that is, it proceeds as the sentence is being read and is complete when the end of the sentence is reached. It is based on graph transformations and a lexicon which associates words with graph transformation rules that implement the incremental construction process. Our main result states that, for an effectively nonterminal-bounded reader $\mathcal{R}$ and a Millstream system *MS* based on monadic second-order logic, the correctness of $\mathcal{R}$ with respect to *MS* can be checked: it is decidable whether all graphs generated by $\mathcal{R}$ belong to the language of configurations specified by *MS*.

1. Introduction

Millstream systems simultaneously model several aspects of language structure in a parallel and co-ordinated way [3,4]. A Millstream configuration of a sentence represents the analysis of that sentence with respect to those aspects, including appropriate links between the analyses. As aspects to be considered, morphology, syntax and semantics come to mind immediately. However, other aspects can be modelled as well. An important point is that the separation of aspects can lead to simple models for each of them; the connections between the models, the links, are established by, hopefully, also simple conditions. While the formal notions developed in this paper as well as the results obtained are independent of the number and types of linguistic aspects considered, we illustrate them by rather small examples that cover only syntax and semantics – and even these in a very restricted way neglecting many linguistic details – because our aim is to convey the principles and the potential of our approach rather than to present a full-blown implementation of a system for linguistic analysis of sentences. Nevertheless, the implementation of such a system is one of the long-term goals of this research, and we hope that our presentation shows that such an implementation would be both desirable and possible.

Various psycholinguistic and cognitive neuroscience-based studies (see [34] for example) show that humans do not postpone the analysis of an utterance or sentence until it is complete; they rather start to process the sentence immediately when they have heard or read the first words or parts of words. Along these lines, we present results regarding the

☆ This article is a revised and extended version of [7].

* Corresponding author.
E-mail addresses: suna@cs.umu.se (S. Bensch), drewes@cs.umu.se (F. Drewes), hjj@csd.uwo.ca (H. Jürgensen), abvdm@cs.sun.ac.za (B. van der Merwe).

http://dx.doi.org/10.1016/j.tcs.2014.02.006
0304-3975/

incremental syntactic and semantic analysis of natural language sentences using Millstream systems as part of our ongoing work on this formalism for the description and analysis of language.

Incremental language processing is an intensively studied topic both in the context of compiler construction for programming languages and in the context of natural language parsing. Of the vast literature related to incremental parsing we mention only a few selected studies: Work on incremental LR-parsing for programming languages by Ghezzi and Mandrioli [17] and by Wagner and Graham [35]; studies of incremental parsing for natural languages using various grammar models and various computing paradigms by Beuck et al. [8], Costa et al. [10,9], Hassan et al. [22], Huang and Sagae [24], Lane and Henderson [28], Nivre [29] and Wu et al. [36]. In these and similar studies one constructs a structural representation of an utterance, a sentence, or a program by building partial structures as one progresses reading or hearing the input and by combining them or rejecting already constructed structures. The structural representation is intended to reflect all relevant aspects as described by a single formal grammar. In his 1960 paper *Grammar for the hearer* [23], Hockett discusses the natural understanding of spoken language and the implied constraints on parsing models. What Hockett calls a "hearer" would be called a "reader" in our setting.

We propose that various linguistic levels like phonology, morphology, syntax and semantics should be considered simultaneously and not successively. Hence, we base our work on Millstream systems [3,4], a generic mathematical framework for the description of natural language. These systems describe linguistic aspects such as syntax and semantics in parallel by separate modules and provide the possibility to express the relation between the aspects by so-called interfaces. Roughly speaking, a Millstream system[1] consists of a finite number of *modules* each of which describes a linguistic aspect and an *interface* which describes the dependencies between these aspects. The modules need not be of the same mathematical nature: one aspect might be adequately modelled by a context-free grammar while, for another aspect, a Montague grammar might be preferable. Each module defines a tree language which describes one linguistic aspect in isolation. The interface establishes links between the trees given by the modules, thus turning unrelated trees into a meaningful whole called a *configuration* and filtering out analyses which make sense with respect to some linguistic aspects, but not all of the ones modelled.

In contrast, if one were to use a single type of grammar to model all aspects simultaneously, the resulting construct would be unmanageable, as is well known and can be seen in the admirable attempt of [27] to model German.

Consider – for simplicity – a Millstream system containing only two modules, a syntactic and a semantic one, which model the syntax and semantics of a natural language. A configuration of the Millstream system consisting of two trees with links between them represents an analysis of a sentence. An obvious question is how to construct such a configuration from a given sentence. Such a procedure would be a step towards automatic language understanding based on Millstream systems. This paper continues the work begun in [6], where we proposed to use graph transformation for that purpose. We mimic the incremental language processing performed by humans to construct a Millstream configuration by a step-by-step procedure while reading the words of a sentence from left to right.[2] The idea is that the overall structure of a sentence is built incrementally, word by word. With each word, one or more lexicon entries are associated. These lexicon entries are graph transformation rules whose purpose it is to construct an appropriate configuration.

For a sentence like *Mary loves Peter*, for example, we first apply a lexicon entry corresponding to *Mary*. This results in a partial configuration representing the syntactic, semantic and interface structure of the word. We continue by applying a lexicon entry for *loves*, which integrates the syntactic, semantic and interface structure of this word into the configuration. Thus, after the second step, we have obtained a partial configuration representing *Mary loves*. Finally, the structure representing *Peter* is integrated into the configuration, resulting in the Millstream configuration for the entire sentence.

We call such a sequence of graph transformation steps a *reading* of the sentence. The graph transformation system itself, which consists mainly of the lexicon, is called a *reader*. Since words can appear in different contexts, alternative lexicon entries for one and the same word may co-exist. In general, this may result in nondeterminism or even ambiguity; the former occurs when two or more rules are applicable, but only one will finally lead to a complete reading; the latter arises when two or more readings of the sentence are possible. These effects are inevitable as they are caused by inherent properties of natural language. In many situations, however, only one lexicon entry will be applicable because its left-hand side requires certain partial structures to be present, to which the new part is added. This corresponds to the situation of a human reader who has already seen part of the sentence and can thus rule out certain lexicon entries associated with the next word.

Given a reader that is supposed to construct configurations of a Millstream system *MS*, an obvious question to ask is whether the reader yields correct configurations, that is, whether the configurations it constructs are indeed configurations of *MS*. The main (formal) result of this paper is Corollary 1 which states that, under certain conditions, this question is decidable for so-called regular MSO Millstream systems, that is, systems in which the modules are regular tree grammars (or, equivalently, finite tree automata) and the interface conditions are expressed in monadic second-order (MSO) logic. In other words, given a regular MSO Millstream system *MS* and a reader satisfying the conditions mentioned, one can determine effectively whether all readings yield correct configurations of *MS*.

[1] The term *Millstream system* refers to the place at which the notion was created; thus, it has no direct connection to language theory.

[2] Instead of "left to right" one might prefer to say "in their spoken (or natural) order."