Syntactic Control for Compositional Vector Space Models

Michael Moortgat[1] and Gijs Wijnholds[2]

[1] Utrecht Institute of Linguistics
M.J.Moortgat@uu.nl
[2] ILLC Amsterdam
gijswijnholds@gmail.com

Abstract. The framework of Compositional Distributional Semantics unifies vector space models for lexical meanings with a compositional account of how these meanings combine into phrases and larger units. The syntactic engines that have been used to drive the interpretation process (Lambek grammars, pregroups) are problematic in two respects (overgeneration and undergeneration) compromising the accuracy of the quantitative values associated with a derivation. We address these problems by moving to a non-symmetric, non-associative, non-unital type logic with a tree-building tensor operation, generating *phrases* rather than strings. Composition (tensor) and decomposition of phrases (cotensor) are treated on a par. Reordering and restructuring are controlled by adjoint pairs of modalities, the grammatical analogues of Linear Logic's '!'. We discuss the categorical structures for this model of syntax and the associated graphical language. We identify some empirical areas where the model leads to improved performance.

In the field of natural language semantics, the compositional distributional framework of [3] and subsequent work (see [9] for an overview of results obtained so far) has achieved remarkable progress by unifying vector space models for lexical meanings with a compositional account of how these meanings combine into phrases and larger units. Interpretation takes the form of a *functorial* transition from Form to Meaning: a structure-preserving map that associates the operations for building syntactic structure with vector composition operations, thus assigning quantitative values to these structures.

The quality of the quantitative values thus obtained is determined by the accuracy of the syntactic engine driving the interpretation process. Compositional Distributional Semantics has used type logics for that purpose: Lambek's original Syntactic Calculus (**L**), and its more recent Pregroup incarnation (**PG**). Categorically, these are systems with a (non-symmetric) monoidal bi-closed or compact closed structure, respectively.

As models of natural language syntax, these calculi are lacking in two respects: overgeneration and undergeneration. Both **L** and **PG** model the composition of phrases with an *associative* multiplicative tensor operation, claiming in fact that no aspect of grammatical organization beyond linear order can affect

wellformedness. Lambek [7] was the first to recognize that as a result of global associativity certain structures are wrongly identified as wellformed. The problem is aggravated in a monoidal setting, where the multiplicative tensor comes with a unit.

As for undergeneration, both **L** and **PG** have the recognizing capacity of context-free grammars. Cases of information flow between physically detached parts of an utterance requiring expressivity beyond context-free are well documented in the literature on so-called mildly context-sensitive formalisms [5]. Discontinuous dependencies of that kind are beyond the analytical reach of type logics such as **L** or **PG**.

Our strategy for dealing with these issues has three ingredients:

1. Our basic syntactic engine is a non-symmetric, non-associative, non-unital type logic. The tensor in this setting becomes a tree-building operation, generating *phrases* rather than strings.
2. We treat grammatical composition (fusion) and decomposition (fission) on a par, restoring the distinction between tensor (multiplicative product) and cotensor (multiplicative sum). In the degenerate case of **PG** these operations are identified.
3. We enrich the formula language with syntactic control modalities — substructural relatives of the '!' of Linear Logic — and show that in moving to a more discriminating grammar logic, no expressivity is lost.

$$(1)\ \dfrac{\dfrac{B \to A\backslash C}{A \otimes B \to C}}{A \to C/B} \qquad (2)\ \dfrac{\dfrac{C \oslash A \to B}{C \to B \oplus A}}{B \osl C \to A} \qquad (3)\ \dfrac{A \to \Box^{\downarrow} B}{\Diamond A \to B}$$

Fig. 1. The three ingredients of the base system: residuation laws

The combination of (1) and (2) on this agenda leads to the Lambek-Grishin calculus (**LG**) of [8, 2, 1]. Its categorical structure is worked out in [10], together with a pictorial calculus for which *coherence* (soundness, completeness) is proved. Structures (and co-structures) of **LG** are non-associative objects, reflecting the fact that meaning composition in natural language takes hierarchical constituent information into account. Empirical support for this position comes from disambiguation studies (reported on in [9]) showing that direct objects have a stronger influence on the meaning of a transitive sentence than subjects. In non-associative **LG**, this is exactly what is predicted by a transive verb type $(n_{subj}\backslash s)/n_{obj}$. In associative **PG** (or **L**) the effect has to be stipulated by imposing extra equations, because the distinction between $(n_{subj}\backslash s)/n_{obj}$ and $n_{subj}\backslash(s/n_{obj})$ is lost.

For item (3) on the agenda, we equip the grammar logic with the counterpart of the '!' operation of Linear Logic. In **LL**, $!A$ singles out a renewable (stable)

piece of information of type A within an otherwise resource-conscious world. Copying and/or deletion of resources are brought back in a *controlled* form in the sense that they have to be explicitly licensed by this modality.

For the logic of grammar, control regards not so much the multiplicity of the resources, but rather the structural aspects of their composition: precedence (linear order), dominance (hierarchical structure). To achieve control in these domains, [6] decompose the linear '!' into an *adjoint pair* of operations ($\Diamond, \Box^{\downarrow}$ in Fig 1) together with structural postulates keyed to these modalities. A set of embedding theorems then shows that from the non-associative base logic, the expressivity of **LL** can be regained.

We work out the fine-structure of this approach. Following the example of [4], where '!' is decomposed into '$!_c$' (copying) and '$!_w$' (deletion), we further factorize the grammatical control modalities into their elementary parts, providing restricted forms of commutativity and associativity or their combination. The elementary components, in isolation, represent small local rewirings that leave the form-meaning correspondence intact. Chained together by sequential composition they create global effects: non-local information flow through a grammatical structure. The types of non-local composition that arise depend on the basic building plan of a language — its 'structural fingerprint'.

In line with the above we extend the categorical framework of [10] to accommodate the control modalities, and provide the associated graphical language. We discuss empirical validation for the extended model.

References

1. A. Bastenhof. *Categorial Symmetry.* PhD thesis, Utrecht University, 2013.
2. R. Bernardi and M. Moortgat. Continuation semantics for the Lambek-Grishin calculus. *Inf. Comput.*, 208(5):397–416, 2010.
3. B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394, 2010.
4. B. Jacobs. Semantics of weakening and contraction. *Annals of Pure and Applied Logic*, 69(1):73 – 106, 1994.
5. L. Kallmeyer. *Parsing Beyond Context-Free Grammars.* Springer, 2010.
6. N. Kurtonina and M. Moortgat. Structural control. In P. Blackburn and M. de Rijke, editors, *Specifying Syntactic Structures*, pages 75–113. CSLI, Stanford, 1997.
7. J. Lambek. On the calculus of syntactic types. In R. Jacobson, editor, *Structure of language and its mathematical aspects*, pages 166–178. American Mathematical Society, 1961.
8. M. Moortgat. Symmetric categorial grammar. *J. Philosophical Logic*, 38(6):681–710, 2009.
9. R. Piedeleu, D. Kartsaklis, B. Coecke, and M. Sadrzadeh. Open system categorical quantum semantics in natural language processing. *CoRR*, abs/1502.00831, 2015.
10. G. Wijnholds. Categorical foundations for extended compositional distributional models of meaning. MSc Thesis, ILLC, University of Amsterdam, 2014.