Annals of Pure and Applied Logic 164 (2013) 1079–1100

Contents lists available at SciVerse ScienceDirect

Annals of Pure and Applied Logic

www.elsevier.com/locate/apal

Lambek vs. Lambek: Functorial vector space semantics and string diagrams for Lambek calculus

Bob Coecke[a], Edward Grefenstette[a], Mehrnoosh Sadrzadeh[b,*]

[a] *Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, United Kingdom*
[b] *School of Electronics and Computer Science, Queen Mary University of London, Mile End Road, London, E1 4NS, United Kingdom*

ARTICLE INFO

Article history:
Received 12 December 2011
Accepted 13 March 2013
Available online 17 June 2013

MSC:
03
18
68
81

Keywords:
Categorial grammars
Vector spaces
Monoidal categories
Lambek calculus
Pregroups
Distributional semantics

ABSTRACT

The Distributional Compositional Categorical (DisCoCat) model is a mathematical framework that provides compositional semantics for meanings of natural language sentences. It consists of a computational procedure for constructing meanings of sentences, given their grammatical structure in terms of compositional type-logic, and given the empirically derived meanings of their words. For the particular case that the meaning of words is modelled within a distributional vector space model, its experimental predictions, derived from real large scale data, have outperformed other empirically validated methods that could build vectors for a full sentence. This success can be attributed to a conceptually motivated mathematical underpinning, something which the other methods lack, by integrating *qualitative* compositional type-logic and *quantitative* modelling of meaning within a category-theoretic mathematical framework. The type-logic used in the DisCoCat model is Lambek's pregroup grammar. Pregroup types form a posetal compact closed category, which can be passed, in a functorial manner, on to the compact closed structure of vector spaces, linear maps and tensor product. The diagrammatic versions of the equational reasoning in compact closed categories can be interpreted as the *flow of word meanings* within sentences. Pregroups simplify Lambek's previous type-logic, the Lambek calculus. The latter and its extensions have been extensively used to formalise and reason about various linguistic phenomena. Hence, the apparent reliance of the DisCoCat on pregroups has been seen as a shortcoming. This paper addresses this concern, by pointing out that one may as well realise a functorial passage from the original type-logic of Lambek, a monoidal bi-closed category, to vector spaces, or to any other model of meaning organised within a monoidal bi-closed category. The corresponding string diagram calculus, due to Baez and Stay, now depicts the flow of word meanings, and also reflects the structure of the parse trees of the Lambek calculus.

1. Introduction

Language is both empirical and compositional: we learn meanings of words by being exposed to linguistic practice, and we form sentences we've never heard before by composing words according to the rules of grammar. Various mathematical and formal models have sought to capture facets and aspects of language learning and formation. Compositional type-logical approaches [30] represent sentence formation rules based on formal syntactic analysis, using formalisms such as context free grammars [10,41], Lambek calculus [30], or Combinatorial Categorial Grammar [56]. Such formal approaches to grammar

[*] Corresponding author.
E-mail addresses: bob.coecke@cs.ox.ac.uk (B. Coecke), edward.grefenstette@cs.ox.ac.uk (E. Grefenstette), mehrnoosh.sadrzadeh@eecs.qmul.ac.uk (M. Sadrzadeh).

0168-0072/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.apal.2013.05.009

align well with Frege's notion of compositionality, according to which the meaning of a sentence is a function of the meaning of its parts [19], but eschew the empirical nature of language, requiring pre-defined mathematical structures, domains and valuations to make sense.

Orthogonal to formal logical models, empirical approaches to semantics construct representations of individual words based on the contexts in which they are used. These models are often referred to as *distributional* or geometric models of semantics and are sometimes considered to be in line with the "meaning is use" view of Wittgenstein's philosophy of language [59]. Distributional models have been applied successfully to tasks such as thesaurus extraction [20,15], automated essay marking [36], and other semantically motivated natural language processing tasks. While these models reflect the empirical aspects of language learning that type-logical models lack, they in turn lack composition operations which would allow us to learn meanings of phrases based on the meanings of their parts. Developing models that could combine the strengths of the above two approaches has proved to be a challenge for computational linguistics and its applications to natural language processing (NLP).

The distributional compositional categorical (DisCoCat) model of meaning, developed in [13,14], provides a solution to the above problem. This framework, which realised the challenge proposed in [11] enables a combination of the type-logical and distributional models of meaning and resulted in a procedure for compositionally computing meaning vectors for sentences by exploiting the grammatical structure of sentences and the meaning vectors of the words therein. The framework was inspired by the category-theoretic high-level framework for modelling quantum protocols [2], were the corresponding string diagram calculus exposes flows of information between the systems involved in multi-system protocols such as quantum teleportation [12]. The DisCoCat model has meanwhile been experimentally validated for natural language tasks such as word-sense disambiguation within phrases [21,22].

The DisCoCat model relies on Lambek pregroups [32] as its base type-logic. In category-theoretic terms, these have a (non-symmetric) compact closed structure when considering types as objects and type reductions as morphisms. The DisCoCat exploits the fact that finite dimensional vector spaces can also be organised within a compact closed category. The first and mainly technical goal of this paper is to stress that the choice of a compact or monoidal type-logic is not crucial to the applicability of the procedure. To achieve this goal, we have tweaked the distributional compositional model of previous work from Lambek pregroups to Lambek monoids, hence developing a vector space model for the meaning of natural language sentences parsed within the Lambek calculus. In this paper, we develop a similar homographic passage via a functor from a monoidal bi-closed category of grammatical types and reductions to the symmetric monoidal closed category of finite dimensional vector spaces.

This functorial passage is another contribution of this paper and gives rise to an interesting analogy with Topological Quantum Field Theory (TQFT) [4,5,28]. A TQFT is also a monoidal functor from the category of cobordisms into the category of vector spaces and linear maps. From the perspective of TQFT, our DisCoCat models form a 'Grammatical Quantum Field Theory' obtained by replacing the monoidal category of cobordisms in a TQFT by a certain partially ordered monoid which accounts for grammatical structure. This analogy of the compositional distributional model of meaning with TQFT was first pointed out by Louis Crane at a workshop in Oxford, August 2008. Similar to the original model-theoretic framework of meaning by Montague [41], this semantic framework is obtained via a homomorphic passage from sentence formation rules to compositions of meanings of words. However, contrary to the Montague's model, meanings of words and sentences are expressed in terms of vectors and vector compositions rather than in terms of sets and set-theoretic operations.

As a result of the compactness of Lambek pregroups, the mechanism of how meanings of words interact to produce meanings of sentences has a purely diagrammatic form, which admits an intuitive interpretation in terms of *information flow*. By *information flow* we mean the topology of the two-dimensional graphical representation of the operations that produces the meaning of sentences from the meaning of words. Mathematically, these are expressed in the graphical language of the particular category in which we model the meaning of words and sentences [54], a practice tracing back to Penrose's work in the early 1970s [47], that was turned into a formal discipline by Joyal and Street in the 1990s [25]. These diagrams, for the particular case of compact closed categories, were extensively exploited in the earlier DisCoCat models. Here we show how the clasp-string calculus of Baez and Stay [6] can be used to provide diagrams for information flows that arise in the Lambek monoids, which are not compact. Our ambition is to use this work as a starting point for providing vector space meaning for more expressive natural language sentences such as those parsed with Combinatorial Categorial Grammars (CCGs) or Lambek–Grishin calculus [44,8]. The expressive powers of these grammars go beyond that of Lambek grammars, which are context free.

Finally, drawing a connection with games seems appropriate in the context of this special issue. Application of games to interpreting and formalising natural language traces back to the 'dialogical logic' of Lorenz and Lorenzen [39] who used the dialogue analogy to develop a game semantic model for formulae of intuitionistic logic. Later, a classical logic version of the theory with a model theoretic focus was developed by Hintikka and Sandu [23]. A proof-theoretic approach led to the use of linear logic, and proof nets, e.g. see [29,37]. Independently, another line of research was pursued by linguists who also used the term 'dialogue games' to provide a semantic model for real-life human–computer dialogues. One of the original proposals of this line was based on Grice's pragmatic philosophy of language and used component programs and specifications to model dialogues and queries; the setting was applied to online sale tools [38]. Later on, a formal model based on belief revision and Bayesian update was developed for this approach [52]. Our work can be seen as bridging these two (abstract logical and applied linguistic) communities. Our starting point is a Lambek calculus, with a proof theory similar to that of intuitionistic multiplicative linear logic. In this calculus, the grammatical structure of a sentence is represented as