Productivity and Generalization

David R. Mortensen

January 21, 2025

A Formalism for Tokenization and Word Segmentation

Before we continue, let's establish formal definitions for tokenization and word segmentation. Both sequences of graphemes/phonemes and sequences of token IDs/morpheme IDs are mad of symbols. Let's call the alphabet of grapheme/phoneme level of Hockett's dual articulation framework Σ. Let's call the alphabet of token IDs or morpheme IDs Δ. The set of possible grapheme/phoneme sequences we will call Σ^* and the set of possible token ID/morpheme ID sequences we will call Δ^*. Note that a linguistic sign (which could be a single word, a single sentence, or a whole document/discourse) is a pairing of a signifier $\sigma \in \Sigma^*$ and a meaning $\delta \in \Delta^*$.

Tokenization (or morpheme segmentation) can be formalized as a function $\tau : \Sigma^* \rightarrow \Delta^*$. Detokenization (or EXPONENCE/spell-out) can be formalized as a function $\kappa : \Delta^* \rightarrow \Sigma^*$. Note that, in tokenization schemes like BPE, κ is the inverse of τ, that is,

$$\kappa(\tau(\sigma)) = \sigma \text{ for all values of } \sigma \in \Sigma^* \tag{1}$$

In other words, κ and τ are both lossless. This is not necessarily the case for morpheme segmentation (and probably not the case for tokenization).

Consider the English word ⟨picks⟩. Out of context, this sequence is ambiguous. It can either be a noun referring to a tool or a verb meaning 'to select,' 'to gather by plucking,' or 'to dig into.' For a human, the root ⟨pick⟩ actually represents two different signs, and we can imagine a "linguistic" κ than maps then onto different token IDs drawn from Δ. In this case, κ would not be a function since it would be a one-to-many relation. τ could still be a (surjective) function, though. Likewise, one could image a κ such that ⟨entity⟩ and the ⟨entiti⟩ in ⟨entities⟩ are represented by the same token ID, since they represent—in linguistic terms—signifiers for the same sign. This would imply that our τ must be non-functional (but our κ still might be a—surjective—function).

In general, in tokenization, we want to minimize the vocabulary size ($|\Delta|$) and the average length of tokenized sequences. We might also say that we want to minimize the entropy of the training corpus C when tokenized by κ. In one formulation

$$l_{avg} = \frac{1}{|C|} \sum_{\sigma \in C} |\kappa(\sigma)|. \tag{2}$$

Conjecture: Human language learners parse speech and text, looking for subwords, in a way analogous to how tokenizers learn κ and τ. Morphemes are a project of compression (encoding form as meaning) and decompression (decoding meaning as form) coupled with linguistic and social context.

The biggest difference is that κ and τ are, for human language users, non-functional relations yielding lossy encoding/decoding.

Introduction to Productivity and Generalization

In 1958, Jean Berko Gleason reported what would become perhaps the most important experiment in the history of PSYCHOLINGUISTICS: the Wug Test[1]. In this experiment, small English-speaking children were asked to generalize a morphological construction to a NONCE WORD[2] Berko Gleason found that even very small children were able apply some English morphology (complete with rules of allomorphy). For example, when asked to complete a frame in which the plural of *wug* was required, children said *wugs* (pronounced /wʌgz/, with a ⟨z⟩ sound at the end).

Psycholinguistics is the study of how language is learned, processed, and produced by the brain.

[1] Jean Berko. The child's learning of English morphology. *Word*, 14(2-3):150–177, 1958

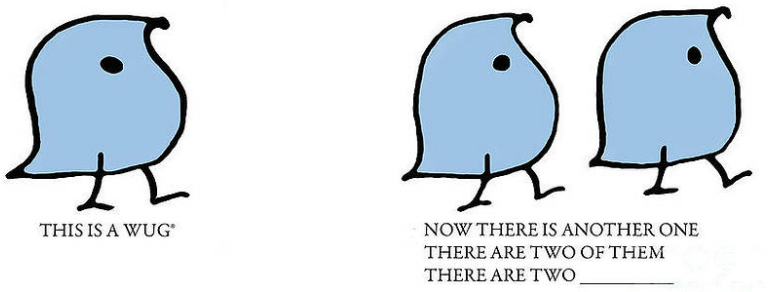

Figure 1: Image from the original Wug Test

[2] A made-up but possible word.

Pluralization with *-s* is PRODUCTIVE in English but pluralization with *-en* as in *oxen* is not. To be productive means that a construction can be applied to new inputs and, thus, yield new outputs. PRODUCTIVITY is the extent to which a morphological construction can apply to new words. Productivity is often correlated with compositionality. Morphological constructions that are more compositional are also more easily generalized to new words. That is why we are talking about it now.

Productivity

Productivity is actually multidimensional, but it might be helpful to start of visualizing it as a continuum, with default morphology on one end (mostly productive), fossilized morphology on the other (mostly not productive) and restricted morphology in the middle (productive under certain conditions). There are different suffixes for forming plurals in English. Of these, one of the least productive is *-en*, which only occurs in *oxen*, *children*, and *brethren*

fossilized ⟵	restricted ⟶	default
plural *-en*	plural *-ī*	plural *-s*

Figure 2: A continuum of productivity with English plurals. *-en* only applies to a couple of words. *-ī* applies only to Latin loanwords ending in *-us*. *-s* applies freely to all words where a more specific construction is not applicable.

Historical Stratum

Some morphological constructions only apply to words with a particular historical origin. For example, there is a negative prefix for English adjectives that is spelled *im-*, *-in*, *-ir* or *-il* depending on what follows it (a case of allomorphy). It occurs in words like the following:

(1) a. im-possible
b. in-tolerable
c. ir-regular
d. il-legal

These are all words that entered English from Latin via Norman French. There is another negative prefix as well:

(2) a. un-selfish
b. un-dutiful
c. un-answerable
d. un-knowable
e. un-knowledgeable

(3) a. un-complimentary
b. un-natural
c. un-substantial

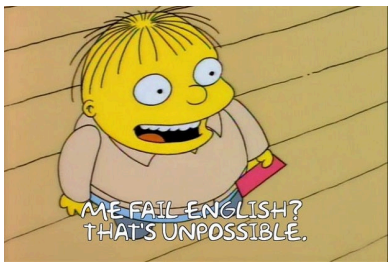

Figure 3: Ralph, a character from the once-popular animated television series, *the Simpsons*, famously defied morphological constraints.

The prefix applies to almost all adjectives from Germanic (from Old English) that you could negate, as exemplified in (2) as well as a number of words that entered English from French (3). Interestingly, the Germanic words cannot take the Latinate prefix (**in-knowable* is not okay) and the less-assimilated Latinate words (as exemplified in (1) cannot take the *un-* prefix. That is to say, **unpossible* is unpossible.

Upshot: While *un-* is a default, it applies to a historically-defined class of words. Something similar can be said of *-ness* (Germanic) and *-ity* (Latinate).

Semantic Class

Some morphological constructions only apply to words with particular SEMANTIC properties (a particular class of meanings). For example, the plural suffix 们 *-mén* in Chinese can only be attached to human nouns and pronouns:

(4) Acceptable

a. 我 -们
wǒ -mén
1SG -PL
'we'

b. 同志 -们
tóngzhì -mén
comrade -PL
'comrades'

(5) Unacceptable

a. 书 -们
shū -mén
book -PL
'books' (intended)

Morphological Constraints

Some morphological constructions only apply to inputs that have been produced in a particular way (morphologically speaking). In Russian, the FEMININE suffix *-ja* only affixes to bases that are produced by adding the suffix *-um* to a root (etc.). The feminine for all other nouns use a different suffix (*-ka*, *-ša*, *-inja*, or *-isa* [3]:

[3] Natalia Yulievna Shvedova et al. *Russkaya grammatika [Russian grammar]*. Moscow: Institute of the Russian Language, Russian Academy of Sciences, Moscow, 1980

Table 1: Words with the Russian feminine suffix *-ja*

Verb		Noun (MASC)		Noun (FEM)
govor-it^{j}	talk	govor-un	talker	govor-un^{j}-ja
beg-at^{j}	run	beg-un	runner	beg-un^{j}-ja
pljas-at^{j}	dance	pljas-un	dancer	pljas-un^{j}-ja
lg-at^{j}	lie	lg-un	liar	lg-un'-ja

Phonological Constraints

Some morphological constructions are sensitive to the sound structure (or phonology) of words to which they might apply. For example, the English suffix *-er* that makes ordinary adjectives into comparative adjectives can freely apply to words with one SYLLABLE and to many words with two syllables (following a rule we cannot yet describe in this course) but typically cannot apply to words with more than two syllables.

A syllable, for our current purposes, is a vowel preceded by zero or more consonants and following by zero or more vowels. It is defined in terms of sounds, not letters. Words are divided into syllables in such a way as to maximize the number of syllables that start with a consonant, minimize the number of syllables that end in a consonant, and minimize the number of consonant sequences within a syllable.

(6) Acceptable

a. smarter
b. faster
c. cleverer

d. speedier
e. brainier

(7) Unacceptable
a. *confuseder
b. *crowdeder
c. *arroganter
d. *differenter
e. *intelligenter
f. *expeditiouser
g. *adventurouser

Productivity and Psychological Reality

(8) Some assumptions that are sometimes made about grammar (including morphology)
a. Language and grammar are psychological phenomena
b. If a grammatical pattern is productive, it must be psychologically real
c. If a pattern in not productive, it must not be psychologically real (it is just fossilized historical leftovers)
d. Therefore, patterns that are not productive are not part of the language

This is problematic, though, in that it is not clear that language is just a psychological phenomenon. Certainly, natural language processing is concerned with language but is expressly not concerned with psycholinguistics. The same may be said of SOCIOLINGUISTICS and sociology of language. Furthermore, as we have seen, productivity is a continuous space—there is no solid boundary between productive patterns and fossilized or accidental ones.

In the approaches to language associated with (8), grammar is a discrete, self-contained system that defines a set of words or sentences that are in the language and (its complement) a set that are not. This view points, associated with thinkers like Noam Chomsky and Morris Halle, is still very important in theoretical linguistics and was once important in computational linguistics and NLP. However, with the empirical turn in language technologies, starting the in the 1990s and continuing up to the present, computational linguists have become much more inclined to view language as a continuous and gradient system rather than discrete one. Under such an approach, local regularities and gradient productivity cease to be anomalies. Indeed, perfect, unconstrained productivity becomes the anomaly (matching empirical observations from years of work).

Here is a working hypothesis:

(9) Morphology is, like subword tokenization, a kind of compression
   a. Frequent words are encoded as single symbols (e.g., token IDs)
   b. Infrequent words are encoded as series of symbols
   c. The less frequent the word, the greater the number of symbols

(10) Other things being equal, symbols correspond to morphemes/signs
   a. Frequent words are encoded as unitary signs
   b. Infrequent words consist of composed signs
   c. Models/humans generate new signs by adding morphemes to bases with similar representations to those with which they occur in the training data/linguistic experience.

(11) A morphological construction/affix will be productive just in case it occurs in many, relatively low frequency word types.

References

Jean Berko. The child's learning of English morphology. *Word*, 14(2-3): 150–177, 1958.

Natalia Yulievna Shvedova et al. *Russkaya grammatika [Russian grammar]*. Moscow: Institute of the Russian Language, Russian Academy of Sciences, Moscow, 1980.