Combining successor and predecessor frequencies to model truncation in Brazilian Portuguese*

Mike Pham and Jackson L. Lee
University of Chicago
{mpham, jsllee}@uchicago.edu

October 27, 2014

Abstract

Brazilian Portuguese exhibits word truncation: e.g. *vagabunda* 'slut' > *vagaba*, where the theme vowel *-a* is added to the truncated stem *vagab*. Gonçalves 2011 claims that truncated words preserve the rightmost syllable's onset of the first binary foot. Alternatively, Scher 2012 proposes a Distributed Morphology account involving reanalysis of internal morphological structure without actual truncation: *cerveja* is reanalyzed as $\sqrt{\text{CERV}}$-*ej*-*a*, with the new root $\sqrt{\text{CERV}}$ as a truncated stem to derive *cerva*. We argue instead that derivation of the truncated stem is better modeled by SUCCESSOR FREQUENCIES and PREDECESSOR FREQUENCIES (SF and PF, respectively; Harris 1955; Hafer and Weiss 1974) optimizing phonological truncation and original word recovery. More specifically, a model incorporating both SF and PF outperforms one that uses only one or the other, as well as a binary foot model, in predicting truncated stems in Brazilian Portuguese. Locating the best SF-PF trade-off point can be viewed as the best morpheme boundary of a given word, which can in turn serve as the basis of a potential morpheme segmentation model, a fully unsupervised strategy that does not a priori assume (i) directionality of affixation and (ii) consistency among morphemes.

1 Introduction

Like many languages, Brazilian Portuguese (BP) can to a certain extent shorten words into truncated forms (TF), which are commonly organized into at least four different types (Scher, 2012). Type 1 TFs are formed by taking the initial morpheme in the full form and deleting the following material; Type 2 TFs retain part or all of the root from the original full form, ending in a vowel from this original root; Type 3 TFs 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 TFs are identical to Type 3 TFs, with the difference being that the inserted suffixal material is either *-as* or *-(i)s*. Note that virtually all truncation in Brazilian Portuguese that we have come across in our data is left-anchored with respect to the preserved material and right-anchored with respect to deleted material; in other words, the beginning of words are preserved with deletion occurring at the end of words.

(1) **Type 1: preserve first morpheme**

*Technical report, Department of Computer Science, University of Chicago

a. psicologia, 'psychology' → psico
b. odontologia, 'dentistry' → odonto
c. fonoaudiologia, 'speech therapy' → fono

(2) **Type 2: preserve (part of) root**
a. prejuízo, 'loss (of money)' → preju
b. bijuteria, 'bijou' → biju
c. depressão/deprimido, 'depression/depressed' → deprê

(3) **Type 3: preserve (part of) root ending with consonant and appending** *-a*
a. cerveja, 'beer' → cerv-a
b. vagabunda, 'slut' → vagab-a
c. delegado, 'sheriff' → deleg-a

(4) **Type 4: preserve (part of) root ending with consonant and appending** *-as/-(i)s*
a. saudades, 'homesickness' → saud-as
b. bermuda, 'shorts' → berm-as
c. bobeira, 'silliness' → bob-(i)s

In this paper we distinguish the two terms TRUNCATED FORM and TRUNCATED STEM. We refer to the entire word on the right side of the arrow as the truncated form, which comprises a truncated stem (TS): for example, in *cerveja*, the truncated stem is *cerv-* to which the theme vowel *-a* is added (making it Type 3), an independent nominalizing suffix in BP. Our focus is to model the derivation of the TS (i.e. *cerv-*), 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 normal nominal morphophonological processes operating on a derived TS: we assume that whatever morphological process that results in the full word *cerveja* having the theme vowel *-a* is the same process that produces the TF *cerva* from the TS *cerv-*. As our approach does not make reference to a priori morpheme boundaries, we do not make further distinction between the different types of truncation in BP.

Previous approaches to truncation in BP have either been phonological (Belchor, 2006, 2009; Gonçalves, 2006, 2009, 2011; Gonçalves and Vazquez, 2004) or morphosyntactic (Scher, 2011, 2012). Gonçalves 2011 is an example of the former, where there is a prosodic process that preserves the onset of the right-edge syllable in the first binary foot. This is essentially equivalent to saying that truncation keeps the first binary foot of the original word, up until the second syllable's vowel; one of the problematic predictions of such an analysis, however, is that it predicts only bisyllabic TFs, which is clearly not the case, as can be seen in *odonto* (1b) and *vagaba* (3b).

Alternatively, Scher 2012 derives the TF of Type 3 and 4 words by decomposing the morphological structure of the original word. For example, she analyzes the TF *cerva*, from *cerveja* (3a), as having the following morphosyntactic structure: $\sqrt{}$CERV-*ej-a*. She defends the odd claim that $\sqrt{}$CERVEJ- can be further decomposed by providing data that show that *-ej-* (along with *-am-* and *-at-*) are unrelated suffixes in other contexts – this essentially makes her analysis for these examples based on reanalysis/back formation based strictly on phonological identity to another morpheme (not unlike the tongue-in-cheek English example *history* > *his-tory* > *her-story*). This presents a

very strict environment in which truncation in BP can take place: it is limited by the capacity to reanalyze parts of the root based on phonologically identical affixes elsewhere in the language; this is even more problematic within the Distributed Morphology framework she utilizes, where Late Insertion prevents any phonological material from being visible within the same Spell Out domain.

We take a different approach, one that is purely phonologically based and models TS derivation as optimizing maximal deletion of the original word and maximal likelihood of word recovery by the hearer: 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, as it were. We consider the deleted material to be phonological segments (rather than morphemes, for example). 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 formalized and computationally implemented as a version of Zellig Harris's successor frequencies (PFs) (Harris, 1955; Hafer and Weiss, 1974). SFs provide a quantitative basis for estimating the optimal truncation point for maximal likelihood of word recovery, and PFs for maximal deletion. We compare various models involving SFs and PFs against each other to predict TSs in BP.

The remainder of this paper is organized as follows: in §2 we describe our methodology and how each model predicts what the TS should be. In §3 we provide the results of running each model on a gold standard list of attested TFs in BP, as well as how we evaulate the accuracy of each model. In §4 we discuss our results, and provide a more general outlook on our work's implications for morphology in §5. We conclude in §6.

2 Methodology

In this section, we discuss our methodology: first by explicitly defining what successor frequency and predecessor frequency mean in our models, then by elaborating on our data source, and finally by outlining the four models of truncation in BP that we construct and evaluate in this paper.

2.1 Successor frequencies and predecessor frequencies

To construct models of truncation, we employ a version of Harris 1955's successor frequencies and predecessor frequencies—originally proposed for word and morpheme boundary discovery—to predict the optimal truncation point of BP gold standard nouns with attested truncated forms (details on the gold standard list and data in the following section, §2.2). We define successor and predecessor frequency as follows:

(5) a. **Successor frequency (SF):** the number of words in a lexicon that contain and begin with a given string of symbols; i.e. ABCDE is a successor of ABC, as it contains and begins with the string ABC

b. **Predecessor frequency (PF):** the number of words in a lexicon that end with a given string of symbols; i.e. VWXYZ is predecessor of XYZ, as it contains and ends with the string XYZ

For each BP word with an attested truncation, we consider 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 calculate its SF and PF, viz. the number of words in our BP lexicon beginning/ending with the left/right-most substring. For example, given a potential TS **vagabun* from *vagabunda*, SF is the number of words in the lexicon that begin with the string *vagabun*; the correlating PF would be the number of words that end with the string *da*.

2.2 Data

Our data source comprises two main components: the first is a BP lexicon of about 350,000 words that was created from a Brazilian Portuguese corpus. The second is a set of 96 gold standard nouns with attested TFs that were pulled from data in Scher 2012 as well as the appendix of Vilela et al. 2006; proper names were excluded – hypocoristics have been shown to potentially be derived from independent processes than other types of truncation (John Goldsmith, personal communication) – as were the handful of 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 for now.

For the gold standard words, we took the actual attested TS to be the maximal preserved material corresponding to the original word for truncations of Types 1 and 2 (no phonological material added after deletion): i.e. the attested TF for *bijuteria* is *biju*, and the TS is also *biju*. For Types 3 and 4, we considered the TS to be the maximal preserved material corresponding to the original word excluding the final *-a(s)* or *-(i)s*, as we take those suffixes to be inserted due to an independent regular morphological process in the language: i.e. the attested TF for *vagabunda* is *vagaba*, and the TS is *vagab*. Cases where the TF appears to be ambiguous between being Type 1/2 (non-suffixed) or Type 3/4 (suffixed) were grouped with Type 3/4: the TF *pija* from *pijama* has the TS *pij*.

An advantage of using BP is that the orthography and actual phonology are relatively isomorphic – as compared to English, for example. However, our current models do not take into account orthographic conventions correlating to specific phonological forms. For example, the word *manequim* has the attested TF *maneca*, where the orthographic "q" and "c" are phonologically equivalent (/k/) in both forms. In such a case, we decided that the attested TS was *maneq*; though it should be obvious that the lack of orthographic identity will skew the SF and PF values for these words. In total, of the 96 gold standard nouns, 10 involved some mismatch in orthography and phonology correlating to a regular orthographic rule. Further work will better account for orthographic conventions.

2.3 The tested models

In this paper, we test various models of truncation in BP against each other in order to determine which most accurately predicts the attested TSs. The four models tested are (i+ii) the SF- and PF-only models, which predict the TS to terminate at (and including) the symbol at the point of maximal curvature of their respective elbow curve; (iii) the SF+PF combined model, which predicts TS termination after the symbol closest to the SF and PF curve intersection; and (iv) the binary foot model, which predicts TSs to terminate before the second vowel/nucleus, following Gonçalves

2011, who observed that TFs in BP preserve the onset of the second syllable of the first binary foot.

Our Python script for the four models, including our evaluation metrics plus data files and outputs, is available online at: `https://github.com/JacksonLLee/successor-predecessor-freq/`

3 Results and Evaluation

Graphing the SF and PF values calculated by our algorithm provides two curves for each word, which were then used to derive predicted TSs for the three models based on SF and PF: (i+ii) the predicted TS for SF- and PF-only models terminates at (and including) the symbol at the point of maximal curvature of their respective elbow curve; (iii) the predicted TS for the SF+PF model terminates at (and including) the symbol closest to the intersection of the SF and PF curves.

We evaluated the accuracy of the predicted TS for each model (including the binary foot model) for each word by counting the distance in number of symbols between the predicted TS and the attested TS. Distance errors were then summed up for each model, with SF+PF having the lowest value of errors and therefore being the best model to predict TSs.

3.1 Results

For each word in our gold standard list, we calculated the SF and PF values for each potential stem, as outlined above, giving us a table like the one below for each word. Log distributions are also provided due to highly skewed distributions in lexical statistics (Baayen, 2001)

(6)

TRUNC:	V	A	G	A	B	u	n	d	a
SF:	8652	1924	146	69	26	21	21	21	7
LOG(SF):	3.9371	3.2842	2.1644	1.8389	1.4150	1.3222	1.3222	1.3222	0.8451
PF:	1	1	1	14	20	69	464	6461	52303
LOG(PF):	0.0	0.0	0.0	1.1461	1.3010	1.8389	2.6665	3.8103	4.7185

In (6), the top row shows the original word, with the symbols comprising the attested TS in capital letters. For SF, the number in each column shows the SF value for each potential TS formed from the symbols to the left of and including the symbol heading that column: i.e. as can be seen in the column headed by "V", there are 8,652 words in the lexicon that begin with the string *v*; there are 1,924 words that begin with the string *va*; 146 beginning with *vag*; etc. PF values are the inverse: starting from the right edge, we can see that there are 52,303 words in the lexicon that end with the string *a*, 6,461 that end with the string *da*, 464 with *nda*, and so on.

Plotting the LOG(SF) and LOG(PF) values provides a graph, as seen below in Figure 1:
As can be seen, starting from the left edge and moving rightwards, SF begins with a high value and steeply declines as the potential TS gets longer. PF mirrors this: it begins with a high value on the right edge and steeply declines moving rightwards as the potential deleted material gets longer. This intuitively makes sense as the number of words that contain a given string will decline as that string gets longer.

For the SF- and PF-only models, the potential TS is calculated by finding the "elbow" point along the respective curves, or the point of maximal curvature. The reasoning is that this is the point at which original word recovery will begin to become more difficult proportionate to

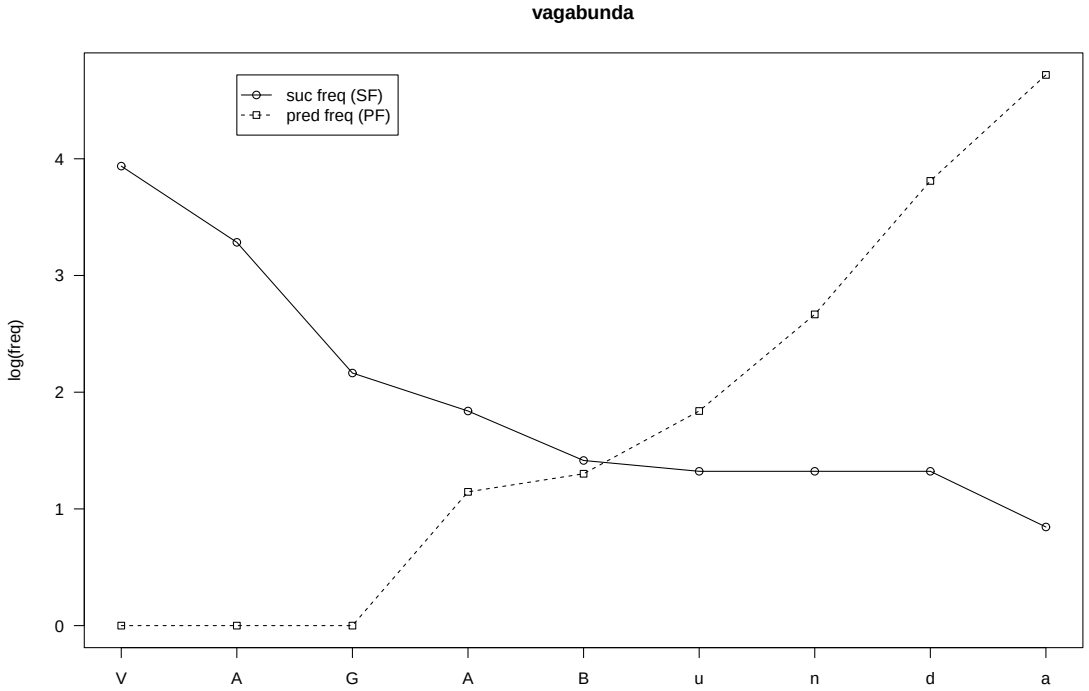


Figure 1: *vagabunda* graph

the number of greater number of possible SFs or PFs corresponding to that point. To be more concrete, consider the SF-only model: looking at the raw SF values in (6), we can see that there are 7 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 21 successors; the jump from 7 to 21 is relatively minute on the scale of lexicon entries. What is more significant is the point at the symbol *B*, where the number of successors goes from 146 to 1,924 at the first *A* symbol. This increase is much more abrupt than the previously mentioned ones; one might expect a speaker only looking at SF to consider this the TS that best optimizes deleting right-edge material without introducing too many possible words that the TS can be reconstructed as.

As mentioned, the LOG(PF) curve more or less mirrors the SF curve in terms of general shape, though of course is not the identical curve in reverse. However, rather than representing the number of possible reconstructions of a potential TS, as SF does, the LOG(PF) curve represents the number of words that also end in the deleted material. Another way of looking at this is to say that SF provides a metric for the informativity of the preserved material while PF provides a metric for the informativity of the deleted material. As such, right-edge material that has a high PF value, such as the final *a* in (6) can be seen as relatively uninformative, as there 52,303 words that end with that. Looking at the PF alone, then, allows us to predict how much of the right-edge material can be deleted. The elbow of the LOG(PF) 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. To calculate where these elbow points

are, we calculated the symbol with the greatest second derivative value for each curve. In (1), this means that the *G* point is the predicted SF elbow over *B*.

For the SF+PF model, we consider the LOG(SF) and LOG(PF) curves together instead of considering them separately: rather than directly considering the elbow points of each curve to predict the TS, this model uses the intersection of both curves. As the LOG(SF) and LOG(PF) curves are roughly mirrored shape, this point of intersection will be somewhere towards the middle of the word. However, while the elbow points of the curves are inherently located at a symbol by virtue of the SF and PF values being aligned with symbols, the intersections of the LOG(SF) and LOG(PF) curves rarely line up exactly with a symbol.

Given this, for the SF+PF model, we use the symbol closest to the intersection of the curves for the termination point of the predicted TS. The closest symbol is determined by calculating the shortest x-axis distance from the point of intersection to the symbol on either side of the intersection point. In Figure 1, this point is easily eyeballed by looking at the graph, but looking at the graph for *transação*, with the attested TS *trans*, in Figure 2, we can see that it is necessary to have this more precise measure of closest symbol to intersection; in the case of Figure 2 for *transação*, it turns out to "S".

(7)

TRUNC:	T	R	A	N	S	a	ç	ã	o
SF:	17191	4543	2411	1025	759	63	5	1	1
LOG(SF):	4.2353	3.6573	3.3822	3.0107	2.8802	1.7993	0.6990	0.0	0.0
PF:	1	1	1	11	29	1941	2731	6510	50164
LOG(PF):	0.0	0.0	0.0	1.0414	1.4624	3.2880	3.4363	3.8136	4.7004

Finally, the TSs as predicted by the binary foot in the fourth and last model were hand coded in our dataset for automated evaulation. In the next section, we discuss the evaluation of the four models.

3.2 Evaluation

The most basic evaluative measure of our four models is to compare the percentage of TS accurately predicted by each model, with the following results:

(8) % of TSs accurately predicted

	% correct
(i) SF:	16.7
(ii) PF:	18.9
(iii) SF+PF:	33.3
(iv) binary foot:	28.9

At first glance, we can see that no model is as accurate at predicting TSs as we would hope – ideally, the best model would be able to predict the attested TS 100% of the time. Crucially for us, though, it should be noted that the SF+PF model is still the most accurate of all the models tested, including the binary foot one.

Beyond this cursory binary measure of error, however, we use a more detailed metric of error based on DISTANCE ERROR, as defined below:

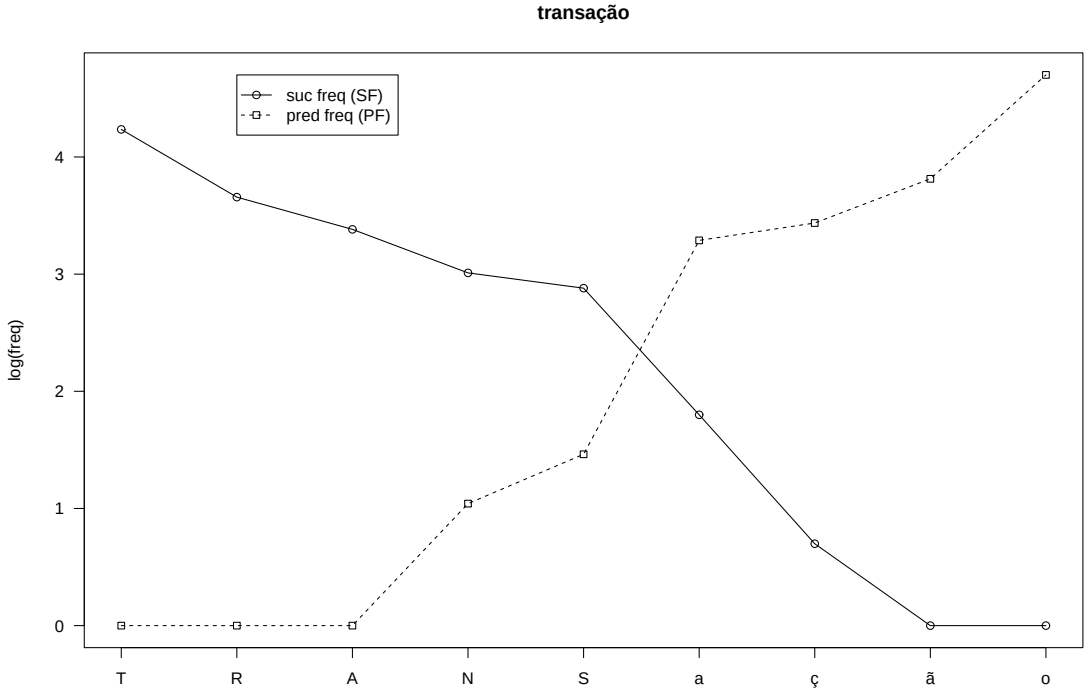


Figure 2: *transação* graph

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

Consider the example *preguiça*, 'laziness', with the TS (and TF) *pregui*:

(10)

TRUNC:	P	R	E	G	U	I	ç	a
SF:	25678	6153	2245	101	42	16	9	2
LOG(SF):	4.4111	3.7891	3.3512	2.0043	1.6233	1.2041	0.9542	0.3010
PF:	2	2	2	4	6	79	652	52303
LOG(PF):	0.3010	0.3010	0.3010	0.6021	0.7782	1.8976	2.8143	4.7185

Numbering each of the symbols on the x-axis starting with 1 on the left edge ("P"), we can see that the attested truncated stem terminates at 6, corresponding to the symbol "I". That is, we simply assign an x-axis position number to each symbol, in order to acquire a numeric integar value for the termination point (or length) of the attested and predicted TSs.

With these numbers we then calculate the distance error for each model, using the formula below:

(11) $E = TS_0 - TS_x$

The distance error E is equal to the numeric value of the predicted truncated stem (TS_x) subtracted from the attested truncated stem (TS_0) for a given word. E is a numerical integar value

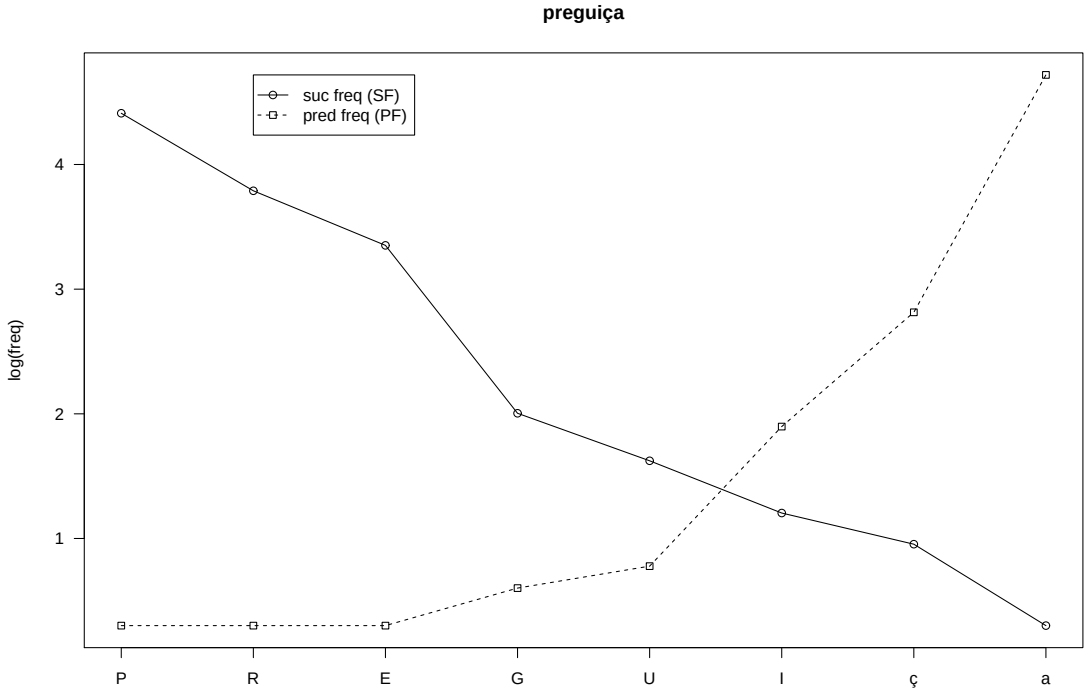


Figure 3: *preguiça* graph

corresponding to the error of the predicted TS when compared with the attested TS: i.e. a distance error of 2 means that the predicted TS terminated two symbols to the left of the attested TS, or, alternatively, that the predicted TS is two symbols shorter in string length than the attested TS. Note that this method of calculating E allows for negative values as well, which occurs when the predicted TS_x terminates to the right of the attested TS_0 ($TS_x > TS_0$) – that is, the predicted TS is longer than the attested TS.

The attested and predicted TS and distance errors for each model for *preguiça* are as follows:

(12) a. TS_0 and TS_x for *preguiça*

Attested TS:	6
(i) SF:	4
(ii) PF:	7
(iii) SF+PF:	6
(iv) binary foot:	4

b. E for *preguiça*

(i) SF:	2
(ii) PF:	-1
(iii) SF+PF:	0
(iv) binary foot:	2

As can be seen in (12b), (iii) SF+PF has the lowest E of 0; for this particular word, the SF+PF model was the most accurate model in predicting the attested TS – it in fact exactly predicted the attested TS. Compare this to (i) SF and (iv) binary foot, which both have an E of 2, meaning that they predicted a TS two symbols shorter (**preg*) than TS_0. (ii) PF is more accurate, but overpredicts the TS, with a value of -1 (corresponding to the predicted TS **preguiç*).

We calculate the distance error in number of symbols between the predicted TS of the model and the attested TS of each word in the gold standard list. For each model, we then sum the distance errors for each word in the gold standard list to get a total distance error. These total distance errors can then be directly compared to each other to show that the SF+PF model is the most accurate of the four models, with the lowest total distance error.

There are two ways of computing the sum of errors, depending on whether the error direction (that the the predicted TS over- or under-predicts the attested TS, so to speak) is taken into account.

In order to compare the models with each other for overall error E_{total}, $|E|$ for each word with respect to each model is summed. The absolute value of E is taken because it is only the distance of the error that we care about for E_{total}, regardless of which direction the error is in.

(13) $E_{total} = \sum |E_n|$

In (13) above, E_n is the value of E with respect to each word n within the gold standard list. The ideal model would have an E_{total} equal to 0; that is, it perfectly predicts the TS for every word in the gold standard list. Generally speaking, then, the lower the E_{total} is for a model, the better it is.

While E_{total} gives a measure of total distance error without factoring in direction of error, we also calculate the overall skewness of error E_{skew} by using the raw E_n values without taking the absolute value:

(14) $E_{skew} = \sum E_n$

What E_{skew} tells us is not the aggregate value of how inaccurately a model predicts TSs, but rather whether it consistently predicts the TS to be too long or too short. While a model that has $E_{total} = 0$ will also have $E_{skew} = 0$, the inverse is not necessarily true, as the positive and negative distance errors of a given model can cancel each other in E_{skew} but not E_{total}. What $E_{skew} = 0$ tells us instead is that that particular model is just as likely to predict a TS to be too short as it is to predict it to be too long. By extension, having a positive E_{skew} value tells us that a particular model is more likely to predict TSs that are too short, while the inverse is true of a negative E_{skew}.

Using these evaluation metrics of model error, the four models tested in this paper perform as follows:

(15) Overall model error

	E_{total}	E_{skew}
(i) SF:	139	33
(ii) PF:	151	-17
(iii) SF+PF:	**87**	**7**
(iv) binary foot:	126	126

As can be seen in (15), the (iv) SF+PF model not only has the best E_{total} value (lower is better) of 87, but it also has the best E_{skew} value (closer to zero is better) of 7. This means that this model is not only the most accurate overall of the four models tested here, but it is also relatively centered with respect to whether the predicted TSs are too long or too short. The (i) SF model has an E_{total} of 139, making it significantly less accurate than the (iii) SF+PF model. It's E_{skew} score of 33 means that this model has a mild tendency to predict TSs that are too short. The (ii)

PF model is the least accurate one, with an E_{total} of 139; it is also the only model with a negative E_{skew}, meaning that it is the only model that tends to predict TSs that are too long. However, with E_{skew} = -17, this is only a slight tendency. The (iv) binary foot model has the the second best E_{total} value of 126, though its E_{skew} value is also 126, showing that this model always predicted TSs that were too short.

It should be noted that the while the E_{total} scores of (i) SF, (ii) PF and (iv) binary foot model all seem relatively comparable and are of thus similar accuracy, the E_{total} of the (iii) SF+PF model seems substantially better than the other three by both metrics of error computation. In the following section, we discuss why we believe the SF+PF model is the best-performing one.

4 Discussion

Our results show that the (iii) SF+PF model is the most accurate in predicting attested TS of words in Brazilian Portuguese. The (i) SF and (ii) PF 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. The (iv) binary foot model occassionally 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.

At its core, our SF+PF model of BP truncation is based on optimizing speaker deletion and hearer recovery of an original form: the optimal TS will delete as much material as possible without introducing too many options for what the original word could potentially be. As has been previously mentioned, SF gives us a measure of how informative the preserved material is in truncation. The number of words that begin with a certain string drops as the string gets longer; consequently, longer stems are more informative in allowing the speaker to recover the original word. This intuitively means, of course, that a maximal form with no deletion should have perfect recovery – or close to it, as there are perhaps additional suffixes that could potentially be on that stem.

This gives us another perspective onto why the binary foot model seems to capture the minimal possible TS: this tends to align with the point on the SF curve where values jump upwards at a much greater rate than at previous points (moving leftwards on the graph). Deleting any additional material would result in drastically increasing the number of possible successors. If this is the case, then we would expect the (i) SF model and the (iv) binary foot one to have similar E_{total}, as they should be correlated with each other: they are in fact close, (i) E_{total} = 139 and (iv) E_{total} = 126, respectively. However, this is no closer than the E_{total} values for (i) SF and (ii) PF. Moreover, the E_{skew} for (i) SF is drastically closer to 0 than for the (iv) binary foot model, meaning that whereas the latter always underpredicted the length of a TS when it was wrong, the latter was more balanced in direction of error, though also tending towards underprediction.

One way to interpret underprediction is that what the (i) SF model is basing its prediction on is the elbow point of the SF curve. This means that it is a measure of the curve's shape or contour rather than the actual values associated with each point on the curve. Because it is possible to transpose the curve vertically (or horizontally) without affecting its contour, calculating the elbow of the curve is not sensitive to the specific values at each symbol, but only the difference between the values. It happens that the ends of the curve are generally roughly anchored to the same place, but there can in theory be multiple curves with identical shapes, but different numeric values of

points on the curve: that is, it is theoretically possible to have parallel SF (and PF) curves vertically stacked on each other. What this means is that while the curve's elbow may be the optimal point for truncation based solely on curve contour, the actual SF value at that particular point is still too high – that is to say, there are still too many potential recovery options. If this is the case, then the actual truncation point will occur to the right of what the (i) SF model predicts, in order to further drive down the SF value and ensure hearer recovery of the original word. This could be what we are seeing for the (i) SF model in terms of its positive E_{skew} value.

Because the (ii) PF model is essentially the (i) SF model in reverse, we expect that its E_{skew} should be also be reversed, and that its E_{total} might be comparable. This is in fact the case: its E_{total} = 151, higher than in the (i) SF model but still in the same ballpark; and its E_{skew} = -17. Recall that this negative E_{skew} score translates to the (ii) PF model tending to predict longer TSs than what is attested. On the surface, this is a nice result given that we expect the (i) SF and (ii) PF models to be rough mirrors of each other.

The (ii) PF model is interpreted in a slightly different way than the (i) SF model, however: while the latter is a measure of relative informativity of the TS, the former is actually a measure of the relative informativity of the deleted material. As a result, they are of course related to each other, but still have a large degree of autonomy in their behavior. When the PF value is high, as in word-final "a", then that particular string is interpreted as relatively uninformative material, as its ubiquitous distribution does relatively less in helping the hearer discern what the intended form is to be. This is exactly the type of material we expect to be targeted for deletion in truncation.

The reason that the (iii) SF+PF model is more successful than the others, then, is because it has the advantage of combining the measures provided by the (i) SF and (ii) PF models. The former provides information on how much material should be preserved for optimal recovery purposes while the latter provides information on what material can be optimally deleted. Put in simpler terms, given some word, the speaker can begin deleting the uninformative material from the right edge, but stopping when it begins to introduce too many possible reconstructions to ensure that the hearer will successfully recover the original word. This optimal point is precisely what the intersection of the SF and PF curves is modeling. Note that the (iii) SF+PF model is based on looking at the SF and PF curves together, but doesn't directly care about their respective elbow points, which is what the (i) SF and (ii) PF models each use to make predictions. It is the ability of the (iii) SF+PF model to combine both of these metrics that allows for its vastly improved results over the other models.

In the following section, we discuss what we believe the implications of this analysis of BP truncation to be for morpheme segmentation in general.

5 Implications for morpheme segmentation

The goal of our model is to find the optimal truncation point within a given word. This can be alternatively seen as a morphological segmentation model: given some form, what is the best place to create a morphological boundary? Importantly, our model makes no a priori assumptions about the internal structure of the words it looks at. It essentially treats all word forms as being monomorphemic at the outset, as it were, and decides where the optimal boundary should be. Additionally, as a model of morphological segmentation that allows for the decomposition of monomorphemes, we also speculate as to how this model can be extended to morphological reanalysis.

5.1 Morphological segmentation

Unlike other models of morphological segmentation (cf. Goldsmith 2010; Hammarström and Borin 2011), our model of truncation in BP does not assume morpheme consistency from the outset (see Lee 2014; Lee and Goldsmith 2014 for stem extraction). In a sense, our model determines a single morpheme boundary independently for each word; if some substring X is predicted to be a(n optimal) morpheme for a given word then the identical substring X may not be considered an optimal morpheme for another word. The model does not learn morphemes so to speak incrementally and look for those morphemes given a form, but is rather always being forced to create a morpheme boundary without prior knowledge of what morphemes look like in the language.

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 phonological representation of words. Our approach, then, can be interpreted as a way of potentially modeling reanalysis and backformations.

In her analysis of BP truncation, Scher (2012) seems to implicitly assume that reanalysis occurs based on phonological similarity. Consider the following words and their proposed morphological decomposition:

(16) a. *cerveja*, 'beer' > $\sqrt{}$CERV-ej-a
b. *pijama*, 'pajamas' > $\sqrt{}$PIJ-am-a
c. *burocrata*, 'burocrat' > $\sqrt{}$burocr-at-a

(Scher, 2012)

In each of the examples in (16), "the forms *-ej-*, in *cerveja*, *-am-* in *pijama* or *-at-* in *burocrata* are not supposed to be considered separated morphemic in these words" (Scher, 2012). However, Scher tentatively proposes that (present day) speakers of BP are treating these pieces as derivational suffixes based on their surface similarity to other, independent derivational suffixes in BP: *-ej* is a dimunitive suffix, *-am* is a collective suffix, *-at* is a nominalizing suffix:

(17) *-ej*, diminutive
a. *broto-ej-a*, 'rash'
b. *sertan-ej-a/o*, 'woman/man from *sertão*, 'badlands"
c. *pardal-ej-a/o*, 'small sparrow'
d. *grac-ej-o*, 'joke'

(18) *-am*, collective
a. *dinher-am-a*, 'lots of money'
b. *poeir-am-a*, 'lots of dust'
c. *cabel-am-a*, 'lots of hair'
d. *burac-am-a*, 'lots of holes'

(19) *-at*, nominalizer
a. *passe-at-a*, 'parade'
b. *seren-at-a*, 'serenade'
c. *son-at-a*, 'sonata'
d. *brav-at-a*, 'arrogant threat'

(Scher, 2012)

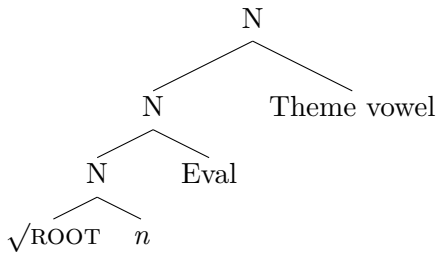
However, other than perhaps the nominalizing suffix *-at*, these are not the meanings of the proposed suffixes in (16): *cerveja* does not mean a dimunitive of beer, nor does *pijama* mean lots of pajamas. Essentially, Scher is claiming reanalysis of a single morpheme into a new decomposed morphological 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 SF and PF in this paper, we are able to model this reanalysis that underlies Scher's story of morphological decomposition in BP truncation. We have shown that the SF+PF model predicts with the best relative accuracy what a TS of a word might be. In doing so, it is creating the new morpheme boundary that shows up in reanalysis. This is due to two factors: (1) UNIQUENESS: the TS is the optimal string of the original word that should remain in truncation to allow for identification of a root; (2) SURFACE MATCHING: with respect to determining a morpheme boundary, phonological sequences that are superficially identical to actual morphemes will treated as if they are actually these morphemes. Surface matching falls out of our model, as it is only considering distrubtional patterns of phonological sequences without taking into account their syntactic or semantic representations.

5.2 A potential model of morphological reanalysis

In her Distributed Morphology analysis, Scher 2012 claims that the evaluative meaning of truncation – which we do not attempt to directly account for in this paper – is due to a morphosyntactic head Eval that intervenes between a nominalizing categorical head (*n* below) and the theme vowel, which we paraphrase in the following simplified schema:

(20)



While we won't recapitulate her entire analysis here, the most relevant part of Scher's analysis for our purposes is that the reanalyzed morpheme, such as *-ej*, can now be associated as the exponent for the added Eval head.

More generally, what this implies for our model of truncation/morphological segmentation is that given some phonological form, a speaker/hearer can morphologically resegment that form and associate new morphemes with different syntactic/semantic representations. Imagine that Scher's analysis is correct and that evaluative meaning is due to the presence of an Eval head in the morphosyntax. While this head may be strictly functional at first, with no phonological exponent, because this evaluative meaning is associated with truncation, morphologically (re)segmentation happens, resulting in two new morphemes: ignoring the theme vowel, *cervej* > *cerv-ej*. At this point, rather than having a single phonological form for the root node and a null phonological form for the Eval head, there are conveniently two phonological forms that can each be associated with their own terminal node. In other words, *-ej* has been reanalyzed as an Eval head.

Because our model also does not assume morpheme consistency, the phonological sequence that is associated with the Eval head in each case of truncation may not be identical. However, sequences that occur with higher frequency, such as those that are homophonous with independent

morphemes in the case of *-ej*, will probably have a greater chance of stabilizing as a new morpheme, as opposed to an arbitrary sequence that does not occur frequently.

In sum, our model of truncation in BP is also a potential model of morphological segmentation that doesn't a priori assume morpheme consistency. It provides a novel way of thinking about how reanalysis of phonological sequences into new morphemes can occur, as well as how the new sequences that result from the new morpheme boundary can become associated with syntactic and semantic structures, such as is potentially the case with the Eval head.

6 Conclusion

Brazilian Portuguese truncation is best modeled as optimizing word recovery: uninformative right-edge material is deleted (minimizing PREDECESSOR FREQUENCIES, PFs) while constraining possible reconstructions (minimizing SUCCESSOR FREQUENCIES, SFs). Beyond truncation, locating the best SF-PF trade-off point can be viewed as what finds the best morpheme boundary of a given word. This can serve as the basis of a potential morpheme segmentation model, a fully unsupervised strategy that does not a priori assume (i) directionality of affixation and (ii) consistency among morphemes. This in turn provides a small first step to understanding how reanalysis works and how new morphemes might arise.

Further work will include fine-tuning our model with an expanded gold standard list, as well as controlling for orthographic inconsistencies. Additionally, our model should be tested not only on more languages, but also other types of truncation. Recall that we have only considered truncation where right-edge material is deleted and preserved material is left-anchored; we have not considered cases of right-anchored truncation, or truncation that might be affected by stress. Furthermore, our model on truncation should also shed some insight onto blends, which are a combination of left and right-anchored truncation.

References

Baayen, Harald R. 2001. *Word Frequency Description*, volume 18 of *Text, Speech, and Language Technology*. Dordrecht: Kluwer.

Belchor, A. P. V. 2006. O encurtamento de formas com a preservação do morfema à esquerda: uma análise otimalista. *Revista de Estudos de Linguagem* 4.

Belchor, A. P. V. 2009. Construções de formas truncadas no português do brasil: análise estrutural à luz da teoria da otimalidade. Masters dissertation, UFRJ/Faculdade de Letras, Rio de Janeiro.

Goldsmith, John A. 2010. Segmentation and morphology. In *Computational linguistics and natural language processing handbook*, ed. Alex Clark, Chris Fox, and Shalom Lappin, chapter 14, 364–393. Wiley-Blackwell.

Gonçalves, C. A., and R. Vazquez. 2004. Fla x flu no maraca: uma análise otimalista das formas truncadas no português do brasil. In *Questôes de morossintaxe*, ed. J. P. Silva, volume 8, 56–64. Rio de Janeiro: Cifefil.

Gonçalves, C. A. V. 2006. Usos morfológicos: os processos marginais de formação de palavras em português. *Gragoatá (UFF)* 21:219–242.

Gonçalves, C. A. V. 2009. Retrospectiva dos estudos em morfologia prosódica: de regras e circunscrições à abordagem por ranking de restrições. *Alfa (ILCSE/UNESP), Araraquara* 44.

Gonçalves, C. A. V. 2011. Construções truncadas no português do brasil: das abordagens tradicionais à análise por ranking de restrições. In *Língua e linguagem: perspectivas de investigação*, ed. Gisela Collischonn and Elisa Battisti, 293–327. Porto Alegre: EDUCAT.

Hafer, M. A., and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10:371–385.

Hammarström, Harald, and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37:309–350.

Harris, Zellig. 1955. From phoneme to morpheme. *Language* 31:190–222.

Lee, Jackson L. 2014. Automatic morphological alignment and clustering. Technical Report TR-2014-07, Department of Computer Science, University of Chicago.

Lee, Jackson L., and John A. Goldsmith. 2014. Complexity across morphological paradigms: a minimum description length approach to identifying inflectional stems. Poster at the MorphologyFest: Symposium on Morphological Complexity, Indiana University, Bloomington, June 16-20, 2014.

Scher, Ana Paula. 2011. Formas truncadas em português brasileiro e espanhol peninsular: descrição preliminar. *ReVEL, edição especial* URL `www.revel.inf.br`.

Scher, Ana Paula. 2012. Concatenative affixation in Brazilian Portuguese truncated forms. In *The Proceedings of GLOW in Asia IX*, ed. Nobu Goto, Koichi Otaki, Atsushi Sato, and Kensuke Takita.

Vilela, Ana Carolina, Luisa Godoy, and 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* 14:149–174.