Chris Hayduk
Math A4900
Research Paper Summaries
12/15/2020

## 1. Introduction

The main paper in this research summary, *Lambek vs. Lambek: functorial vector space semantics and string diagrams for Lambek calculus*, was published in the Annals of Pure and Applied Logic in June 2013 and was written by Bob Coecke, Edward Grefenstette, and Mehrnoosh Sadrazdeh [1]. Bob Coeke is a currently theoretical physicist at the University of Oxford. He received his doctorate degree from Vrije Universiteit Brussel in 1996 and went on to perform postdoctoral work at Imperial College, London and McGill University in Montreal. During the time that this paper was written, Bob Coeke was a professor of Quantum Foundations, Logic, and structures at Oxford. He stills holds this position today. His research interests include category theory, logic, and diagrammatic reasoning [4].

Edward Grefenstette is currently a research scientist at Facebook AI and an honorary Associate Professor at University College London (UCL). He received his PhD in Computer Science from the University of Oxford in 2013. Following his PhD, he worked as a postdoctoral research assistant in the Computer Science department at Oxford. Grefenstette then moved to DeepMind in London before finally joining Facebook AI. His research primarily focuses on applying category theory and deep learning to natural language processing problems [5].

Mehrnoosh Sadrazdeh is currently an Associate Professor in the Programming Principles, Logic, and Verification Group at University College London. From 2008 to 2013, she was a Postdoctoral Fellow and a Research Fellow in the Computer Science Department of the University of Oxford. Following this, she became a lecturer in the School Electronic Engineering and Computer Science at Queen Mary University of London from 2013 to 2019. Her research interests include natural language processing (NLP) and high-level logical models for computer systems [6].

Their paper, *Lambek vs. Lambek: functorial vector space semantics and string diagrams for Lambek calculus*, is an application of categorial grammar, which is a view of a natural language syntax which states that the components of a sentence behave as functions [7]. The authors of the paper make use of Lambek's pregroup grammar in order to model the type-logic of the sentences. Lambek's pregroup grammar is a simplification of Lambek calculus, which has been used in various contexts in mathematical and computational linguistics. This pregroup grammar provides a computational procedure when can be used, in conjunction with the definitions of the words, to construct a meaning for any given sentence. More popular approaches in natural language processing currently purely use word meanings based on context in order to construct the meanings of sentences. These models are known as empirical models of semantics, and they have been very successful in quite a few applications, such as thesaurus extraction. In this paper, the authors create the distributional compositional categorical (DisCoCat) model of meaning, which combines the logical framework of Lambek's pregroup grammar with the contextual word definitions of empirical models in order to improve performance in specific natural language processing tasks, such as word-sense disambiguation [1].

## 2. Summary

In this paper, the authors construct a form of categorial grammar using Lambek's pregroup grammar. They begin with a description of Lambek calculus. It consists of a non-commutative binary operation on partially ordered sets. The authors then go on to define a residuated monoid as follows,

**Definition 2.1** [1, §2.1] A residuated monoid is a partially ordered set $(L, \leqslant)$ equipped with a monoid structure $(L, \cdot, 1)$ that preserves the partial order, that is for all $a, b, c \in L$, we have,

$$\text{If } a \leqslant b \text{ then } a \cdot c \leqslant b \cdot c \text{ and } c \cdot a \leqslant c \cdot b$$

Note that a monoid has a single associative binary operation and an identity element [8]. A partially ordered residuated monoid is called a Lambek monoid in the context of this paper [1, §2.1]. In addition, we have the following definitions for a natural language and a Lambek grammar,

**Definition 2.2** [1] For $\Sigma$ the set of words of a natural language and $\mathcal{B}$ a set of basic grammatical types, a Lambek type-dictionary is binary relation $D$ defined as

$$D \subset \Sigma \times T(\mathcal{B})$$

where $T(B)$ is the free Lambek monoid generated over $\mathcal{B}$

**Definition 2.3** [1, §2.1] A Lambek grammar $G$ is a pair $\langle D, S \rangle$, where $D$ is a Lambek type-dictionary and $S \subset \mathcal{B}$ is a set of designated types, containing types such as that of a declarative sentence $s$ and a question $q$.

Thus, as we can see, a Lambek-type dictionary assigns each word in the language a structure from the free Lambek monoid generated over $\mathcal{B}$. The Lambek grammar then assigns another type to each ordered pair in $D$.

Lambek later developed a simplification for his Lambek grammar in which he used a pregroup structure. The definition is as follows,

**Definition 2.5** [1, §2.1.1] A Lambek pregroup is a partially ordered unital monoid where each element has a left and a right adjoint $(P, \leqslant, \cdot, 1, (-)^{\ell}, (-)^{r})$. That is, for every $p \in P$, there is a $p^{r}$ and a $p^{\ell}$ in $P$, which satisfy the following four inequalities:

$$p \cdot p^{r} \leqslant 1 \leqslant p^{r} \cdot p$$
$$p^{\ell} \cdot p \leqslant 1 \leqslant p \cdot p^{\ell}$$

The authors gave the following example for a Lambek pregroup grammar,

**Table 2**
Type assignments for the toy language $\Sigma$ in a Lambek pregroup.

| men | dogs | cute | kill | to kill | do | not |
|---|---|---|---|---|---|---|
| $n$ | $n$ | $n \cdot n^{l}$ | $n^{r} \cdot s \cdot n^{l}$ | $\sigma^{r} \cdot j \cdot n^{l}$ | $n^{r} \cdot s \cdot j^{l} \cdot \sigma$ | $\sigma^{r} \cdot j \cdot j^{l} \cdot \sigma$ |

Each word in the language is assigned a specific grammatical role, represented by some combination of elements in the Lambek pregroup. The meaning of a sentence can then be determined by composing these elements together as functions in the order in which they appear in the sentence [1, §2.1.1].

The authors then went on to describe distributional models of meaning, which model words based on their meaning in context. The language is depicted as a vector space, and usually several important words are selected in order to form a basis for the vector space in which each word vector

lives. The angle between word vectors is used as a proxy for the closeness in meaning between those two words, as the angle is smaller for words that appear in similar context. For example, the word vectors "cat" and "dog" would have a smaller angle than the word vectors "cat" and "democracy" because "cat" and "dog" appear in very similar contexts [1, §3]. The authors give the following image on p. 7 of the paper in order to graphically explain this concept,

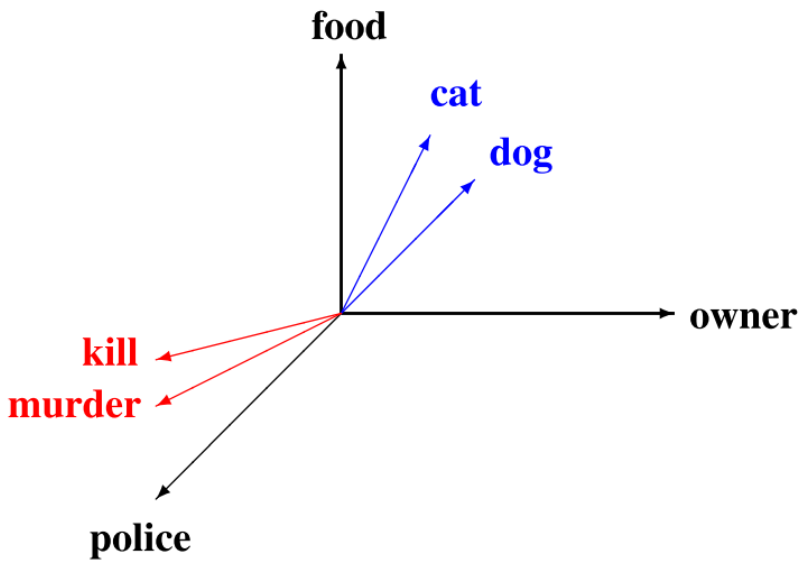


**Fig. 1.** A subspace of a real vector space model of meaning.

These models ignore the grammatical structure of the language, unlike the Lambek pregroup grammar discussed previously [1, §3]. There is also an issue with sentences of differing lengths. For example, a distributional model of meaning cannot compare the sentences "I walk the dog" and "I walk" because they live in different spaces [1, §3]. The authors discuss that a way to resolve these issues is to combine the pregroup grammar with a distributional model. This allows the model to produce vectors for the meanings of sentences, taking into account both the grammatical structure of the sentence and the word vectors contained within it [1, §4]. The paper proposes the following definition for their DisCoCat model,

**Definition 4.1** [1, §4.1] The distributional compositional categorical (DisCoCat) model of meaning is a category $FVect \times Preg$, equipped with a tensor product given by pointwise tensor of $Preg$ and $FVect$, that is $(V, p) \otimes (W, q) = (V \otimes W, p \cdot q)$, whose unit is $(\mathbb{R}, 1)$.

where $FVect$ is the finite dimensional vector space of word meanings and $Preg$ is he pregroup grammar. Both the pregroup grammar and the finite dimensional vector space form compact closed categories [1, §4.1]. The authors evaluated the DisCoCat model on sentence data sets, performing tasks such as "disambiguation", which when the model was distinguish between multiple possible meanings for a word within a sentence [1, §5.2]. DisCoCat performed very well in this setting, showing the promise of this approach to natural language processing [1, §5.2].

## 3. Reflection

This paper adds onto the current literature in categorial grammar by extending the DisCoCat model from pregroups to Lambek monoids [1]. This allows the model to more closely relate to semantic models of language [1]. In addition, the method of constructing the DisCoCat model can be extended to other domains, and in particular has been used in Topological Quantum Field Theory. The authors describe their work here as a "grammatical quantum field theory for Lambek monoids" [1]. I personally had a strong interest in this paper due to my interest in language learning and linguistics. I find the diverse structures of natural language and the rules that govern its grammar to be very interesting, and I believe mathematical linguistics is important in analyzing the general rules that are common among natural language grammars. Moreover, I think Natural Language Processing is one of the most exciting fields in Machine Learning, so seeing the connections between algebra and category theory with more traditional NLP approaches was great.

The authors of this paper made use of a significant number of definitions and images in order to explain the background of this topic. I found this to be very helpful in building up some understanding for and intuition of the objects that were being discussed. Unfortunately there was still a large amount of background knowledge that I was lacking which made parts of the paper difficult to parse. However, overall, I think the writing style of the paper and clarity of exposition made this paper easier to digest than many other papers that I have seen.

## Appendix A. Follow-up Papers

In the paper, *A generalised quantifier theory of natural language in categorical compositional distributional semantics with bialgebras*, the authors expand on the work in categorical compositional distributional semantics which was discussed in *Lambek vs. Lambek*. The authors in this paper formalize and generalize this approach to natural language modeling [2]. The authors of this paper include Mehrnoosh Sadrzadeh, who was an author on the *Lambek vs. Lambek* paper, as well as Jules Hedges. Hedges received his PhD from the University of Oxford and, at the time of the paper's publication, was a postdoctoral researcher at the Max Planck Institute for Mathematics in the Sciences [9].

In the paper, *Generalized Relations in Linguistics and Cognition*, the authors expand the categorical compositional models of natural language which were discussed in *Lambek vs. Lambek* to conceptual space models of cognition [3]. This builds on what was stated in *Lambek vs. Lambek*, that the categorical compositional model described a generalized approach which could be applied to other domains [1]. The authors of this paper included Bob Coecke, who was included on the *Lambek vs. Lambek* paper. In addition, the authors included Fabrizio Genovese, who was a PhD student at the University of Oxford at the time of the paper's publication [10]. We also have Martha Lewis, a research assistant at Oxfrod with research interests in modeling how humans represent and use concepts in their minds [11]. Lastly, Dan Marsden is a research assistant at Oxford whose research interests include category theoretic techniques to model systems in areas such as quantum computation and natural language semantics [12].

## References


[1] Coecke, Bob; Grefenstette, Edward; Sadrzadeh, Mehrnoosh. *Lambek vs. Lambek: functorial vector space semantics and string diagrams for Lambek calculus*. Ann. Pure Appl. Logic 164 (2013), no. 11, 1079–1100.

[2] Hedges, Jules; Sadrzadeh, Mehrnoosh. *A generalised quantifier theory of natural language in categorical compositional distributional semantics with bialgebras*. Math. Structures Comput. Sci. 29 (2019), no. 6, 783–809.

[3] Coecke, Bob; Genovese, Fabrizio; Lewis, Martha; Marsden, Dan. *Generalized Relations in Linguistics and Cognition*. Logic, Language, Information, and Computation (2017), pp. 256–270.

[4] "Bob Coecke." Wikipedia, Wikimedia Foundation, 18 Oct. 2020, `en.wikipedia.org/wiki/Bob_Coecke`.

[5] "Edward Grefenstette." 18 Oct. 2020, `https://www.egrefen.com/`

[6] "Mehrnoosh Sadrzadeh." 18 Oct. 2020, `https://msadrzadeh.com/`

[7] "Categorial Grammar." Wikipedia, Wikimedia Foundation, 30 Aug. 2020, `en.wikipedia.org/wiki/Categorial_grammar`.

[8] "Monoid." Wikipedia, Wikimedia Foundation, 18 Oct. 2020, `en.wikipedia.org/wiki/Monoid`.

[9] "CV." *Jules Hedges*, 21 Aug. 2020, `julesh.com/cv/`.

[10] "Fabrizio Romano Genovese." 19 Oct. 2020, `www.cs.ox.ac.uk/people/fabrizio.genovese/`.

[11] "Martha Lewis." 19 Oct. 2020, `https://www.cs.ox.ac.uk/people/martha.lewis/`.

[12] "Dan Marsden." 19 Oct. 2020, `https://www.cs.ox.ac.uk/people/dan.marsden/`.