Glossa a journal of general linguistics

Pham, Mike and Jackson L. Lee. 2018. Mincing words: Balancing recovery and deletion in word truncation. *Glossa: a journal of general linguistics* 3(1): 36. 1–19, DOI: https://doi.org/10.5334/gjgl.269

RESEARCH

Mincing words: Balancing recovery and deletion in word truncation

Mike Pham and Jackson L. Lee
University of Chicago, 1115 East 58th street, Chicago, IL 60637, US
Corresponding author: Jackson L. Lee (jsllee@uchicago.edu)

Brazilian Portuguese exhibits word truncation: e.g., the word *cruzeiro* 'cruise' results in the truncated form *cruza*, where the vowel *-a* is added to the truncated stem *cruz-*. Gonçalves (2011) claims that truncated words preserve the onset of the rightmost syllable of the first binary foot. We argue from a corpus-based perspective instead that the truncated stem is better predicted by optimizing two opposing forces: original word recovery and phonological deletion. These are formalized and implemented as *right-complete counts* (RC) and *left-complete counts* (LC), based primarily on the analysis of blends and subtractive word formation in Gries (2006) and taking into consideration the informativity of the deleted material as well as the preserved material. Specifically, a model incorporating both RC and LC outperforms one that uses only one or the other, as well as prosodic models based on binary feet, in predicting truncated stems in Brazilian Portuguese. Beyond truncation, our model has implications for morpheme segmentation as well as the mechanics of morphological reanalysis.

Keywords: truncation; recoverability; optimization; segmentation; morphology; Brazilian Portuguese

1 Introduction

Many languages exhibit truncation (or clipping), whereby a word is shortened, and considerable research is on the prediction of the truncated output of a given word (Kreidler 1979; 2000; Katamba 2005; López Rúa 2006; Gries 2006). This paper concerns Brazilian Portuguese, whose truncations Scher (2012) organizes into four different types (1). She mentions that truncation in Brazilian Portuguese is associated with an evaluative, appreciative reading, though, like her, we consider this semantic contribution of truncation to be beyond the scope of this paper.

Type 1 truncations are formed by taking the initial morpheme in the full form and deleting the following material; Type 2 truncations retain part or all of the root from the original full form, ending in a vowel from this original root; Type 3 truncations are similar to Type 2, with the difference being that the part of the root that remains ends in a consonant, followed by insertion of *-a*; Type 4 truncations are identical to Type 3, with the difference being that the inserted suffixal material is either *-as* or *-(i)s*. Except for Type 1, truncation may occur at an intra-morphemic point. This observation is central to a morpheme-agnostic perspective on truncation, which will become clear as the paper unfolds.

(1) Four types of truncation in Brazilian Portuguese (Scher 2012)
  a. Type 1: preserve first morpheme
    i. psicologia, 'psychology' → psico
    ii. odontologia, 'dentistry' → odonto
    iii. fonoaudiologia, 'speech therapy' → fono
  b. Type 2: preserve (part of) root
    i. prejuízo, 'loss (of money)' → preju
    ii. bijuteria, 'bijou' → biju
    iii. depressão/deprimido, 'depression/depressed' → deprê
  c. Type 3: preserve (part of) consonant-ending root and append *-a*
    i. cerveja, 'beer' → cerv-a
    ii. vagabunda, 'slut' → vagab-a
    iii. cruzeiro, 'cruise' → cruz-a
    iv. burgês, 'burgess' → burg-a
  d. Type 4: preserve (part of) consonant-ending root and append *-as/-(i)s*
    i. saudades, 'homesickness' → saud-as
    ii. bermuda, 'shorts' → berm-as
    iii. bobeira, 'silliness' → bob-(i)s

While truncation is not restricted to preserving initial material and deleting final material, this pattern forms the majority of truncation in English and other languages (Mattiello 2013). This tendency is reflected in our data, where there were not enough examples of right-anchored truncation, or other types of truncations that preserve some intra-word material, for a thorough analysis of all forms of truncation. Dressler (2005) shows that the beginning of a word is more salient, which likely strongly influences the predominance of left-anchored truncation. Due to data limitations, we restrict the scope of our paper to only those truncations which preserve left-edge material and delete right-edge material.

In this paper we distinguish the two terms *truncated form* (TF) and *truncated stem* (TS). We refer to the entire word on the right side of the arrow in (1) as the truncated form, which comprises a truncated stem (TS), possibly plus a theme vowel: For example, in *cerveja*, the truncated stem is *cerv-* to which *-a* is suffixed (making it Type 3). Evidence for *-a* being suffixed in these truncation types, rather than being retained from the end of original word can be seen in (1c–iii) and (1c–iv), where the original words do not contain an /a/ segment – *cruzeiro* and *burgês* – yet the attested truncated forms contain a final *-a*: *cruza* and *burga*, respectively. For the scope of this paper, we consider this to be evidence that truncated forms can be derived from truncated stems by appending suffixal material, though we do not make any claims for whether this final *-a* is a morphological suffix or a phonological repair for illicit consonant-final words.

Our focus is to model the derivation of the TS from the original word (i.e., *cerv-* from *cerveja*), rather than the full TF (i.e., *cerva*). Note, however, that the TS and TF can be identical, as in the case of Type 1 and Type 2 truncations, where a word such as *bijuteria* (Type 2) derives the TS *biju* and the identical TF *biju*. For Types 3 and 4 where the TF differs from the TS, we assume the derivation of the TF to generally be handled via independent morphophonological processes operating on a derived TS, which may be connected to gender, such as the *-o/a* in *amig-o/amig-a* 'friend' or phonotactic restrictions, such as Portuguese word-final consonants being limited to *-s* and *-r* (as well as orthographic *-l*).[1] As our approach does not make reference to a priori morpheme

[1] Thanks to an anonymous reviewer for pointing out this phonotactic observation of Portuguese.

boundaries, we do not make further distinction between the different types of truncation in Brazilian Portuguese.

Previous approaches to truncation in Brazilian Portuguese have either been phonological (Belchor 2006; 2009; Gonçalves 2006; 2009; 2011; Gonçalves & Vazquez 2004) or morphosyntactic (Scher 2011; 2012). Gonçalves (2011) is an example of the former, where there is a phonological process that drops part of the last foot of the original word, but preserves that foot's onset. As Scher (2012) points out, this does not account for data such as the TF *adrena* from the original word *adrenalina*, where the onset of the last foot's first syllable (the /l/) is not preserved. More generally, analyses that are strictly phonological have difficulty accounting for when onsets are preserved in a TS, resulting in a trisyllabic TF such as *vagaba* from *vagabunda*, and when they are deleted, resulting in a disyllabic TF such as *cerva* from *cerveja*. These theories by themselves do not distinguish when these onsets should be preserved or deleted during truncation.

Alternatively, Scher (2012) derives the TF of Type 3 and Type 4 words by decomposing the morphological structure of the original word. For example, she analyzes the TF *cerva*, from *cerveja* (1c–i), as having the following morphosyntactic structure: √CERV-*ej*-*a*. In her analysis, √CERVEJ- is further decomposed by providing data that show that -*ej*- (along with -*am*- and -*at*-) are unrelated suffixes in other contexts – essentially a reanalysis account based strictly on phonological identity to another morpheme (not unlike the tongue-in-cheek English example *history* > *his-tory* > *her-story*).

While this presents a strict environment in which truncation in Brazilian Portuguese can take place, Scher's account is problematic within the Distributed Morphology framework she utilizes, where Late Insertion prevents any phonological material from being visible within the same Spell Out domain. In other words, Scher's analysis depends on morphological reanalysis based on homophony. However, phonological material in Distributed Morphology is not inserted into the structure until after (morpho) syntactic derivations have already happened. As such, nothing in the morphosyntactic structure should be sensitive to phonology prior to Vocabulary Insertion. Given this, it should not be possible in Scher's approach for the morphosyntactic reconfigurations associated with reanalysis – i.e., insertion and projection of a new functional head within the noun – to be sensitive to phonological identity unless we claim that there is a word-internal Spell-Out domain (or phase).

We take a different approach, one that is based on segments (rather than larger prosodic units or morphemes) as well as generalizations induced from data distributions. Our approach models TS derivation as optimizing two opposing forces: maximal deletion and maximal recoverability of the original word. The speaker deletes as much of the original form as possible while ensuring that the hearer has enough material in the TS to successfully recover the original form. Under this model, *vagabunda* produces the TS *vagab*-, which is the point at which the most original phonological material has been deleted without overly hindering recovery of the original word; the potential TS **vagabu*- can undergo further deletion, while the potential TS **vaga*- has not preserved enough material to make the original word reasonably recoverable.

The two opposing constraints – deleting as much material as possible and maintaining ease of word recovery – are an extension of Gries (2006), who provides an analysis of subtractive word formation processes based on uniqueness points, the point of a word at which it can be uniquely identified from a set of candidate words, and recognition points, the point of a word at which a majority of speakers can recognize it with high probability. Similarly, recoverability of the original word from derived preserved material has been used in the analysis of blends (Gries 2004; Cook 2010): e.g., *brunch* from *breakfast* and *lunch*. However, while these models incorporate the idea of recoverability of the original

source word(s) based on their string similarity to the output word, they do not take into account optimization of deleted material in conjunction with the original word's recovery. For Kemmer (2003), there is competition between recoverability and the prosodic similarity of the output word to its source word that is balanced in the process of blending. This input-output similarity, however, cannot be a motivating factor for deletion in truncation, as the output truncated form by definition must be phonologically and prosodically smaller than the original source form. Rather than motivating deletion as an indirect means of maximizing original word recoverability via preserving prosodic similarity, we view deletion of phonological material in truncation to be independently motivated as removal of substrings with low informativity. In the following, we show that a truncation model that incorporates both maximal deletion and maximal recoverability of the original word outperforms a model that has one but not the other.

While we treat each model of truncation, including approaches in previous literature, as being independent distinct models, we believe that a more complete analysis from original word to truncated form will be influenced by all the factors discussed within this paper as well as in the previous literature. For instance, phonotactic and morphological features can likely help to explain consonant cluster preservation and sensitivity to morpheme boundaries; concrete examples will be discussed in §5. However, we focus on the truncation tactics independently to demonstrate the strong influence recoverability and deletability have in truncation, when not augmented with other grammatical considerations. We leave a more comprehensive model of truncation for further research.

The remainder of this paper is organized as follows. In §2 we describe our methodology and the different models of TS prediction under consideration. In §3 we provide the results of each model on a gold standard list of nouns with attested TFs in Brazilian Portuguese and their evaluation in §4. In §5 we discuss why a truncation model that combines maximal word deletion and recoverability outperforms the other models under consideration, and provide a more general outlook on our work's implications for morphological segmentation and reanalysis. We conclude in §6.

2 Methodology

In this section, we discuss our methodology – first by defining what right-completes and left-completes and their respective counts mean in our models, then by elaborating on our data source, and finally by outlining the seven models of truncation in Brazilian Portuguese that we construct and evaluate in this paper.

2.1 Right-completes and left-completes

We borrow and modify Gries's (2006) concept of recovery points to predict the optimal truncation point of Brazilian Portuguese nouns with attested truncated forms. Central to our work is the notion of a *complete* of a string *s*:

(2) **Complete of *s*:** an entire word in a lexicon that can be formed by concatenating a string of symbols to a given string *s*. For example, if "abcde" is a word in the lexicon, then it is a complete of the string "abc", as it can be formed by concatenating the suffixal string "de" to the given string, *s*, "abc". Similarly, "abcde" is also a complete of the string "de", as it can also be formed by concatenating the prefixal string "abc" to the given string "de". For brevity, we use the term *complete* with the implicit understanding that it is always relative to a specific string *s*.

We further specify two types of completes: *right-completes (R-completes)* and *left-completes (L-completes)*. R-completes are the subset of completes that can be formed by concatenating a string to the *right* of a given string – e.g., words formed from attaching

suffixes. L-completes are the subset of completes that can be formed by concatenating a string to the *left* of a given string – e.g., words formed from attaching prefixes.

(3) Let a language L be a set of strings w:
   a. The right-completes of a string p are the set $RC_p = \{w \in L : \exists\ s$ such that $w = ps\}$. For instance, *spree* is an R-complete of "spr", as it is formed by concatenating "ee" to the right of "spr".
   b. The left-completes of a string s are the set $LC_s = \{w \in L : \exists\ p$ such that $w = ps\}$. For instance, *spree* is an L-complete of "ee", as it is formed by concatenating "spr" to the left of "ee".

Given the above definitions, we can define the *R-complete count (RC)* and the *L-complete count (LC)* as the following:

(4) a. **R-complete count (RC):** the number of R-completes in a lexicon for a given string
   b. **L-complete count (LC):** the number of L-completes in a lexicon for a given string

Another way of thinking about RC and LC is that RC is the number of words in a lexicon that begin with a certain string, and LC is the number of words in a lexicon that end with a certain string. We note that this way of characterizing RC and LC makes them related to the much earlier work by Harris (1955) using successor and predecessor frequencies for word and morpheme boundary discovery. Our present work differs in that RC is the number of *words* that begin with a given string, whereas Harris's successor frequency is the number of *symbols* (phonemes or letters) that begin a given string instead; the same contrast applies to LC in this paper and Harris's predecessor frequency.

2.2 Data

Our data source comprises two main components. The first is a Brazilian Portuguese lexicon of about 750,000 word types (from a corpus of about 340 million word tokens).[2] The second is a set of 107 gold standard nouns with attested TFs that were pulled primarily from data in Scher (2012) and the appendix of Vilela et al. (2006), with additional data added from personal communication with a native speaker consultant. Proper names were excluded for the divergent and highly idiosyncratic possibilities in truncations (e.g., *Elizabeth* → *Eliza*, *Liz*, *Beth*, *Betsy*, etc.), as were the relatively few TFs that were not aligned with the left edge of the original word. Restricting ourselves to only considering left-aligned truncation is a practical matter, and we leave a more thorough investigation of more truncation types to future research.

All datasets were used as-is in their ordinary Brazilian Portuguese orthography, as the language has fairly high grapheme-to-phoneme correspondence (compared to, say, English). To be sure, some digraphs such as "ch", "lh", and "ss" have consistent grapheme-to-phoneme mappings and could have been replaced. Also, there are graphemes such as "gu" (for /g/ or /gw/) and "c" (for /s/ or /k/) that might have been handled in some way. However, the issue of whether a more phonetic dataset or a more orthographic one should be used is not trivial. In addition, as we use large datasets (e.g., the lexicon with 750,000 word types), there would be practical issues for how to, for example, efficiently

[2] The Brazilian Portuguese lexicon with word frequency information is from https://github.com/hermitdave/FrequencyWords. It is derived from a corpus of movie subtitles from http://www.opensubtitles.org/ – highly representative of the spoken language.

replace "gu" with /g/ or /gw/ (which cannot be straightforwardly automated because of ambiguity) in a huge amount of words.[3]

2.3 The tested models

In this paper, we test various models of truncation against each other in order to determine which most accurately predicts the attested TSs in Brazilian Portuguese. The seven models tested are (i) the RC-only model; (ii) the LC-only model; (iii) the RC + LC combined model; (iv) the right to left binary foot model (binRL), following Gonçalves (2011), who observed that TFs in Brazilian Portuguese preserve up to the onset of the second syllable of the first binary foot, building feet from the right; (v) the left to right binary foot model (binLR), which predicts TSs to terminate before the second vowel from the left; (vi) the algorithm by Gries (2006); (vii) a baseline model by random sampling. The first two models only consider RC or LC independently, while the RC + LC model predicts TSs by looking at both RC and LC simultaneously. §3 further elaborates on the exact mechanics of these corpus-based models (the RC-only, LC-only, RC + LC combined, and Gries models). Below we discuss the two binary foot-based models as well as the baseline model.

2.3.1 The binRL model

The right to left binary foot model (binRL) is included as an implementable interpretation of Gonçalves (2011), as summarized by Scher (2012). This is the most charitable prosodic foot-based model that can account for both disyllabic and trisyllabic TFs in BP truncation. For instance, our gold standard list shows that both types of TFs are attested: *cerveja* → *cerva* (disyllabic); *vagabunda* → *vagaba* (trisyllabic). This variation between disyllabic and trisyllabic TFs presents a potential problem for a model that derives TSs from binary feet: If the feet are constructed from the left edge rightwards, however the rules or constraints are formulated, they will favor either a TS that results in a disyllabic or trisyllabic TF, but not both. This is because the feet on the left edge of the original word are created procedurally before any rightwards material, meaning that word length can have no effect on the leftmost feet.

The binRL model builds binary feet from the right edge of the word. It is able to derive either disyllabic or trisyllabic TFs by preserving all but the final rhyme of the first binary foot in addition to a potential defective non-binary foot on the left edge of the original word. In fact, this is the only way in which a model based on prosodic binary feet can derive both disyllabic and trisyllabic TFs. Consider how the TSs of the words *baterista* 'drummer' and *Bermuda* 'shorts' are handled by the binRL and binLR models:

(5) binRL: disyllabic and trisyllabic TFs
   a. *baterista* → (*ba.te*).(*ris.ta*) → predicted TS/TF = **bat/*bata*
      (actual TF = *batera*)
   b. *bermuda* → *ber.*(*mu.da*) → predicted TS/TF = **bermud/*bermuda*
      (actual TF = *bermas*)

(6) binLR: only disyllabic TFs
   a. *baterista* → (*ba.te*).(*ris.ta*) → predicted TS/TF = **bat/*bata*
      (actual TF = *batera*)
   b. *bermuda* → (*ber.mu*).*da* → predicted TS/TF = *berm/*berma*
      (actual TF = *bermas*)

[3] We have experimented with the arguably less controversial replacements for digraphs (ch → S, lh → L, nh → N, ss → s, rr → R), available as an option for all tested models of truncation in our code (footnote 6). The results bear no qualitative differences from those reported in this paper.

As (5) shows, because the binRL model builds binary feet from the right, it is able to derive TSs that result in both disyllabic TFs (5a) and trisyllabic TFs (5b). In (5a), the leftmost foot *(ba.te)* is a binary foot, and so the cut is made before the vowel of the second syllable, resulting in the predicted TS *bat*. In the case of *bermuda* in (5b), the initial syllable *ber* is a defective non-binary foot, and so it is ignored; instead, the model looks to the leftmost binary foot *(mu.da)* and makes the cut before the vowel of the second syllable, resulting in the predicted TS *bermud*. Looking at (6), we can see that the binLR model is unable to predict any TSs except those that result in disyllabic TFs. This is because it builds binary feet from the left, and will thus always make its cut before the vowel of the second syllable of the word. In this case, it accurately predicts that the TS of *bermuda* is *berm*, but inaccurately predicts that the TS of *baterista* is **bat* (attested TS is *bater*).

The binRL model makes a strange prediction about how disyllabic and trisyllabic TFs occur: In this model they are ultimately based on whether the syllable count of the original word is odd or even. Words with an even number of syllables will derive disyllabic TFs and those with an odd number of syllables will derive trisyllabic TFs. Crucially, this analysis does not reference word length in any way. This results in an undesirable prediction that the length of a TF should vary back and forth between having two or three syllables as its original word gains syllables.

2.3.2 The binLR model

While the binLR model is not based on any previous literature on truncation in Brazilian Portuguese, we included it to contrast the binRL model with another prosodic binary foot-based analysis of Brazilian Portuguese truncation that appears to correctly derive attested TSs with some degree of accuracy. The binLR model, then, essentially considers truncation that produces *disyllabic* TFs to be the default pattern, and successfully derives those cases, but has nothing to say about the trisyllabic TF cases. This would be somewhat analogous to formulating an elsewhere rule and assuming that the exceptions to this rule form a minority of the empirical data, and can be handled by more specific rules; only in our case, we have not included what these more specific rules that can account for trisyllabic TFs are within the binLR model.

2.3.3 The baseline model of random sampling

For the purposes of comparison, we implemented a baseline model based on random sampling of the true truncation points. For each word in the gold standard list, the normalized true truncation point was computed. For example, if the true truncation point for TS is after the second segment in a 10-segment word, then the normalized true truncation point is $2/10 = 0.2$, where 2 is the length of the attested TS and 10 is the number of segments of the original word.[4] The mean and standard deviation of all the normalized true truncation points were calculated, which are 0.57 and 0.12, respectively. For each word in the gold standard list, the truncation point predicted by random sampling is the mean of 10,000 random samples drawn from the normal distribution $N(0.57, 0.12^2)$.

3 Results

In this section, we explain and discuss the results of the four corpus-based models: the RC- and LC-only models, the RC + LC combined model, and the Gries model.

[4] The implementation was based on a suggestion from a reviewer, for which we are grateful. We acknowledge that random sampling of this sort depends on the representativeness of the data, which could be improved in further work with a larger, more carefully collected list of gold standard words.

For each word in our gold standard list, we calculated the RC and LC values for each potential stem, as described here. We considered every potential TS derived from iteratively deleting right-edge material. Given a non-truncated word of length *n*, all left-aligned substrings of lengths {1, 2, ..., *n* – 1} are considered potential TSs. For each potential TS, we calculated its RC and LC. The RC is the number of words in the Brazilian Portuguese lexicon ending with the given potential TS, whereas the LC is the number of words in the lexicon that begin with what was deleted to form the given potential TS. For example, given a potential TS **vagabun* from *vagabunda*, RC is the number of words in the lexicon that begin with the string *vagabun*; the corresponding LC would be the number of words that end with the string *da*. The results can be tabulated as follows for each word. The log-transformed counts are also provided due to highly skewed distributions in lexical statistics (cf. Baayen 2001).

In Table 1, the top row shows the original word, with the symbols comprising the attested TS in capital letters. For RC, the number in each column shows the RC value for each potential TS formed from the symbols to the left of and including the symbol heading that column. As can be seen in the column headed by "V", there are 17,979 words in the lexicon that begin with the string *v*; there are 4,393 words that begin with the string *va*; 315 beginning with *vag*; etc. LC values are the reverse: Starting from the right edge, there are 107,925 words in the lexicon that end with the string *a*, 11,171 that end with the string *da*, 1,019 with *nda*, and so on.

Plotting the log(RC) and log(LC) values provides a graph; see Figure 1. Starting from the left edge and moving rightwards, RC begins with a high value and declines as the potential TS gets longer. LC mirrors this: It begins with a high value on the right edge and declines moving leftwards as the potential deleted material gets longer. The number of words that contain a given string declines as that string gets longer. Put another way, looking left to right, the RC is monotonically decreasing and the LC is monotonically increasing.

For the RC- and LC-only models, the potential TS is calculated by finding the "elbow point" along the respective curves, or the point of maximal curvature where the maximum change of the curve occurs; this is computed as the point with the greatest second derivative value for each curve. For our purposes of modeling TS prediction, we interpret this point as one where original word recovery will begin to become more difficult proportionate to the greater number of possible RCs or LCs corresponding to that point. To be more concrete, consider the RC-only model: Looking at the raw RC values in Table 1, we can see that there are 8 possible words in the corpus that begin with *vagabunda*, meaning that the speaker can be fairly certain that the hearer can recover the original word based on this TS, as there are relatively few options. Deleting the final *a* results in the potential TS *vagabund*, which has 40 R-completes; the jump from 8 to 40 is relatively minute on the scale of lexicon entries. What is more significant is the point at the symbol *G*, where the number of R-completes goes from 315 to 4,393 at the first *A* symbol. This increase is much more abrupt than those associated with the symbols to the right; one might expect a speaker only looking at RC to consider this the TS that best

Table 1: RC and LC values for *vagabunda*.

	V	A	G	A	B	u	n	d	a
RC	17979	4393	315	129	57	47	41	40	8
log (RC)	4.25	3.64	2.5	2.11	1.76	1.67	1.61	1.6	0.9
LC	2	2	2	12	28	137	1019	11171	107925
log (LC)	0.3	0.3	0.3	1.08	1.45	2.14	3.01	4.05	5.03

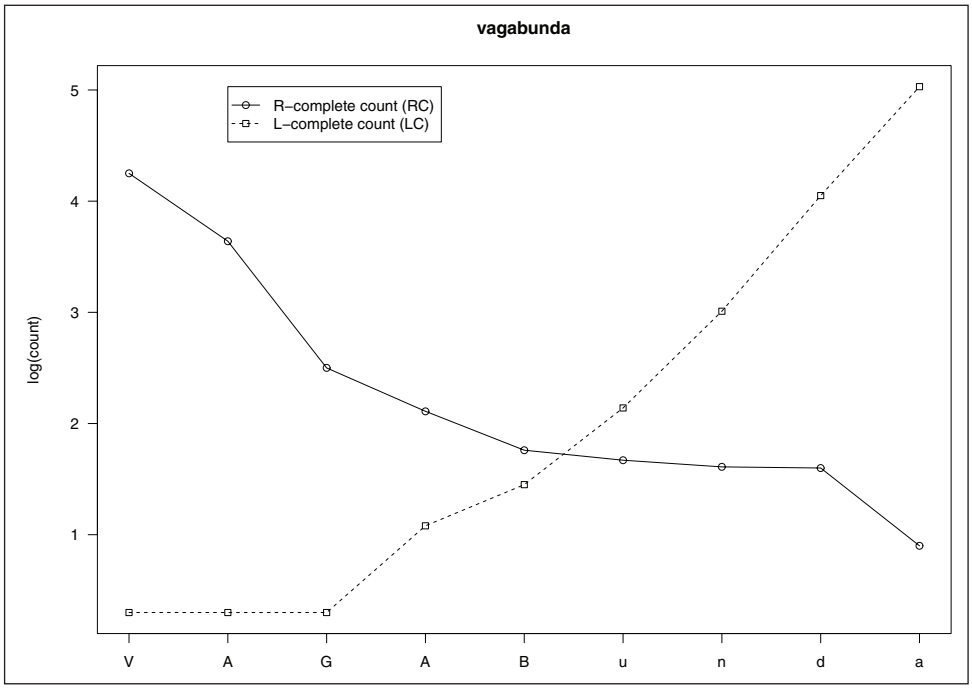

Figure 1: Log-transformed R- and L-complete counts of *vagabunda*.

optimizes deleting right-edge material without introducing too many possible words that the TS can be reconstructed as.

The log(LC) curve more or less mirrors the log(RC) curve in terms of general shape. Rather than representing the number of possible reconstructions of a potential TS, as RC does, the log(LC) curve represents the number of words that also end in the deleted material. Another way of looking at this is to say that RC provides a metric for the informativity of the preserved material while LC provides a metric for the informativity of the deleted material. As such, right-edge material that has a high LC value, such as the final *a* in Table 1 can be seen as relatively uninformative, as there are 107,925 words that end with that. Using the graph, then, allows us to predict the point where the optimal amount of right-edge material can be deleted. The elbow point of the log(LC) curve is exactly this point, where high-frequency material to the right of that point can easily be deleted, while the relatively more informative material to the left of that point will have greater resistance to deletion. For *vagabunda* (whose true truncation point is "b", the fifth letter), both the computed RC and LC elbow points are "g" (the third letter).

For the RC + LC model, we consider both the log(RC) and log(LC) curves together instead of considering them separately. Rather than directly considering the elbow points of each curve to predict the TS, this model predicts the TS to preserve material from the left edge to the symbol that has the minimum difference between RC and LC. The symbol at which the minimum difference between RC and LC is attained is mathematically equivalent to the symbol closest to the intersection of the two curves. As can be visually identified in Figure 1, the RC + LC model predicts the letter "b" (closest to where the RC and LC curves intersect) to be the truncation point, which is also the true truncation point.

With the models based on RC and LC explained above, we are ready to introduce the last model implemented, which is the algorithm by Gries (2006) for estimating truncation (as observed in blends or truncations more generally). The Gries algorithm is similar to the RC model, as the latter is an adaptation of it, in terms of looking at every possible left-aligned

Table 2: Type and token frequencies of words beginning with beginnings of *agitation*, based on Table 2 in Gries (2006: 543).

Potential TS	Number of right-completes	Frequency rank of *agitation* among these right-completes
a	4,347	595
ag	137	24
agi	12	1
agit	8	1
agita	8	1
agitat	8	1
agitati	3	1
agitatio	2	1
agitation	2	1

TS and checking what the right-completes are for each potential TS. The crucial difference between the two models is that our RC model derives the predicted truncation point by the elbow point computation, as explained above, whereas Gries makes use of word token frequency information instead. To concretely illustrate Gries's method, Table 2 shows his example of English *agitation*.

In Gries's method, each potential TS is associated with its set of right-completes. These right-completes each have their word frequency information available. For each potential TS, we check the frequency rank of the original word in question (*agitation* here) among the associated set of right-completes. The method takes as the predicted TS (i.e., Gries's "selection point") the shortest potential TS where the original word is the most frequency word among the relevant right-completes. For *agitation* in Table 2, the potential TS "a" has 4,347 right-completes, with *agitation* ranking 595th among these right-completes for word frequency. As we consider longer potential TS, the number of the right-completes in question decreases while the original word remains in the set of right-completes and its frequency rank climbs. As Table 2 shows, the potential TS "agi" is the shortest one where *agitation* is the most frequent word among the relevant right-completes, and is considered the predicted TS in this method.

4 Evaluation

4.1 Overall accuracy

The basic evaluation metric of the seven models is to compare the percentage of TS accurately predicted by each model, as in Table 3 (the best result is bolded).

At first glance, we can see from Table 3 that while no model is perfectly accurate at predicting TSs, the relative accuracies are quite clear. The RC + LC model is the most accurate of all the models tested. Its accuracy of 43% is higher than the accuracy of 37.4% for the baseline model by random sampling. Because accuracy is a rather crude measure, in that just one segment off from the true truncation point makes a predicted TS categorically wrong, we examine the TS truncation results in greater detail with more refined metrics in the following.

4.2 Distance errors

Beyond the cursory measure of accuracy, we use an array of more detailed metrics based on *distance error*, as defined below, for understanding the nuanced picture of TS predictions by the models:

(7) **Distance error:** the number and direction of symbols between the attested truncation point and the predicted truncation point.

Consider the example *metaleiro* 'metalhead (fan of metal music genres)' with the TS (and TF) *metal*. Table 4 shows the RC and LC values for *metaleiro*, whereas Figure 2 plots the log-transformed values.

In order to acquire a numeric value for the termination point (or length) of the attested and predicted TSs, we assign each symbol an integer equal to the length of the potential

Table 3: Percentages of TSs accurately predicted.

Model	% correct
RC	24.3
LC	24.3
RC+LC	**43.0**
binRL	25.2
binLR	32.7
Gries	21.5
Baseline	37.4

Table 4: RC and LC values for *metaleiro*.

	M	E	T	A	L	e	i	r	o
RC	50090	8254	1104	387	88	4	3	3	2
log (RC)	4.7	3.92	3.04	2.59	1.94	0.6	0.48	0.48	0.3
LC	1	1	7	26	98	1280	1767	6258	95398
log (LC)	0.0	0.0	0.85	1.41	1.99	3.11	3.25	3.8	4.98

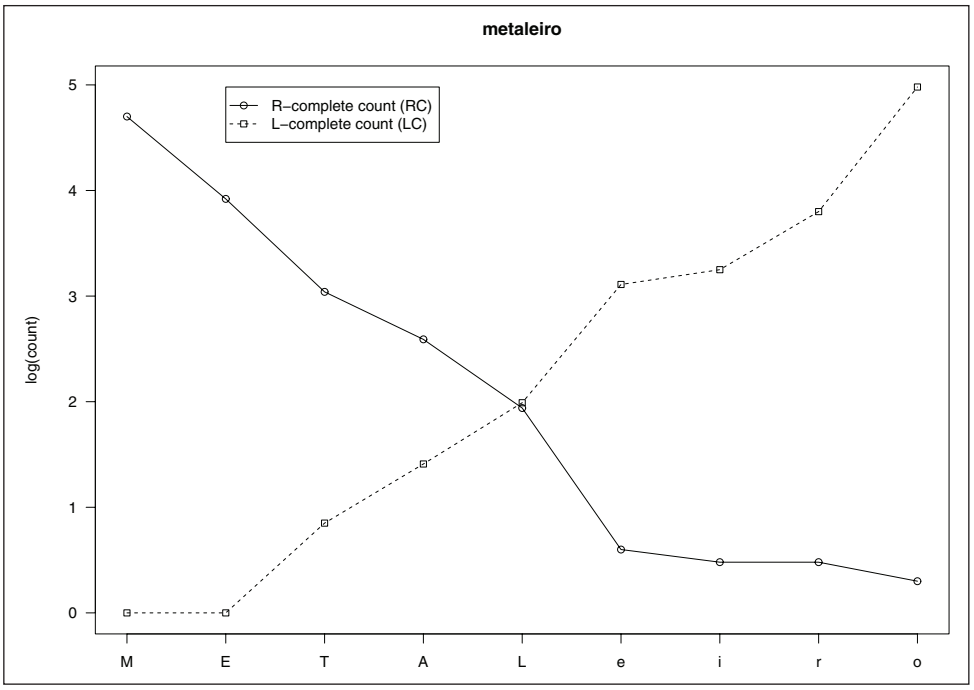

Figure 2: Log-transformed R- and L-complete counts of *metaleiro*.

TS up to and including that symbol. For example, starting with 1 on the left edge ("M"), we can see that the attested truncated stem terminates at "L", yielding a string of length 5 (*metal*).

Using these string length values for each potential TS for a given word, we calculate the distance error, *E*, for each model, using the formula below:

$$E = |TS_x| - |TS_0| \tag{8}$$

The distance error *E* is equal to the length of the attested truncated stem ($|TS_0|$) subtracted from the length of the predicted truncated stem ($|TS_x|$) for a given word. *E* carries both a sign and a magnitude. If *E* is positive, the predicted TS is longer than the attested TS; a negative *E* means the predicted TS is shorter instead. The magnitude of *E* is the number of symbols by which the lengths of the predicted TS and attested TS differ. As an example, Table 5 shows the distance errors for all possible TS of *metaleiro* (attested TS: *metal*).

Table 6 shows the attested and predicted TS and distance errors for each model for *metaleiro*.

As can be seen in Table 6, the TS prediction by both the RC + LC and baseline models has an *E* of 0. This means that both models tied as the most accurate models in predicting the attested TS for *metaleiro*, as they exactly predicted the attested TS. Compare this to the RC or Gries model, which has an *E* of 1, meaning that it predicts a TS one symbol longer (**metale*) than the attested TS *metal*. The other three models all underpredict the TS, as can be seen by the negative *E* values.

We calculate the distance error in number of symbols between the predicted TS of all the models and the attested TS of each word in the gold standard list. Doing so allows us to look at the distribution of distance errors for each model, as in the boxplot in Figure 3.

Two intuitions follow from Figure 3. First, an *accurate* model will have distance errors centered at or around zero. Second, a *consistent* model will have densely distributed errors, as opposed to sparsely distributed, "spread-out" errors. These intuitions guide us through the interpretation of the error distributions. Eyeball examination of this boxplot suggests that the RC + LC model is the best, as its distance errors are the least spread out and centered around zero. While the boxplot provides a good visual comparison of the models' performances, we would like to be able to compare them more quantitatively. For this, the models can be compared along two measures of accuracy: (i) the mean (μ), (ii) standard deviation (σ).

Table 5: *E* values for each potential TS of *metaleiro*.

	M	E	T	A	L	e	i	r	o
$\|TS_x\|$	1	2	3	4	5	6	7	8	9
E	−4	−3	−2	−1	0	1	2	3	4

Table 6: *E* values for each model's TS prediction for *metaleiro*.

	RC	LC	RC+LC	binRL	binLR	Gries	Baseline
$\|TS_x\|$	6	2	5	3	3	6	5
E	1	−3	0	−2	−2	1	0

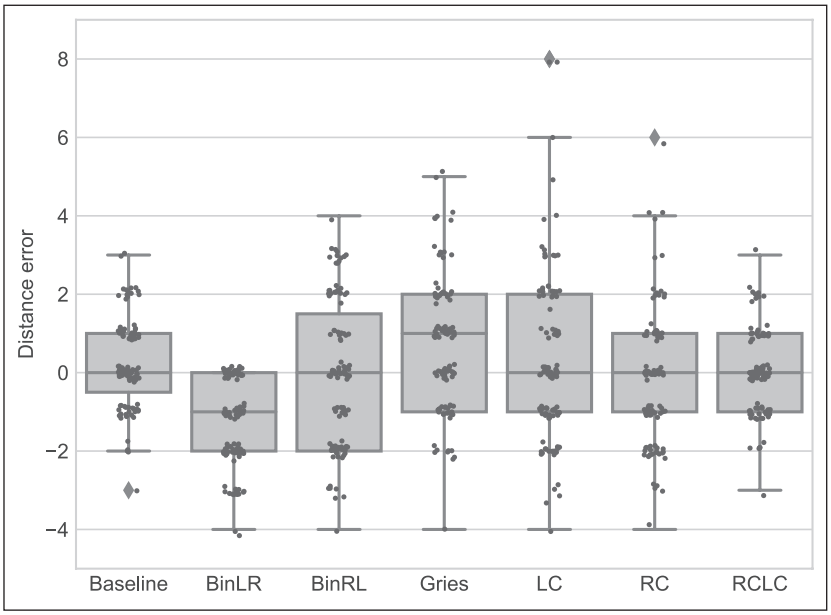

Figure 3: Error distribution of the seven models.

Table 7: Error evaluation.

	RC	LC	RC+LC	binRL	binLR	Gries	Baseline
%	24.3	24.3	**43.0**	25.2	32.7	21.5	37.4
μ	−0.12	0.30	0.04	**0.02**	−1.28	0.65	0.22
σ	1.70	2.14	**1.04**	1.82	1.10	1.70	1.12
%: higher is better			μ: closer to 0 is better			σ: lower is better	

In order to determine whether or not a model biases towards underprediction or overprediction, we calculate the mean μ of all the distance errors of a model. A better model should have the mean closer to zero. The second measure we take into account is standard deviation, σ, a measure of the spread of the distance errors. In terms of our models, a low standard deviation means that the distribution has a small spread, and that most of the values lie closer to the mean point. On the other hand, a model with a high standard deviation value is one with a large spread. A model with a lower standard deviation is more desirable for consistency in making predictions.

With these evaluation metrics of model error, Table 7 shows the performance of the seven models tested in this paper. According to the evaluation, the RC+LC model is the best performing one in terms of overall accuracy (% = 43.0) and standard deviation ($\sigma = 1.04$). Although the binRL model appears to be the best with respect to overprediction/underprediction bias ($\mu = 0.02$), the RC+LC model is very similar in this regard ($\mu = 0.04$). Because the mean involves directly summing errors, it is affected by the sign of each individual E value, and thus reflects how much a model overpredicts or underpredicts TSs on average. A mean distance error of 0 may only tell us that the particular model is just as likely to predict a TS to be too short as it is to predict it to be too long.

5 Discussion

In this section, we elaborate in §5.1 on why the RC+LC model outperforms the other models of TS prediction under consideration. In §5.2, we discuss implications of our work for morphological reanalysis.

5.1 Why right-completes and left-completes together work

Our results show that the RC + LC model is the most accurate among the tested models in predicting attested TS. The RC- and LC-only models respectively provide measures of optimal preserved material and optimal deleted material in truncation, and combining the two measures provides an intuitively – and testably, as we have shown – better result than considering either independently.

One way to interpret the underprediction of the RC model is that it is not the absolute value of the right-complete counts that matters, but their relative relationship to each other – specifically, the rate of change in their values. What this means is that while the curve's elbow point may be the optimal point of truncation from the point of view of minimizing the number of words that begin with the same string as TS, it appears that this often does not preserve sufficient material for recoverability. Better recoverability is ensured when we also maximize confidence about how the string beginning with the TS will end. If this is the case, then the actual truncation point will occur to the right of what the RC model predicts, in order to further drive down the RC value and ensure hearer recovery of the original word. This could be what we are seeing for the RC model in terms of its negative mean value (underprediction).

Another reason RC tends to underpredict could be the fact that it does not take phonotactics into account. Because the RC, LC, and RC + LC models consider candidate truncated stems segment by segment, they can predict a truncated stem that splits a consonant cluster: RC + LC predicts the TS **buroc* instead of *burocr* for *burocrata* 'bureaucrat'. If there were a phonological preference to preserve consonant clusters – as is the case in English Pig Latin, for example, which derives *ate-skay* from *skate* rather than **kate-say* – then we would expect fewer RC underpredictions. This is evidence that a more complete model of truncation needs to incorporate knowledge of phonotactics. Because the LC model is essentially the RC model in reverse, we expect that it should have a positive mean rather than a negative one. This is indeed the case, with $\mu = 0.30$, meaning that the LC model tends to overpredict TSs.

It is also instructive to compare the results between the RC-only and LC-only models. In general, the LC model fares worse than the RC one, based on the measures of the mean error and standard deviation in Table 7. In other words, if one had to pick only RC or LC to pay attention to in truncation, then RC would be a better choice. This is reasonable, as the RC model is a measure of recovery. After all, if the original is not recoverable from a TS, then any pragmatic or semantic effect triggered by deletion is moot. Given that the beginning of words are more salient (Dressler 2005), this preference of recovery over deletion is likely to be an important factor in why truncation primarily preserves from the left and deletes from the right cross-linguistically (Mattiello 2013): If the left-edge is asymmetrically more salient than the right edge of a word, then preserving the leftmost material of a word in truncation will independently maximize recoverability of the original.

Another aspect worth noting about comparing RC and LC is that, with respect to *vagabunda* in Figure 1, both the RC and LC models predict its TS to be **vag* instead of the attested *vagab*. This suggests that optimizing RC and LC is not about balancing the two models by splitting the distance, so to speak, between their predicted truncation points. Optimizing between RC and LC does not depend as much on their individual elbow points as it does on where the two curves intersect.

The binLR model occasionally makes accurate predictions when the TF is bisyllabic, but consistently fails to capture the fact that not all truncation results in a bisyllabic form. Rather, it seems to be an analysis of the minimal possible TS, analogous to a minimal prosodic word requirement.

Although the Gries model is the inspiration for the RC model, it tends to overpredict the TS (the RC model tends towards underprediction), with a positive μ (0.65). The reason for overprediction by the Gries model is potentially due to the fact that Brazilian Portuguese is inflectionally more complex than English (the language under study in Gries 2006), with many of these inflectional morphemes being suffixes. Because there are relatively more candidate word forms that share the same root but differ in suffixal material, the predicted TS within the word where the original word is sufficiently unique and recoverable tends to be longer. A relatively inflectionally robust language like Brazilian Portuguese has more word types that share the same root than a relatively inflectionally impoverished language like English; if these inflections are primarily suffixal, we expect to see a Gries-style model overpredict due to the existence of more candidate word types that share phonological material until the right edge of the word. The Gries model also makes use of word token frequency information, something that is important for avoiding issues associated with the pure use of word types for a language like Brazilian Portuguese with more complex inflectional morphology.[5] Further research could possibly implement a model of truncation that takes into account the opposing influences of RC and LC, as well as the interaction of word token frequencies and word type ranks as Gries's analysis does.

Finally, for the given dataset, the performance of the best model, RC + LC, is still far from being perfect. This is likely due to the fact that the RC + LC model is purely based on the segments, and does not have phonotactic and morphological knowledge at all. An examination of the nouns for which the RC + LC model made the incorrect TS predictions reveals that this is indeed the problem for quite a number of cases.

For instance, *extraordinário* 'extraordinary' has the attested TS (and TF) as *extra*, but was incorrectly predicted to have the TS **extraord*- by the RC + LC model, an error that morphological information might have helped avoid. As it stands, the RC + LC model tends to disprefer a TS like "extra" due to the relatively high number of right-completes following a morphological prefix. Also, the complete ignorance of the consonant-vowel distinction as well as phonotactics has also led to other errors. The case of TS **buroc*- for *burocrata* (attested TS: *burocr*-) discussed above is an example of incorrectly splitting a consonant cluster. *Travesti* with the predicted TS **trave*- (attested: *trav*-) is an example where a better model of truncation would potentially benefit from a preference of making a cut immediately after a consonant rather than a vowel. Further research with the higher-order goal of a more comprehensive model of truncation would likely have to address these issues.

5.2 Implications for morphological reanalysis

The goal of our model is to find the optimal truncation point within a given word. This can be alternatively seen as a model of morphological reanalysis: i.e., how do phonologically similar sequences within other words affect where a speaker creates an internal boundary within a word? Importantly, our RC + LC model makes no a priori assumptions about the internal structure of the words it looks at; it treats all word forms as being monomorphemic at the outset and decides where the optimal boundary should be.

Another characteristic is that unlike other computational models of morphological segmentation (see Goldsmith et al. 2017), our model of truncation does not assume

[5] In our models involving RC and LC, it is the word types that are counted, and it is legitimate to ask if word token frequencies should be incorporated in these models. A variant of these models is available in our code (footnote 6) where each word type is not counted as 1 (by default) but as log(token frequency of that word type) instead; if this latter number is zero (log(1) = 0, for a word type whose frequency is 1), use 0.1 instead. The results are not qualitatively different from those reported in this paper.

morpheme consistency. It determines a single morpheme boundary independently for each word: If some substring X is predicted to be a(n optimal) subpart for a given word, then the identical substring X may not be considered an optimal subpart for another word. Instead, a morpheme boundary within a word can be created based on comparing it to other words in the lexicon, as the model is only looking at the segments of words. Our approach, then, can be interpreted as a way of potentially modeling reanalysis.

To see the connection between our work and reanalysis, we return to the analysis of Brazilian Portuguese truncation by Scher (2012) discussed in the introductory section of this paper. In Scher's analysis, she implicitly assumes that reanalysis occurs based on phonological similarity. Consider the following words and their proposed morphological decomposition:

(9) a. *cerveja*, 'beer' > √CERV-ej-a
b. *pijama*, 'pajamas' > √PIJ-am-a
c. *burocrata*, 'burocrat' > √BUROCR-at-a
(cf. (20) in Scher 2012)

In each of the examples in (9), "the forms *-ej-*, in *cerveja*, *-am-* in *pajama* or *-at-* in *burocrata* are not supposed to be considered separate morphemic units in these words" (Scher 2012). Scher proposes that speakers of Brazilian Portuguese are treating these pieces as derivational suffixes based on their surface similarity to other, independent derivational suffixes in the language: *-ej* is a diminutive suffix, *-am* is a collective suffix, and *-at* is a nominalizing suffix. Essentially, Scher is claiming reanalysis of a single morpheme into a new, decomposed morphological representation based on phonological similarity. She extends this claim to other phonological sequences as well, besides the three mentioned above: *-un*, *-und*, *-ul*, *-ar*, *-et*, and *-ab*.

By looking at RC and LC in this paper, we are able to provide strongly empirical motivation for Scher's analysis of morphological decomposition in Brazilian Portuguese truncation. We have shown that the frequency of a string across the lexicon, as represented by our RC + LC model, has a significant influence on where a speaker might place a boundary within a word. In doing so, they may create a new morpheme boundary that shows up in reanalysis. Within this hypothesis, the English word *alcoholic*, which originally was morphologically parsed as *alcohol-ic*, might be reparsed as *alc-oholic* due to RC + LC type frequency effects. At some stage after this, speakers might identify these reanalyzed elements as *alc-*, 'alcohol' (this in itself may be sensitive to *alcohol* being the more frequent and/or salient right complete of *alc-*), and *-oholic*, 'person addicted to X'. This in turn can lead to the newly reanalyzed suffix *-oholic* to be applied in novel constructions, such as *shop-oholic*.

6 Conclusion

When looking at several truncation strategies in Brazilian Portuguese independently, we have found that truncation is best modeled as optimizing original word recovery (minimizing the *right-complete counts*) and deletion of uninformative right-edge material (minimizing *left-complete counts*). We show that a model that considers both right- and left-complete counts together not only outperforms a model that only considers each (or a variant) independently, but also outperforms prosodic models based on binary feet. We take this to be evidence that frequency-based informativity should be incorporated into a complete theory of truncation, in conjunction with other phonological and/or morphological constraints.

Our model of truncation is sensitive to the distribution of surface similarities between words, and is thus affected by the morphological composition of words. While the model itself is not a priori aware of morphological structure, it is sensitive to the presence of these morphemes indirectly through surface similarities, which may have implications for further work on models of morphological segmentation. Because this sensitivity to morphological composition of words in the lexicon is purely segment-based, our model allows independent homophonous morphemes to affect forcible (re)segmentation of a given form, allowing for potential morphological reanalysis of that form.

This paper highlights the importance of taking into consideration the effects of recoverability and deletability in truncation derivation. While the RC + LC model outperforms the other models, it is less than ideal as a general model of truncation prediction. An inclusive model that incorporates the RC + LC approach with prosodic and phonotactic information, as well as some knowledge of morpheme boundaries would be likely to significantly increase the accuracy of truncation prediction. Moreover, a natural area of further work is blending. While previous work on blends have incorporated recovery of original source words from preserved material in their analysis (Gries 2004; Cook 2010; Lignos & Prichard 2015), they have not treated deletion as being independently motivated as our RC + LC model does. As such, our approach to truncation could be tested on other subtractive word-formation processes such as blending.

In the interest of reproducible and extensible research, we have made our complete software package (including all datasets of the lexicon and gold standard wordlist, as well as code for running all models discussed and evaluation results) publicly available to provide a basis for further research on truncation, extension to blending, and beyond.[6]

Our work has shown that linguistic strategies for word-formation likely involve speakers making inferential generalizations based on statistical knowledge. Specifically, it is desirable to model truncation as something that involves generalizations made about the entire lexicon. As the global linguistic knowledge, such as the lexicon, of an individual speaker changes with their experience, these inferential generalizations might also change; further research may reveal if this may lead to changes in linguistic phenomena, such as truncation, for a given speaker as a result. One of the greatest benefits to our perspective here is that it makes these potential variations and changes inherent in a person's grammar, rather than assuming a static set of absolute rules. We do not draw a strong distinction between knowledge and use of language, especially for innovative linguistic processes such as truncation. Time will tell if the blurring or dissolution of this distinction holds for more conventional linguistic processes as well.

Abbreviations

binLR = left to right binary foot model, binRL = right to left binary foot model, LC = left-complete count, RC = right-complete count, TF = truncated form, TS = truncated stem

Acknowledgements

We thank Bruna Elisa da Costa Moreira for discussing Brazilian Portuguese truncation and supplementing the data from the existing literature. Constantine Lignos provided extensive and insightful comments to an earlier draft of this paper. We also benefited from discussion with members of the linguistics community at the University of Chicago, especially John Goldsmith and Jason Riggle. Adam Singerman offered help

[6] https://github.com/jacksonllee/BP-truncation.

with Brazilian Portuguese. This work was presented at the following meetings where participants provided helpful feedback: the Midwest Speech and Language Days at the University of Illinois at Urbana-Champaign on May 2–3, 2014, the 89th Annual Meeting of the Linguistic Society of America in Portland, Oregon, on January 8–11, 2015, and the American International Morphology Meeting 3 at the University of Massachusetts, Amherst, on October 2–4, 2015. We are grateful to the three anonymous reviewers from *Glossa* who provided constructive comments.

Competing Interests

The authors have no competing interests to declare.

References

Baayen, Harald R. 2001. *Word Frequency Distributions* 18 (Text, Speech, and Language Technology). Dordrecht: Kluwer. DOI: https://doi.org/10.1007/978-94-010-0844-0

Belchor, Ana Paula Victoriano. 2006. O encurtamento de formas com a preservação do morfema à esquerda: Uma análise otimalista [Shortening of forms with preservation of the morpheme to the left: An optimality analysis]. *Revista Virtual de Estudos da Linguagem – ReVEL* 4(7).

Belchor, Ana Paula Victoriano. 2009. *Construções de formas truncadas no português do Brasil: Análise estrutural à luz da Teoria da Otimalidade [Truncation constructions in Brazilian Portuguese: A structural analysis in light of Optimality Theory]*. Rio de Janeiro: UFRJ/Faculdade de Letras master's thesis.

Cook, C. Paul. 2010. *Exploiting linguistic knowledge to infer properties of neologisms*. Toronto, Canada: University of Toronto dissertation.

Dressler, Wolfgang U. 2005. Word-formation in Natural Morphology. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 267–284. Dordrecht: Springer. DOI: https://doi.org/10.1007/1-4020-3596-9_11

Goldsmith, John A., Jackson L. Lee & Aris Xanthos. 2017. Computational learning of morphology. *Annual Review of Linguistics* 3(1). 85–106. DOI: https://doi.org/10.1146/annurev-linguistics-011516-034017

Gonçalves, Carlos Alexandre Victório. 2006. Usos morfológicos: Os processos marginais de formação de palavras em português [Morphological uses: Marginal word formation processes in Portuguese]. *Gragoatá (UFF)* 21. 219–242.

Gonçalves, Carlos Alexandre Victório. 2009. Retrospectiva dos estudos em morfologia prosódica: De regras e circunscrições à abordagem por ranking de restrições [Retrospective in studies of prosodic morphology: From rules and constituency to a constraint-ranking approach]. *Alfa* 53(1). 195–221.

Gonçalves, Carlos Alexandre Victório. 2011. Construções truncadas no português do Brasil: Das abordagens tradicionais à análise por ranking de restrições [Truncation constructions in Brazilian Portuguese: From traditional approaches to constraint-ranking analyses]. In Gisela Collischonn & Elisa Battisti (eds.), *Língua e linguagem: perspectivas de investigação*, 293–327. Porto Alegre: EDUCAT.

Gonçalves, Carlos Alexandre Victório & Renato Pazos Vazquez. 2004. Fla x flu no maraca: Uma análise otimalista das formas truncadas no português do Brasil [Fla x flu no maraca: An optimality analysis of truncation in Brazilian Portuguese]. In José Pereira da Silva (ed.), *Questões de morfossintaxe* 8. 56–64. Rio de Janeiro: Cifefil.

Gries, Stefan Th. 2004. Shouldn't it be *breakfunch*? A quantitative analysis of blend structure in English. *Linguistics* 42(3). 639–667. DOI: https://doi.org/10.1515/ling.2004.021

Gries, Stefan Th. 2006. Cognitive determinants of subtractive word formation: A corpus-based perspective. *Cognitive Linguistics* 17(4). 535–558. DOI: https://doi.org/10.1515/COG.2006.017

Harris, Zellig. 1955. From phoneme to morpheme. *Language* 31. 190–222. DOI: https://doi.org/10.2307/411036

Katamba, Francis. 2005. *English words*. 2nd edn. London/New York: Routledge.

Kemmer, Suzanne. 2003. Schemas and lexical blends. In Hubert Cuyckens, Thomas Berg, René Dirven & Klaus-Uwe Panther (eds.), *Motivation in language: Studies in honor of Günter Radden*, 69–97. Amsterdam and Philadelphia: Benjamins. DOI: https://doi.org/10.1075/cilt.243.08kem

Kreidler, Charles W. 1979. Creating new words by shortening. *Journal of English Linguistics* 13. 24–36. DOI: https://doi.org/10.1177/007542427901300102

Kreidler, Charles W. 2000. Clipping and acronymy. In Geert E. Booij, Christian Lehmann, Joachim Mugdan, Wolfgang Kesselheim & Stavros Skopeteas (eds.), *Morphologie-morphology: An international handbook of inflection and word-formation* 1. 956–963. Berlin/New York: Walter de Gruyter.

Lignos, Constantine & Hilary Prichard. 2015. Quantifying cronuts: Predicting the quality of blends. Talk at the 89th Annual Meeting of the Linguistic Society of America.

López Rúa, Paula. 2006. Nonmorphological word formation. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 675–678. 2nd edn. Oxford: Elsevier.

Mattiello, Elisa. 2013. *Extra-grammatical morphology in English: Abbreviations, blends, reduplicatives and related phenomena* (Topics in English Linguistics 82). Berlin/Boston: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110295399

Scher, Ana Paula. 2011. Formas truncadas em português brasileiro e espanhol peninsular: Descrição preliminar [Truncated forms in Brazilian Portuguese and Peninsular Spanish: Preliminary description]. *ReVEL, edição especial* 5.

Scher, Ana Paula. 2012. Concatenative affixation in Brazilian Portuguese truncated forms. In Nobu Goto, Koichi Otaki, Atsushi Sato & Kensuke Takita (eds.), *The proceedings of GLOW in Asia IX*.

Vilela, Ana Carolina, Luisa Godoy & Thaïs Cristófaro Silva. 2006. Truncamento no português brasileiro: para uma melhor compreensão do fenômeno [Truncation in Brazilian Portuguese: For a better understanding of this phenomenon]. *Revista de Estudos de Linguagem* 14(1). 149–174. DOI: https://doi.org/10.17851/2237-2083.14.1.149-174

How to cite this article: Pham, Mike and Jackson L. Lee. 2018. Mincing words: Balancing recovery and deletion in word truncation. *Glossa: a journal of general linguistics* 3(1): 36.1–19, DOI: https://doi.org/10.5334/gjgl.269

Submitted: 27 September 2016 **Accepted:** 21 October 2017 **Published:** 16 March 2018

Copyright: © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press. **OPEN ACCESS**