Realization, Referral, and Construction Morphology

David R. Mortensen

February 20, 2025

Construction Morphology

The most widely practiced approach to word-and-paradigm morphology today is probably Construction Morphology (CxM). This is the framework for Haspelmath & Sims (2010)[1] as well as a lot of other recent work in the area. Unlike some early WP frameworks, it works equally well for derivation and inflection.

The fundamental idea of CxG is that words are signs (like in IA morphology). However, signs are not necessarily compositional combinations of smaller signs. Constructions are like rules or schemas that combine smaller signs into bigger signs (sometimes in non-compositional ways). These signs exist at different levels of abstraction. Concrete words are constructions (instances of a more abstract construction). Constructions themselves may be instances of more abstract (or general) constructions (see Figure 1).

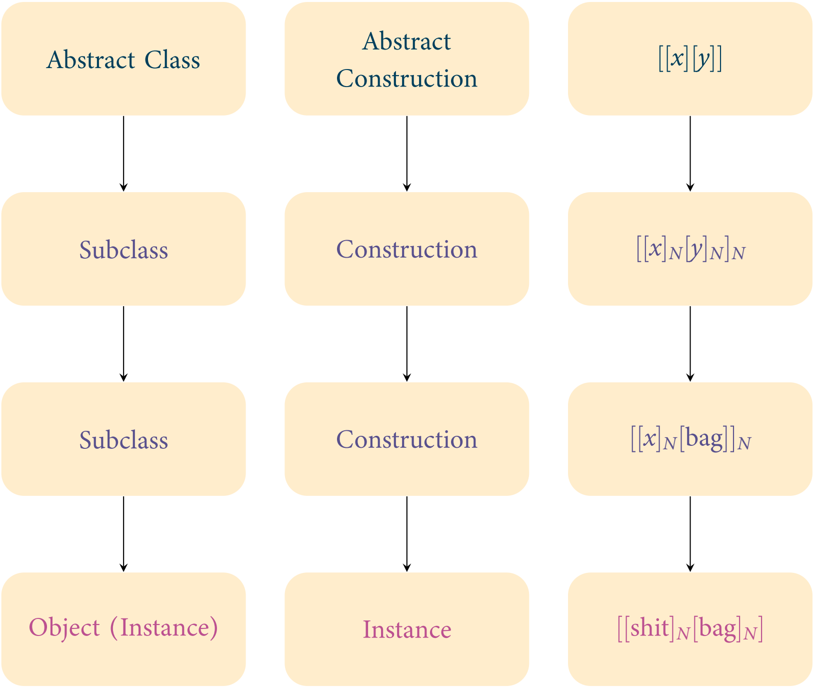

Figure 1: Comparison of Object Oriented Programming and Construction Grammar

Expressions like

(1) a. $[[x][y]]$

b. $[[x]_N[y]_N]_N$

c. $[[x]_N[\text{bag}]]_N$

[1] Martin Haspelmath and Andrea Sims. *Understanding Morphology*. Hodder Education, London, 2nd edition, 2010

d. $[[\text{shit}]_N[\text{bag}]_N]$

are ways of writing the form part of rules of realization. There are two things you should note:

1. This is a rule about compounding, not about inflection

2. It will only be complete if we add a meaning component.

Construction morphology provides a generalization of word-and-paradigm morphology to derivation and compounding (and provides a simple way of capturing the relationship between derivation, inflection, and compounding).

Rules of Realization and Rules of Referral in Construction Grammar

One way of capturing the meaning component and it's relationship to form is to adapt the notation developed by Geert Booij to the problem:

(2) $[[x]_{Ni}[\text{bag}]]_N \leftrightarrow$ unpleasant person who is a metaphorical container for SEM_i

a. dirtbag

b. shitbag

c. douchebag

where the subscript *N* indicates part of speech (noun) and SEM_i indicates the meaning associated with constituent *i*. This is (almost) a rule of realization. The semantic expression at the right side of the arrow is realized with the formal expression on the left side.

Another way of looking at the same construction involves using Haspelmath and Sims[2] notation for rules of referral (adapted):

(3)
$$\begin{array}{ccc} [\ \text{unpleasant thing } x\ ] & & [\ \text{person who is metaphorical container for } x\ ] \\ [X]_N & \leftrightarrow & [[X][\text{bag}]]_N \end{array}$$

These rules look different from the rules of realization and referral that we have examined so far, but the difference is superficial, with two exception. The first big difference is that the signified is expressed in natural language, rather than a collection of symbols like 1, S, and EXCL. This has been done because expressing lexical meaning in terms of attributes is difficult:

(4) a. It is difficult to decompose lexical meaning into a closed number of features, properties, or attributes.

b. Even when it is possible, it is difficult to agree on the features.

c. A simple data structure like a set does not capture scope and function-argument structure, so a more complicated representation is needed.

d. Extensional theories of semantics, based on first order logical and other similar logics, are well understood and can be used to model sentence semantics, but lexical semantics are less well-understood in this type of approach.

[2] Martin Haspelmath and Andrea Sims. *Understanding Morphology*. Hodder Education, London, 2nd edition, 2010

Natural language, though, is not as much a problem if you are modeling construction morphology using large language models that interact with the world through natural language.

The second big difference is that these Construction Morphology "rules" are not actually rules. They are schemata. That means that (2) does not actually mean that any combination of noun + *bag* will automatically refer to a person (consider *doggie bag*, *barf bag*, and *gym bag*) but that, given a word like *douchebag*, the meaning referring to an unpleasant human is licensed[3], not mandated.

Derivation in Construction Morphology

From here, forward, we will use the Haspelmath and Sims-inspired rules. Here is how agent and patient nominalizations might look in Construction Morphology:

(5) $$\underset{\text{X}}{\begin{bmatrix}\text{V}\\ \text{do } x\end{bmatrix}} \leftrightarrow \underset{\text{Xer}}{\begin{bmatrix}\text{N}\\ \text{person who does } x\end{bmatrix}} \leftrightarrow \underset{\text{Xee}}{\begin{bmatrix}\text{N}\\ \text{person to whom } x \text{ is done}\end{bmatrix}}$$

These schemas capture more than an IA analysis of these same data would. Specifically, they tell us that -er and -ee attach to roughly the same set of bases and are in a paradigmatic relationship with one another. Thus, for roughly every word formed with -ee, there is a possible agentive counterpart formed with -er.

attendee	attender
franchisee	franchiser
addressee	addresser
appointee	appointer
retainee	retainer
walkee	walker

Table 1: Agent–patient nominalization pairs in English.

In this framework, you do not simply build up words from morphemes—you infer possible words based on other words in the lexicon. **The "rules" or schemata are simply generalized analogies**. This presents a more computationally tractable and learnable formalism that IA or IP.

Let's relate this to Totonac. Take the following examples of adjective–intensified adjective pairs:

tlánka'	'large'	tlá:nka'	'immense'
tá'wah	'difficult'	tá:'wah	'very difficult'
qáma'	'tasty'	qá:ma'	'very tasty'

Table 2: Totonac adjective–intensified adjective pairs.

These pairs can be modeled with a schema like the following:

[3] In linguistics, we say that X licenses Y if Y is permissible in the presence of X (and that Y may potentially occur without X specifically).

$$\begin{array}{ccc} \left[\text{has property } x\right] & & \left[\text{has property } x \text{ intensively}\right] \\ \text{C(C)(C)VX} & \leftrightarrow & \text{C(C)(C)V:X} \end{array}$$

This is a description of a pattern in the lexicon that can be used to construct new words. It is not, however, a generative rule in the sense of Chomskyan linguistics.

Inflection in Construction Morphology

We can then loop back to inflection. Inflection in Construction Morphology works a lot like inflection in other kinds of Word-and-Paradigm frameworks. Consider the following example from one of the conjugations of Spanish verbs:

$$\begin{array}{cccccc} \text{(6)} & \begin{bmatrix} \text{V} \\ 1 \\ \text{S} \\ \text{Pres} \end{bmatrix} & & \begin{bmatrix} \text{V} \\ 3 \\ \text{S} \\ \text{Pres} \end{bmatrix} & & \begin{bmatrix} \text{V} \\ 1 \\ \text{P} \\ \text{Pres} \end{bmatrix} \\ & \text{Xo} & \leftrightarrow & \text{Xa} & \leftrightarrow & \text{Xamos} \end{array}$$

Philosophical Musings

In some sense, construction morphology seems to have a lot in common with how large language models do morphology (as far as we can tell). There is little reason to believe, either reasoning from first principles or by observing their behavior, that LLMs decompose words into morpheme-like units (even given tokenization) or processes and less evidence that they do morphology with rules. Rather, morphological generation and understanding seem to driven by analogy. In a recent study, we found that ChatGPT's morphological productions in a wug task were actually **far more likely than those from humans** to be based on analogies with similarly-spelled words.[4]

References

Martin Haspelmath and Andrea Sims. *Understanding Morphology*. Hodder Education, London, 2nd edition, 2010.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore, December 2023. Association for Compu-

[4] Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore, December 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.401. URL `https://aclanthology.org/2023.emnlp-main.401`

tational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.401. URL `https://aclanthology.org/2023.emnlp-main.401`.