Enhanced Expressivity for Compositional Distributional Semantics

Gijs Jasper Wijnholds

Institute for Logic, Language and Computation

Introduction

Categorical compositional distributional models of meaning (**CCDMM**) provide a general way of attaching compositional interpretation to distributional semantics. Such models have been developed using either pregroup grammars (**PG**) or Lambek's Syntactic Calculus (**L**) as a syntactic backbone to guide semantic analysis, via a functorial passage to finite dimensional vector spaces (**FVect**).

The typelogical formalisms employed so far (**PG**, **L**) are lacking in two respects: (a) they identify too much structures by admitting associativity, effectively *flattening* constituent structure and (b) they are not capable of recognizing structures beyond context-free. Thus the crucial goal is

to improve the basic CCDMM framework by considering more fine-grained type logics.

We have focused on the Lambek-Grishin Calculus (**LG**): it recognises **local** constituents through a *non-associative* composition operation while facilitating **non-local** information flow. To replicate the architecture of the **CCDMM** framework, we have addressed the following objectives:

Objectives

- To give a categorical characterization of the Lambek-Grishin Calculus,
- Develop a graphical language to reason about morphism structure and equality in terms of *string diagrams*,
- Show that finite dimensional vector spaces are categorically similar to **LG** so that the latter becomes *interpretable*,
- Empirically validate the extended model (future research)

LG Base Logic

$$\frac{}{1_A : A \to A}\,Ax \qquad \frac{f : A \to B \quad g : B \to C}{g \circ f : A \to C}\,T$$

$$\frac{f : A \otimes B \to C}{\triangleright f : A \to C/B}\,R1 \qquad \frac{f : A \otimes B \to C}{\triangleleft f : B \to A\backslash C}\,R2$$

$$\frac{g : A \to C/B}{\triangleright^{-1} g : A \otimes B \to C}\,R1' \qquad \frac{g : B \to A\backslash C}{\triangleleft^{-1} g : A \otimes B \to C}\,R2'$$

$$\frac{f : C \to B \oplus A}{\blacktriangleleft f : B \obslash C \to A}\,CR1 \qquad \frac{f : C \to B \oplus A}{\blacktriangleright f : C \oslash A \to B}\,CR2$$

$$\frac{g : B \obslash C \to A}{\blacktriangleleft^{-1} g : C \to B \oplus A}\,CR1' \qquad \frac{g : C \oslash A \to B}{\blacktriangleright^{-1} g : C \to B \oplus A}\,CR2'$$

$\mathbf{LG^D}$: Linear Distributivities

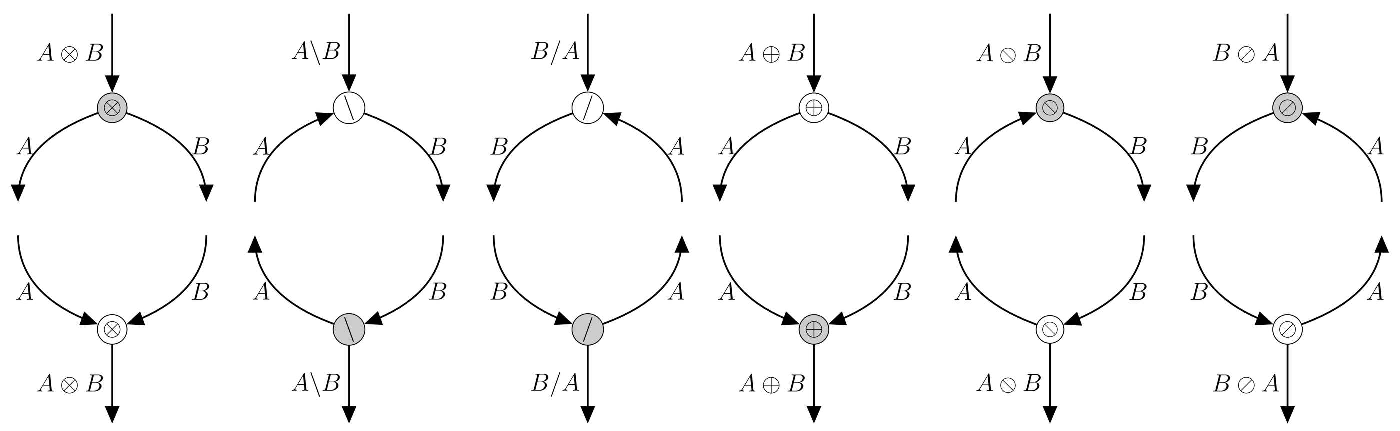
$$\delta^{\obslash} : (A \obslash B) \otimes C \to A \obslash (B \otimes C)$$
$$\delta^{\oslash} : C \otimes (B \oslash A) \to (C \otimes B) \oslash A$$
$$\kappa^{\obslash} : C \otimes (A \obslash B) \to A \obslash (C \otimes B)$$
$$\kappa^{\oslash} : (B \oslash A) \otimes C \to (B \otimes C) \oslash A$$

LG, Categorically

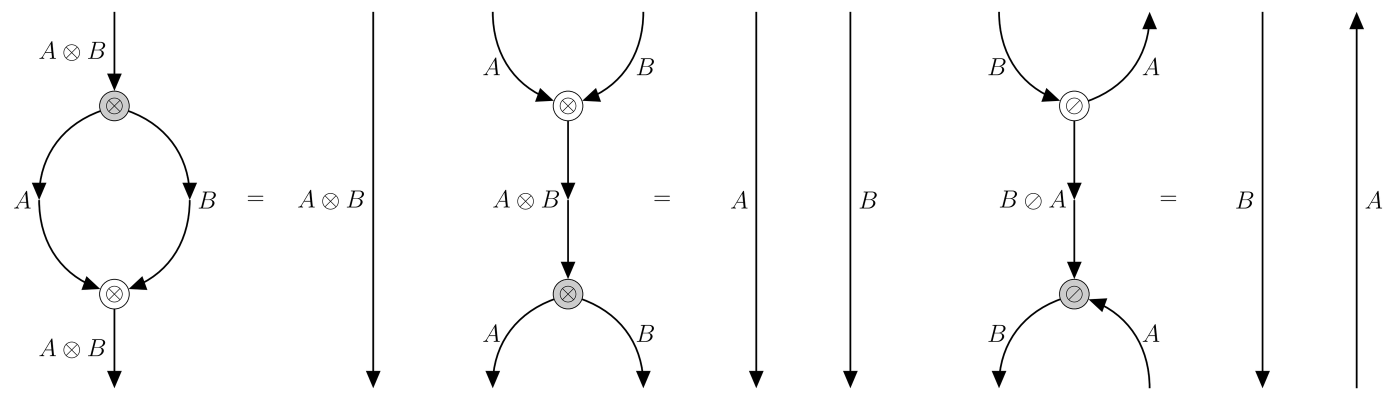
Because **LG** is non-associative and non-unital, it is a tensor category. The Lambek connectives give rise to a biclosed structure, whereas the Grishin connectives give a biopen structure. Thus, after adding interactivity postulates, we have a linearly distributive biclopen tensor category.

PN Defined

The proof net calculus is defined by **links** that can be combined to construct bigger graphs:

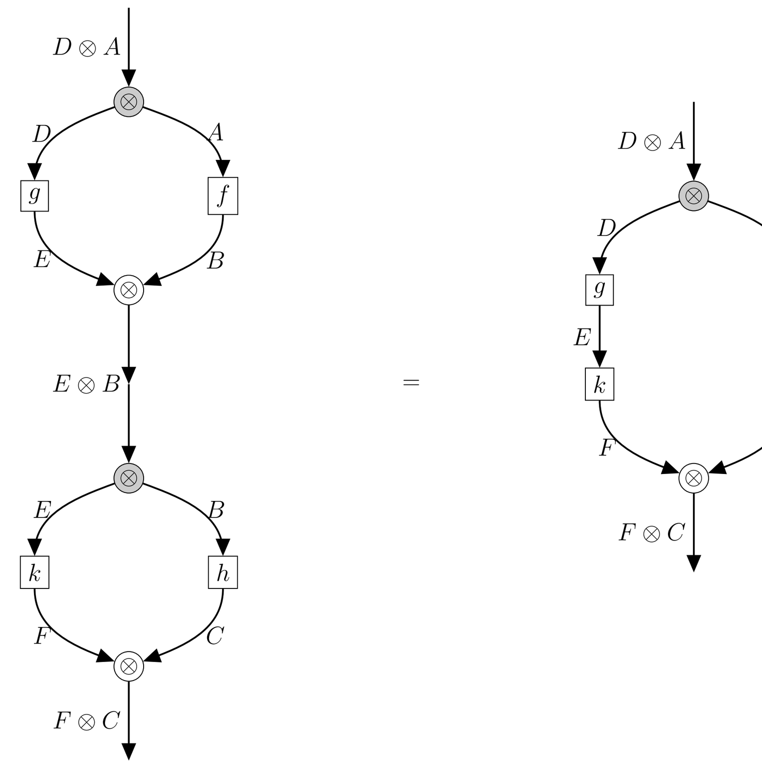


The subclass of proof nets is then defined by appropriate **correction criteria**. Finally, coherence of the proof net calculus is ensured by defining **graphical equations** of the following sort:



Illustration

Bifunctoriality of $\otimes$, i.e. $(k \otimes h) \circ (g \otimes f) = (k \circ g) \otimes (h \circ f)$ is valid by the following equation:

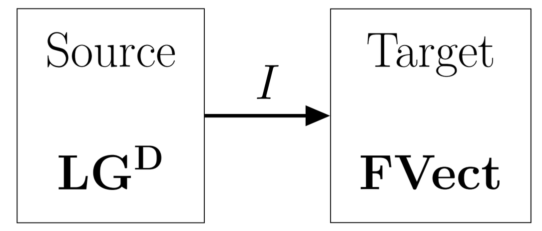


PN, Categorically

Coherence of **PN** is obtained by a freeness theorem, i.e. given an appropriate signature Σ, $\mathbf{PN}(\Sigma)$ is the **free** linear distributive biclopen category over Σ.

Final Model

The final model follows the same pattern as the basic model, thus the functorial transition is depicted as follows:



The non-associativity of the source logic now **blocks** identification of otherwise equivalent derivations while the distributivities **allow** linear restructuring of resources.

Conclusion

The **CCDMM** architecture has been fully developed for the Lambek-Grishin Calculus. More specifically, we have developed novel categorical concepts corresponding to the logic of **LG** and have shown that **FVect** is an instance of this new concept. Moreover we have developed a coherent graphical language for linear distributive biclopen tensor categories. As associativity is now not licensed by the syntactic model, we are able to better distinguish constituent structures even though **FVect** by itself would not accomodate this. This shows that **LG** is a fine-grained grammar logic especially suitable for setting up a **CCDMM** framework.

Future Research

- Empirical validation: replicating experiments of the basic model and/or adding grammar induction
- Incorporating Frobenius structure and/or mixing states
- Richer semantic models: examining target interpretations that preserve more of the source logic structure

[1] Gijs Wijnholds.
Categorical Foundations for Extended Compositional Distributional Models of Meaning.
Master's thesis, Universiteit van Amsterdam, 2014.

[2] Michael Moortgat.
Symmetric Categorial Grammar: Residuation and Galois Connections.
arXiv preprint arXiv:1008.0170, 2010.

[3] Bob Coecke, Edward Grefenstette, and Mehrnoosh Sadrzadeh.
Lambek vs. Lambek: Functorial Vector Space Semantics and String Diagrams for Lambek Calculus.
Annals of Pure and Applied Logic, 164(11):1079–1100, 2013.

[4] Mehrnoosh Sadrzadeh, Stephen Clark, and Bob Coecke.
The Frobenius Anatomy of Word Meanings I: Subject and Object Relative Pronouns.
Journal of Logic and Computation, pages 44–69, 2013.

Acknowledgements

The author is grateful for funding kindly provided by the Institute for Logic, Language and Computation:

Institute for Logic, Language and Computation
www.illc.uva.nl

Contact Information

- Web: gijswijnholds.github.io
- Email: gijswijnholds@gmail.com