Signs, Minimal Signs, and Compositionality

David R. Mortensen

January 16, 2025

Introduction

Consider the word *compositionality*. It is a perfectly understandable English word, but it is likely to be rare in most corpora. To improve the ability of our models to represent it, we may want to break it into parts, as discussed in the previous lecture. We could divide it into sequences of two, for example:

(1) co mp os it io na li ty

Each of these units is likely to be far more common in a training corpus than ⟨compositionality⟩. However, the embeddings of any of these tokens are likely to be uninformative regarding the meaning of this word. That is, the tokens are not SIGNS and they cannot be combined COMPOSITIONALLY.

Signs

A sign is a pairing of form (something observable) with meaning. Consider the road sign in 1.

Figure 1: A road sign meaning "watch for falling rocks."

The form is the observable nature of the sign (a yellow square rotated 45 degrees with a black margin and with assorted black shapes printed on it). The meaning is a warning: rocks fall from the mountainside in this place and you should watch for them. It could even be interpreted as a command: "Watch for falling rocks!"

Signs pervade language and signaling systems. In fact, they are the whole basis of such systems. Linguists (language scientists) and semioticians (people who specialize in the study of signs, generally), have developed a whole vocabulary for discussing signs (a system of signs for discussing systems of signs) but we are going to be make due with a relatively limited set of terminology. The form of the sign (the part you observe) we will call the SIGNIFIER and the meaning of a sign we will call the SIGNIFIED. The important thing to remember is that the sign is not the *signifier* (the yellow square with the black shapes, in our example) and it is not the *signified* (the meaning "beware of falling rocks!"). Rather, it is the *pairing* of signifier and signified.

This becomes clear when we refer back to (1). We are presented with eight forms (pairs of letters, in this case) but none of them is the signifier in a sign (except, perhaps, ⟨co⟩) because none of them (again, except for ⟨co⟩) is paired with a signified.[1]

Note, too, that a sign is not a pairing of a particular physical object—an instance—with a signified. Rather, it is a pairing between classes or types. That is, the "falling rocks" SIGN is not the particular sheet of metal coated in

[1] Even in the case of ⟨co⟩, this relationship is, in context, very tenuous.

reflective paint that is hanging in Sardine Canyon in Northern Utah. Rather, it is the general relationship that allows members of our social community to find meaning in objects of this type.

Compositionality

Note that there are many signs similar to (1) in North American driving culture.

Consider, for example, Figure 2. Like the sign in Figure 1, it consists black objects on a yellow square. And like the earlier sign, this one has a meaning like "warning!" or "watch out!" The object is different, though, and what we are asked to watch out for is also different. In this case, the black figure is a stylized representation of a deer.

Figure 2: Dear warning sign.

It doesn't stop with deer. As Figures 3 and 4 show, the same general "warning" sign can be combined with other signs to make complex signs meaning "beware of bikes" and "beware of pedestrians" as well as "beware of falling rocks" and "beware of deer."

Figure 3: Bike warning sign.

The fact that signs can be combined according to simple principles to yield more complex signs (with predictable form and meaning) is called COMPOSITIONALITY and is an important design feature of language.

In the last lecture, we learned that minimal meaningful units—minimal signs—are called MORPHEMES. Morphemes can combine in predictable ways to produce more complicated signs with predictable forms and meanings—predictable signifiers and signifieds.

Figure 4: Pedestrian warning sign.

Table 1 shows a trivial example of this. There are suffixes that can combine with most English verb roots. These combinations are predictable sums of the parts. In this table, each of the suffixes is rather like the yellow square—it adds

Table 1: Compositionality in English words

INFINITIVE	PAST TENSE	PRES. PART.	AGENT NOUN
kill	kill-ed	kill-ing	kill-er
hunt	hunt-ed	hunt-ing	hunt-er
walk	walk-ed	walk-ing	walk-er
kiss	kiss-ed	kiss-ing	kiss-er

meaning to the more concrete sign with which it is combined. Just as yellow square ('beware') + deer ('deer') yields 'beware of deer', *hunt* ('hunt') + *-er* ('noun that does something') yields *hunter* ('one who hunts').

We can apply the same principle to our original example and arrive, to a first approximation, at (2):

(2) composit ion al ity

The ROOT of this word is *composit* (as in *compositor*, *composition*, *composite*, and so on). The suffix *-ion* makes verbs into "abstract" nouns. The suffix

On historical grounds, we might divided *compos* into *com-*, *pos*, and *-it*, but time has rendered this relationship opaqe.

-al makes nouns into adjectives. The suffix *-ity* makes an adjective *p* into a noun meaning the state of having the property *p*.

The suffixes are like the warning sign in Figure 5: they provide the framing in which the meaning of the more concrete sign (the root or the icon) can be understood.

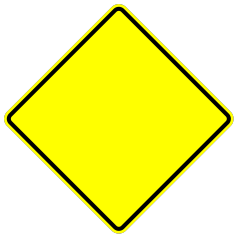

Figure 5: Warning sign.

Constructions

However, some signs go beyond compositionality. Rather than having a simple function that combines signifiers and signified to produce a new sign, they have a CONSTRUCTION-specific function. Take for example, the following names of language families:

(3) a. Sino-Tibetan
b. Tibeto-Burman
c. Indo-European
d. Indo-Iranian
e. Afro-Asiatic
f. Nilo-Saharan
g. Italo-Celtic

Italo-Celtic refers to the language family that includes both the Italic[2] and Celtic branches of Indo-European (a big language family that includes many languages of India, Iran, and Europe).

It is made by combining a special (sometimes truncated) root referring to the first linguistic or geographical category, the FORMATIVE *-o*, and an adjective referring to the second linguistic or geographical category. This adjective consists of the root, followed by either *-ic* or *-an* dependening on the root. The *-o* does not have any referent—it exists solely to complete this pattern (CONSTRUCTION). In fact, there is not a general rule that would allow you to compose these signifiers together to get the right signifieds. The pattern is specific to this particular kind of COMPOUND word.

A construction is a form-meaning pattern that generalizes over multiple words, phrases, sentences, etc. Suffixing *-ed* to form the English past tense is a construction, part of a very general construction for INFLECTING words for tense by adding suffixes. The language family construction, though, is much more specific.

[2] The Italic languages include Latin and it's descendants as well as the other, closely-related, languages of ancient Italy like Oscan and Umbrian.

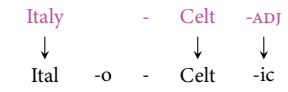

Figure 6: The anatomy of *Italo-Celtic*.

A compound word is a word formed by combining two or more roots or words. *Textbook*, *dishwasher*, and *cowbird* are a compounds.

Consequences

To return to the tokenization problem, we can now place it in semiotic perspective. The surface form of a word (the string of characters) is the signifier. Its embedding, and other deeper vector representations of a word/token, are—it a real sense—the signified (a representation of its meaning). In tokenization, the task is to break words into tokens so that no tokens are ever out-of-vocabulary but reasonable embeddings can be learned for all tokens.

Now is a good time to repeat some hypotheses from the first lecture, in different words:

(4) a. Words should be tokenized so that the tokens are signs.

b. Ideal tokens should either be *words* or *morphemes*.
c. When a token is not a word or morpheme, it should be a sequence of morphemes.
d. Successful tokenization schemes (e.g., BPE and sentencepiece) are successful because they approximate these ideals.
e. It is possible to do better.

Example Exercises

Tz'utujil

Identify the morphemes (both form and meaning) in the following data from Tz'utujil, a Mayan language of Guatemala [3]:

[3] Jon Philip Dayley. *Tzutujil grammar*, volume 107. Univ of California Press, 1985

xinwari	'I slept'	xoqeeli	'we left'
neeli	'he or she leaves'	ninari	'I sleep'
ne7eeli	'they leave'	xixwari	'y'all slept'
nixwari	'y'all sleep'	xe7eeli	'they left'
xateeli	'you left'	xwari	'he or she slept'
natwari	'you sleep'		

Table 2: Verbs from Tz'utujil. The ⟨7⟩ represents a GLOTTAL STOP (the sound between the two vowels in *uh-oh*). The ⟨x⟩ represents the same sound as ⟨sh⟩ in English. 'You' refers to 'singular you' and 'y'all' refers to 'plural you.'

Totonac

Identify the morphemes (form and meaning) in the following data from Totonac. Not that, in some cases, a single "morpheme" my include more than one piece of meaning and that two morphemes may include the same piece of meaning. This is a tricky exercise and at least two solutions are possible.

	Present	Past	Future
1SG	kpaʃ	kpaʃlh	nakpáʃ
2SG	páʃa'	paʃt	napáʃa'
3SG	paʃ	paʃlh	napáʃ
1PL.EXCL	kpaʃá'w	kpaʃw	nakpaʃá'w
1PL.INCL	paʃá'w	paʃw	napaʃá'w
2PL	paʃá:ti't	ʃpaʃá:ti't	napaʃá:ti't
3PL	paʃqó:'y	ʃpaʃqó:'y	napaʃqó:y

Table 3: Inflection of the Totonac verb form 'swim' or 'bathe'. Note that 1, 2, and 3 represent first, second, and third person and SG and PL represent singular and plural (so that 1SG represents 'I/me' and 2PL represents 'ya'll' or 'you guys'. There are two kinds of "we" in this language. Exclusive (EXCL) excludes the person you are talking to. Inclusive (INCL) includes that person. It may be helpful to note that 1SG is always exclusive.

Kikuyu

Identify all of the morphemes in the following data from the Kenyan language Kikuyu, ignoring the tones (indicated by the accents):

		‘look at’	‘send’
1.	‘we are V-ing’	torɔraɣa	totomáɣa
2.	‘we are V-ing him/her’	tomorɔraɣa	tomotomáɣa
3.	‘we are V-ing them’	tomarɔ́raɣa	tomatómáɣa
4.	‘they are V-ing’	márɔ́raɣa	mátómáɣa
5.	‘they are V-ing him/her’	mámórɔraɣa	mámótomáɣa
6.	‘they are V-ing them’	mámárɔ́raɣa	mámátómáɣa
7.	‘we V-ed’	torɔriɾɛ́	totomíɾɛ́
8.	‘we V-ed him/her’	tomorɔriɾɛ́	tomotomíɾɛ́
9.	‘we V-ed them’	tomarɔ́riɾɛ́	tomatómíɾɛ́
10.	‘they V-ed’	márɔ́riɾɛ́	mátómíɾɛ́
11.	‘they V-ed him/her’	mámorɔriɾɛ́	mámótomíɾɛ́
12.	‘they V-ed them’	mámárɔ́riɾɛ́	mámátómíɾɛ́

Table 4: Two Kikuyu verbs. The acute accent represents a high tone. Ignore them on your first pass.

References

Jon Philip Dayley. *Tzutujil grammar*, volume 107. Univ of California Press, 1985.