Category Theory in Theoretical Linguistics: A Monadic Semantics for Root Syntax

Extended abstract

This paper puts forward a monad-based semantics for the branch of theoretical linguistics known as "root syntax." It shows that compared to alternative approaches, the categorical approach presented here not only neatly separates the compositional and noncompositional aspects of natural language meaning but also smoothly extends from classical to generalized root syntax, thus providing a unified mode of composition for lexical and semilexical words. This paper builds on results from an earlier conference paper [15] but has two major updates: (i) it further clarifies the categorical setting of the syntax-semantics interface being assumed; (ii) it discusses the advantages of the monadic approach in comparison with a potential applicative approach.

1 Linguistics (background introduction)

In theoretical linguistics, more exactly in the transformational-generative (i.e., Chomskyan) framework thereof, there is a popular branch of research known as "root syntax." The prototypical issue it tackles is that of content word formation. While content words like nouns and verbs are treated as grammatical primitives in traditional linguistic theory, they are decomposed in root syntax into a root part and a category part (here "category" is used in its linguistic sense). For instance, the verb *eat* is decomposed into a "verbalizer" *v* and a categoryless root EAT. The former encodes categorial information, namely what makes a verb a verb in a specific language, while the latter encodes idiosyncratic information, namely the abstract concept underlying the specific meaning 'eat' (and some phonological index).

Root syntax has had much influence in generative grammar in the past two decades, as represented by its two established incarnations—distributed morphology [8] and exoskeletal syntax [3]. More recently, it has been further extended from the lexical domain to the grammatical domain (e.g., [1], [14], [13]). This extension, called "generalized root syntax" in [14], is motivated by the observation that human language makes use of "semilexical" items in addition to the traditionally recognized lexical and grammatical items. Semilexical items are words or morphemes with both lexical and grammatical content, and hence their meanings cannot be adequately explained either by lexical or by compositional semantics alone—both sides must be considered and somehow combined to explain their linguistic behavior.

A most typical class of semilexical items is classifiers, such as Mandarin Chinese *tiáo* 'for long, thin objects' and Japanese *hon* 'for long cylindrical objects'. Classifiers logically serve to atomize kinds and thereby prepare the ground for counting, but they simultaneously introduce miscellaneous idiosyncratic, conventionalized perspectives onto the atomization process. Another good example of semilexicality is negation particles in Vietnamese, where a single *not* in English can be translated in some ten ways depending on factors like register and speaker attitude, including *không* 'default', *nào* 'colloquial but elevated', *đéo* 'very vulgar', etc. Logically, a negator just denotes $\neg$, but beyond that basic functionality each Vietnamese negator also introduces a particular use-conditional or conventional reading, which is no less salient than the logical reading in native speakers' intuition. Similar phenomena are prevalent in the world's languages (see [16] for a survey), which provides strong empirical motivation for generalized root syntax. The idea is simple: a semilexical item is decomposed into a category part and a root part

Submitted to:
ACT 2022

too, though unlike in classical root syntax, here the category part is a grammatical category (like Cl or Neg) instead of a traditional lexical category (like V or N). As usual, the category part contributes compositional meaning, and the root part, idiosyncratic meaning.

Despite the popularity of root syntax among syntacticians, root-oriented thinking (i.e., ultimate lexical decomposition) has had surprisingly little impact on semanticists' work, where categories are rarely (if ever) severed out of content words, and the issue of semilexicality is not even mentioned. This creates a nontrivial problem for theoretical linguists working on the (morpho)syntax-semantics interface, since a portion of well-defined syntactic representations is systematically glossed over in syntax-to-semantics mapping. Solving this problem is not easy, because it requires one to rethink standardly assumed semantic primitives and break them down into even smaller atoms in a principled, syntax-echoing fashion. Song [15] considers four potential solutions, one from existing literature and three original, and concludes that the category-theoretic, monad-based approach is the most promising among them. Due to space limitations many details are left out in that paper, which the current paper aims to (partially) fill in.

2 Category theory (the application)

Song's [15] application of category theory directly builds on Asudeh & Giorgolo's [2] core idea. A&G use the monad tool to tackle various "enriched meanings." The particular type of enriched meaning that Song analogizes to root meaning is that of conventional implicature, namely the conventional illocutionary effects of certain content words, such as *Yank* and *cur*, which basically mean 'American' and 'dog' but simultaneously carry a negative speaker attitude. Likewise, the idiosyncratic aspect of a (semi)lexical item's meaning can be seen as quasi conventional implicature, in that it, too, is not really logically definable. Logicians may freely use predicate labels like **dog**, but what such labels really mean is a cognitive scientific question. To echo a long-standing view of Chomsky, "a lexical item provides us with a certain range of perspectives for viewing what we take to be the things in the world, or what we conceive in other ways; these items ... themselves do not refer ..." ([4], p.36). Following A&G, Song uses the writer monad to handle semantic composition involving conventionalized, non-truth-conditional information, with the pure function part of the monad being used for logical composition and its side-effect part, for root information logging (see **Definition 6** in [15] for detail). In other words, each "root-supported" syntactic category (in the linguistic sense), be it normally lexical or semilexical, maps to a monadically typed term in the semantics. Due to the fundamental nature of content words in human language, this essentially means that semantic composition is always monadic under root-oriented thinking (examples are left out in this extended abstract; see [15] for some).

Despite the shared writer monad tool, [15] and [2] actually differ greatly in their background assumptions about natural language syntax and the syntax-semantics interface. Hence, their ambient categorical settings also significantly differ from each other. This is mainly because A&G and Song work in different linguistic research paradigms. A&G adopt Moortgat's [11] version of categorial grammar for natural language syntax and an adapted version of Dalrymple et al.'s [6] glue semantics for semantics, the latter of which is in turn based on Moggi's [10] computational λ-calculus (which supports monads) and a fragment of linear logic. Under this setting, the glue-semantic representations can be viewed as inhabitants of a syntactic category (mathematical sense), which can be interpreted in a suitable model-theoretic structure, namely a semantic category (mathematical sense), via a functor (A&G do not really clarify the ambient categorical setting, but see [7] for a relevant model-theoretic discussion).

Due to the categorial grammar foundation of A&G's theory, their ambient categorical setting is likely to be compact closed, as is generally the case in the Lambekian paradigm. By comparison, the categorical

setting in Song's theory is merely cartesian closed. Like A&G, Song does not spell this out, but it is clear from his assumptions about natural language syntax. Song works in the Chomskyan paradigm, where word order is not encoded in the syntax but handled by a dedicated phonological module of the grammar (the modularity in the Chomskyan paradigm is mainly due to its goal of developing a theory of the human mind). Therefore, the left/right-based notions in pregroup grammar become superfluous for the syntax-semantics interface. In fact, since Song basically follows the mainstream formal semantic framework for generative grammar as specified in [12] and [9], which is basically simply typed λ-calculus, a cartesian closed setting is enough. More specifically, this CCC is one of semantic types for natural language, which in turn can be viewed as a tiny subcategory of **Sets**. Under this ambient setting, the writer monad for root syntax proposed in [15] can be easily added.

3 Advantages of the monadic approach

An anonymous conference reviewer commented on an earlier version of this work that it might be worth considering an applicative functor–based alternative approach. In the current paper, I address the monad vs. applicative issue from three angles, arguing that the monadic approach is overall more advantageous. First, while it is true that in computer science an applicative is preferred when the extra power of monad is not needed, the endofunctor for root syntax in [15] is independently definable as a monad whether or not it can be defined as an applicative. Hence, in our case using an applicative functor will not make the monad in the ontological background vanish, and the economy-based counterargument is weakened. Second, we probably do need the extra argument-handling flexibility of a monad, since the syntactic representations in root syntax fall in different patterns. Three patterns already manifest themselves in the limited examples in [15]: (i) nonmonadic head, monadic complement; (ii) monadic head, monadic complement; (iii) monadic intermediate phrase, nonmonadic specifier. Given the [$\pm$monadic] possibilities of various positions in the phrase structure, an applicative-based alternative approach may turn out to be insufficient or too clumsy. Third, we probably also need the extra value-dependent effect flexibility of a monad, since empirically speaking there are semilexical items whose conventionalized idiosyncratic reading depends on the type of the complement they take. For instance, the Mandarin Chinese passive auxiliary *bèi* carries a sarcastic speaker attitude when it operates on an intransitive verb but not when it operates on a transitive verb, so phrases like *bèi xìngfú* 'lit. be happied (i.e., someone is not actually happy but official propaganda says they are happy)' and *bèi sǐwáng* 'lit. be died (i.e., someone is still alive but the media say they are dead)' are sarcastic whereas phrases like *bèi chī* 'be eaten' and *bèi mà* 'be scolded' are not (that is, unless the discourse itself is sarcastic). A similar situation occurs with the Japanese passive suffix *-(ra)reru*, which, when attached to an intransitive verb (or a larger one-place predicate), carries a negative tone (i.e., the indirect or "suffering" passive). When the existence/absence or particular type of conventional implicature on a grammatical function depends on the logical type of its argument, it is probably easier to handle the composition with a monad rather than an applicative.

4 Future prospect

We can notice certain similarity between the core idea of root syntax and that of Coecke et al.'s [5] distributional compositional semantics—namely, compositional semantics and lexical semantics can be separately treated and then somehow combined in a unified formal setting. While the categorical semantics discussed here and that developed by Coecke et al. correspond to very different syntactic theories

(Chomskyan vs. Lambekian), there might be a possibility to combine some insights from the two lines of work in future research.

References

[1] V. Acedo-Matellán & C. Real-Puigdollers (2019): *Roots into functional nodes: exploring locality and semi-lexicality*. *The Linguistic Review* 36(3), pp. 411–436, doi:10.1515/tlr-2019-2019.

[2] A. Asudeh & G. Giorgolo (2020): *Enriched meanings: natural language semantics with category theory*. Oxford University Press, Oxford, doi:10.1093/oso/9780198847854.001.0001.

[3] H. Borer (2005): *In name only*. *Structuring sense* 1, Oxford University Press, Oxford.

[4] C. Chomsky (2000): *Minimalist inquiries: the framework*. In R. Martin, D. Michaels & J. Uriagereka, editors: *Step by step: essays on minimalist syntax in honor of Howard Lasnik*, MIT Press, Cambridge MA, pp. 89–156.

[5] B. Coecke, M. Sadrzadeh & S. Clark (2010): *Mathematical foundations for a compositional distributional model of meaning*, doi:10.48550/ARXIV.1003.4394.

[6] M. Dalrymple, J. Lamping & V. Saraswat (1993): *LFG semantics via constraints*. In S. Krauwer, M. Moortgat & L. des Tombe, editors: *Proceedings of the 6th Conference of the European ACL*, Association for Computational Linguistics, Utrecht, pp. 97–105, doi:10.3115/976744.976757.

[7] M. Gotham (2018): *Making logical form type-logical: glue semantics for minimalist syntax*. *Linguistics and Philosophy* 41(5), pp. 511–556, doi:10.1007/s10988-018-9229-z.

[8] M. Halle & A. Marantz (1993): *Distributed Morphology and the pieces of inflection*. In K. Hale & S. J. Keyser, editors: *Essays in linguistics in honor of Sylvain Bromberger*, *The View from Building* 20, MIT Press, Cambridge MA, pp. 111–176.

[9] I. Heim & A. Kratzer (1998): *Semantics in generative grammar*. Blackwell, Oxford.

[10] E. Moggi (1989): *Computational lambda-calculus and monads*. In: *Proceedings of the 4th Annual Symposium on Logic in Computer Science*, IEEE Press, New York, pp. 14–23, doi:10.1109/LICS.1989.39155.

[11] M. Moortgat (1997): *Categorial type logics*. In J. van Benthem & A. ter Meulen, editors: *Handbook of logic and language*, North-Holland, Amsterdam, pp. 93–177, doi:10.1016/B978-044481714-3/50005-9.

[12] B. Partee, A. ter Meulen & R. Wall (1990): *Mathematical methods in linguistics*. Kluwer, Dordrecht.

[13] C. Pots (2020): *Roots in Progress: semi-lexicality in the Dutch and Afrikaans verbal domain*. Ph.D. thesis, KU Leuven. Available at `https://www.lotpublications.nl/roots-in-progress-semi-lexicality-in-the-dutch-and-afrikaans-verbal-domain`.

[14] C. Song (2019): *On the formal flexibility of syntactic categories*. Ph.D. thesis, University of Cambridge, doi:10.17863/CAM.44842.

[15] C. Song (2021): *On the semantics of root syntax: challenges and directions*. In: *Proceedings of the 18th International Workshop of Logic and Engineering of Natural Language Semantics (LENLS18)*, pp. 61–74. Available at `https://www.juliosong.com/doc/Song2021LENLS18.pdf`.

[16] C. Song (2021): *A typology of semilexicality and the locus of grammatical variation*. Paper presented at International Conference of Formal Linguistics, Nov. 5–7, Fudan University. Available at `https://www.juliosong.com/doc/Song2021ICFL9.pdf`.