Non-Concatenative Morphological Processes

David R. Mortensen

February 6, 2025

Introduction

Most morphology involves concatenating morphemes together:

- Prefixation
- Suffixation
- Compounding

Consider, however, the following examples: In fact, non-concatenative morphological processes are common throughout the languages of the world.

Table 1: Examples of umlaut in English.

SINGULAR	PLURAL
foot	feet
tooth	teeth
goose	geese
man	men
mouse	mice

Generalized Glossing Guidelines

In order to more effectively GLOSS non-concatenative processes, we developed an annotation convention called G3 (Generalized Glossing Guidelines)[1]. It represents non-concatenative processes as string rewrites:

(1) I have two left f{oo>ee}t
1.SG have two left foot{PL}
'I have two left feet'

Similarly, here is an example of umlaut in German:

(2) Ich habe vier Br{u>ü}der
1.SG have.1.SG four brother{PL}
'I have four brothers.'

The same convention can be used to annotate the whole gamut of non-concatenative processes:

We will talk about each of these processes in more detail.

[1] David R. Mortensen, Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja, and Lori Levin. Generalized glossing guidelines: An explicit, human- and machine-readable, item-and-process convention for morphological annotation. In Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors, *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 58–67, Toronto, Canada, July 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.sigmorphon-1.7. URL https://aclanthology.org/2023.sigmorphon-1.7

Infixation

Ulwa, a Misumalpan language of Nicaragua has suffixing infixation:

Type	Example	Glosss
Infixation	s{>um}ulat	write{PFV}
Reduplication	{>su}sulat	write{PROSP}
Transfixation	k{i>u}t{a>u}b	book{PL;1,2}
Apophony	t{u>i}θ	tooth{PL}
Segmental overwriting	{xi>ku}3xi3	eat{IRR}
Tonal overwriting	ku{3>14}ni2	want{NEG}

Table 2: Example forms and glosses for a range of morphological processes.

(3) a. wahai{>ki}
wahai{ki}
brother{POSS::1.SG}
'my brother'

b. sû{>ki}lu
sû{ki}lu
dog{POSS::1.SG}
'my dog'

But there was also infixation in Latin:

(4) ta{>n}g{>o}
ta{n}g{o}
touch{1.SG.PRS.IND}
'I touch.'

Reduplication

(5) Nahuatl reduplication with fixed segmentism:

a. ti- ne:ch- {>teh}te:mowa -0
SUBJ::2S- OBJ::1S- look_for{RED} -PRS.IND.S
'You miss me.'

b. ni- mits- {>ih}ita -0
SUBJ::1S- OBJ::2S- see{RED} -PRS.IND.S
'I visit you.'

The first consonant (if present) and the first vowel are repeated before the stem, followed by /h/.

An example from Mangap-Mbula:

(6) kuk{>uk}
kuk{uk}
bark{PROG}
'be barking'

The final VC is suffixed to the stem.

In Pima, a Uto-Aztecan language of the United State, plurals are formed by infixing a copy of the first consonant of a stem after the first vowel:

(7) a. ma{>m}vit̪
ma{m}vit̪
lion{PL}
'lions'

b. ʧi{>tʃ}mait̪
ʧi{tʃ}mait̪
drum{PL}
'drums'

Classical Latin also featured infixing reduplication of the first vowel and the consonant before it:

(8) s{>po}pond{>ī}
spopondī
perform{1.SG.PRF.IND;1,2}
trans 'I perform'

Conversion

(9) a. I'm going to *swim* across the lake.
b. I'm going to take a *swim* across the lake.

(10) a. Coke comes in a *bottle*.
b. They *bottle* Coke.

Truncation

Murle forms plurals by deleting the last consonant or vowel:

(11) a. nyoo{n>0}
nyoo{}
lamb{PL}
'lambs'

b. wawo{c>0}
wawo{}
white_heron{PL}
'white herons'

Apophony

In Totonac, diminutives are sometimes formed by changing all instances of /ʃ/ into /s/:

ʃku'ta	'sour'	sku'ta	'a little sour'
ʃuːni'	'bitter'	suːni'	'a little bitter'
tʃi'tʃ	'hot'	tsi'ts	'a little hot'

Table 3: Totonac diminutives

Irish also has apophony for forming plurals.

(12) a. c{ea>i}nn
c{i}nn
head{PL}
'heads'

b. m{ui>a}r{>a}
m{a}r{a}
sea{PL;1,2}
'seas'

Tonal Overwriting

In Yoloxochitl Mixtec, habituals are formed by changing to tones on a word:

(13) ta'{3>1>4}bi{>1}4
ta'{4}bi{14}
get-broken{HAB;1,2}
'habitually get broken'

Stress Shift

English:

'object	ob'ject
'reject	re'ject
'conflict	con'flict
'contest	con'test
'insult	in'sult

Table 4: Stress shift in English

Segmental Overwriting

Although it is unusual, some languages have morphological processes where sequences of sounds are "overwritten" by another sequence of sounds.

(14) {xi>ku}3xi3
{ku}3xi3
eat{IRR}
'eat'

Transfixation

A few languages, mostly belonging to the AFROASIATIC family[2], have a kind of non-concatenative morphology called root-and-pattern morphology. The basic operation, in such a system, is a transfix. Transfixes add vowels across bases (not just one place) and may also lengthen consonants.

[2] Afroasiatic is the language family to which Semitic, Berber, Chadic, Cushitic, Omotic, and Egyptian belong. The best known Afroasiatic languages are Arabic, Hebrew, and Amharic.

Morphological Operations as Functions from Sign to Sign

A general Framework

We have implicitly treated a morphological operation as a function having the following type signature

$$f : \langle \Sigma*, G\rangle \rightarrow \langle \Sigma*, G\rangle \rightarrow \langle \Sigma*, G\rangle \quad (1)$$

	Perfect		Imperfect		Participle	
	Active	Passive	Active	Passive	Active	Passive
I	katab	kutib	ktub	ktab	kaatib	ktuub
II	kattab	kuttib	kattib	kattab	kattib	kattab
III	kaatab	kuutib	kaatib	kaatab	kaatib	kaatab
IV	ʔaktab	ʔuktib	ktib	ktab	ktib	ktab
V	takattab	tukuttib	takattab	takattab	takattib	takattab
VI	takaatab	tukuutib	takaatab	takaatab	takaatib	takaatab
VII	nkatab	nkutib	nkatib	nkatab	nkatib	nkatab
VIII	ktatab	ktutib	ktatib	ktatab	ktatib	ktatab
IX	ktab(a)b	ktab(i)b	ktab(i)b			
X	staktab	stuktib	staktib	staktab	staktib	staktab

Table 5: An Arabic paradign for the root *k-t-b* '(related to) writing'.

where $\Sigma*$ represents a string[3] corresponding to a signifier and G represents a graph corresponding to a signified. We have assume that the function f is always roughly like

$$f(\langle a, A\rangle, \langle b, B\rangle) = \langle a \oplus b, g(A, B)\rangle \tag{2}$$

where a and b are signifiers, A and b are signifieds, $\oplus$ is the concatenation operator, and g is a semantic composition function. For any particular instance of prefixation or suffixation, either $\langle a, A\rangle$ or $\langle b, B\rangle$ is going to be constant, so f is actually

$$f : \langle \Sigma*, G\rangle \rightarrow \langle \Sigma*, G\rangle \tag{3}$$

For suffixation of *ing*, f could be defined as

$$f(\langle a, A\rangle) = \langle a \oplus \text{ing}, g(A, \text{GERUND})\rangle \tag{4}$$

However, there is no reason we could not generalize this to

$$f(\langle a, A\rangle) = \langle p(a), g(A, G)\rangle \tag{5}$$

where p is a function in $\Sigma* \rightarrow \Sigma*$ and G is semantic content. p, then, could be any mapping from string to string (including substitutions, deletions, insertions, repetitions, or any combination of these. In other words, **a morphological operation/process is a function from signs to signs.** For example, Nahuatl reduplication with fixed segmentation could be notated as

(15) (C)VX+ → (C)Vh(C)VX+

where C matches a consonant, (C) matches an optional consonant, V matches a vowel, X matches and consonant or vowel, and + is the Kleene plus (one or more repetitions). This would map strings the start with zero or more

[3] We may not think of phonological representations as simple strings, but this suffices for the moment

consonants and one vowel (followed by some number of other segments) to strings starting with the same consonant and vowel, followed by an /h/, followed in turn by the rest of the string (matched by X+).

An Item and Process Tokenizer

In such a framework, where processes (concatenative and non-concatenative) can be seen as functions, words can be seen as the composition of functions and their application to a root. Consider the Nahuatl example we say above:

(16) ti- ne:ch- {>teh}te:mowa -0
SUBJ::2S- OBJ::1S- look_for{RED} -PRS.IND.S
'You miss me.'

Let us say we have four functions:

(17) a. 2SGSUBJ
b. 1SGOBJ
c. RED
d. PRSIND

The root is *te:mowa*. We can view the word as:

$$\textsc{PrsInd}(\textsc{2SgSubj}(\textsc{1SgObj}(\textsc{PrsInd}(\textsc{Red}(\text{te:mowa})))))) \tag{6}$$

or

$$(\textsc{PrsInd} \circ \textsc{2SgSubj} \circ \textsc{1SgObj} \circ \textsc{Red})(\text{te:mowa}) \tag{7}$$

The notation in Formula lends itself naturally to serialization. If each function or root is assigned an ID, then it lends itself naturally to tokenization. Two problems:

(18) a. Learning the rules that correspond to each of these functions is non-trivial. It is an unsolved problem in computational linguistics.
b. Parsing a corpus with such rules would be very computationally intensive and might not scale well.

Upshot

(19) a. Tokenization schemes typically treat language as a sequence of subword units
b. This makes sense if morphology is prefixation ∪ suffixation ∪ compounding (or if morphology is irrelevant)
c. What about non-concatenative processes? How should they interact with tokenization?
d. **Should we move to character-level or token-free models like ByT5?**

e. Will token-free models perform better for languages with lots of non-concatenative morphology like Arabic?

References

David R. Mortensen, Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja, and Lori Levin. Generalized glossing guidelines: An explicit, human- and machine-readable, item-and-process convention for morphological annotation. In Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors, *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 58–67, Toronto, Canada, July 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.sigmorphon-1.7. URL https://aclanthology.org/2023.sigmorphon-1.7.