Modeling Brazilian Portuguese truncation by successor and predecessor frequencies

Mike Pham, Jackson Lee, &
Bruna da Costa Moreira
{mpham, jsllee, bruna}@uchicago.edu

THE UNIVERSITY OF CHICAGO

Big Question

- How can we predict the form of a truncated word?
- **Basic answer:** Optimize (a) deletion and (b) word recovery

Introduction

Brazilian Portuguese (BP) can productively shorten words into truncated forms (TF):

(1) a. cerveja, 'beer' → cerv-a c. bermuda, 'shorts' → berm-as
b. vagabunda, 'slut' → vagab-a d. bobeira, 'silliness' → bob-(i)s

We refer to the entire word on the right side of the arrow as the TF, which comprises a truncated stem (TS) (i.e. *cerv-*) and some final material, which is often the theme vowel *-a*, an independent nominalizing suffix in BP. **Our focus is to model the derivation of the TS**, rather than the full the TF, which we assume to be handled via normal nominal morphophonological processes operating on a derived TS. Also outside of the scope of this current study is the evaluative interpretation of the TF (Scher 2012), which we assume to be a pragmatic question open to further research.

Previous approaches to truncation in BP have either been formally morphosyntactic (Scher 2011, 2012) or phonological (Belchor 2005, 2006, 2007; Goncalves 2006, 2009, 2011; Goncalves & Vasquez 2004). Both types of analyses are similar in deriving the TS/TF by referring to abstract linguistic structure: i.e. morphological decomposition of the original word, applying prosodic constraints. **We take a different approach, modeling TS derivation as optimizing maximal deletion of the original word and maximal likelihood of word recovery by the hearer.**

With the stipulation that left edges have more priority than right edges, phonological segments are iteratively deleted from the right edge, which left-aligns the preserved material. Each stage of deletion is a potential TS. Likelihood of word recovery has two primary factors, successor frequency (SUC-FREQ) and predecessor frequency (PRED-FREQ), which respectively correspond to the number of words that can be formed beginning with a potential TS and the number of words that can be formed ending with the material deleted from a potential TS (Harris 1955, Hafer and Weiss 1974); we also consider a third variable of lexical frequency of the original word within the set of successors for each potential TS (LEX-FREQ).

The TS is derived by deleting uninformative right edge material (relatively high PRED-FREQ values) while minimizing potential competing successors (relatively low SUC-FREQ values; possibly high LEX-FREQ). **The TS is the optimal point in a word where a speaker can delete righwards material and still expect the hearer to recover the original word.** Crucially, our approach differs from previous ones by not referring to any morphosyntactic or morphophonological internal structure other than phonological segments that are the targets of deletion.

Main Claims

- **TSs in BP predicted by two primary factors:**
  - SUC-FREQ: **the number of words that begin with the substring**
  - PRED-FREQ: **the number of words that end with deleted right-aligned substring**
- **The optimal truncation point in the word is located where** SUC-FREQ **and** PRED-FREQ **intersect (±1 segment) on a line graph**

Methodology

Using a BP corpus, we employ a version of Harris's (1955) successor and predecessor frequencies—original proposed for word and morpheme boundary discovery—together with lexical frequency information to predict the optimal truncation point of a given word. Given an original word w of length n, all left-aligned substrings of lengths $\{1, 2, ..., n-1\}$ are considered potential TSs. For each potential truncation point, we calculate the successor and predecessor frequencies, viz. the number of words in a BP lexicon (with 350,000 words) that begin (and end) with the left-hand (and right-hand) substring. Lexical frequency for each word was also provided by the corpus. 20 gold standard TFs provided by a native speaker of BP were examined.

Python script available at:
`https://github.com/JacksonLLee/successor-predecessor-freq/`

Results

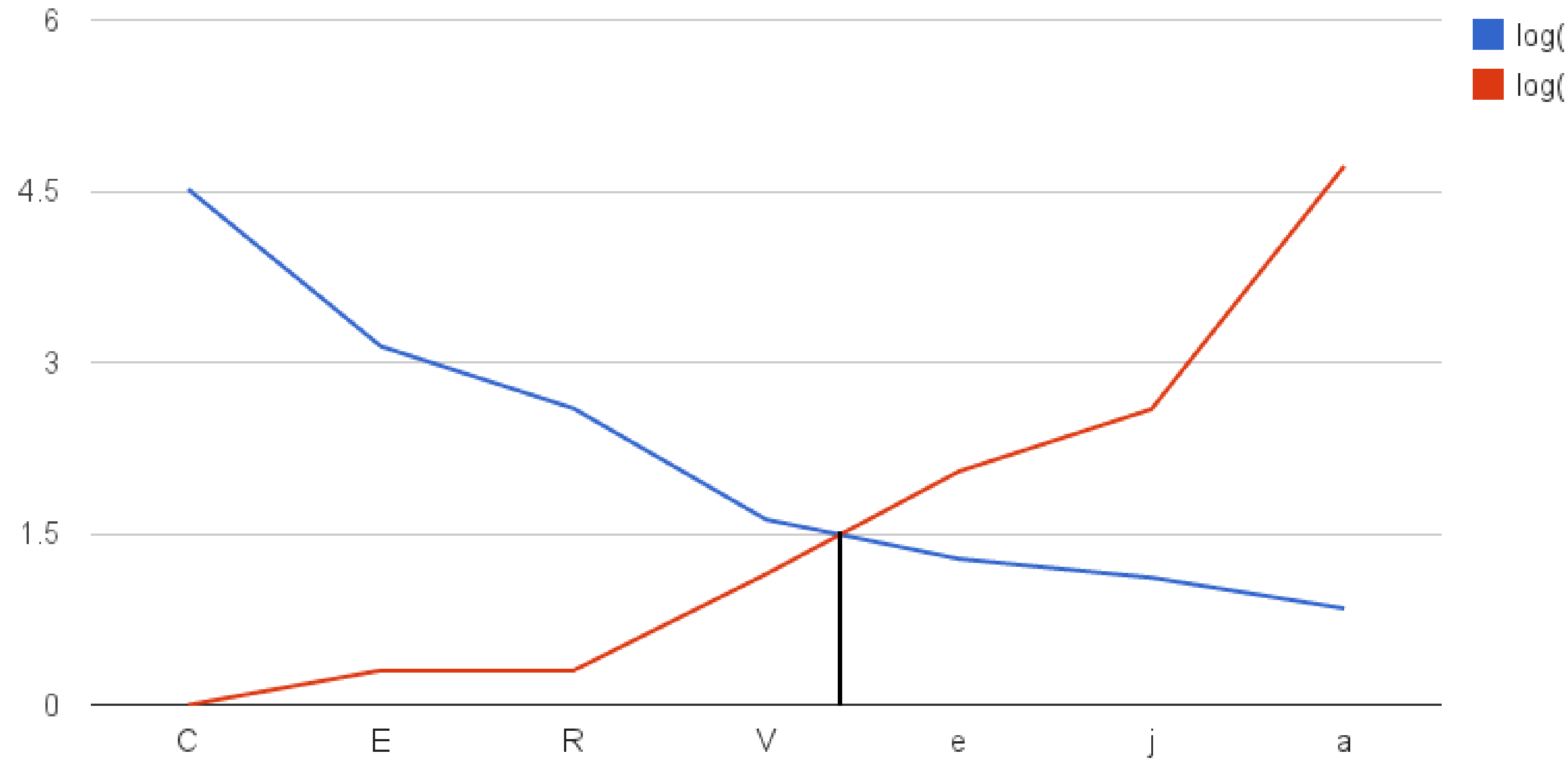
Each example in our data set provides a table as in (2), where the capital letters show the attested TS:

(2)

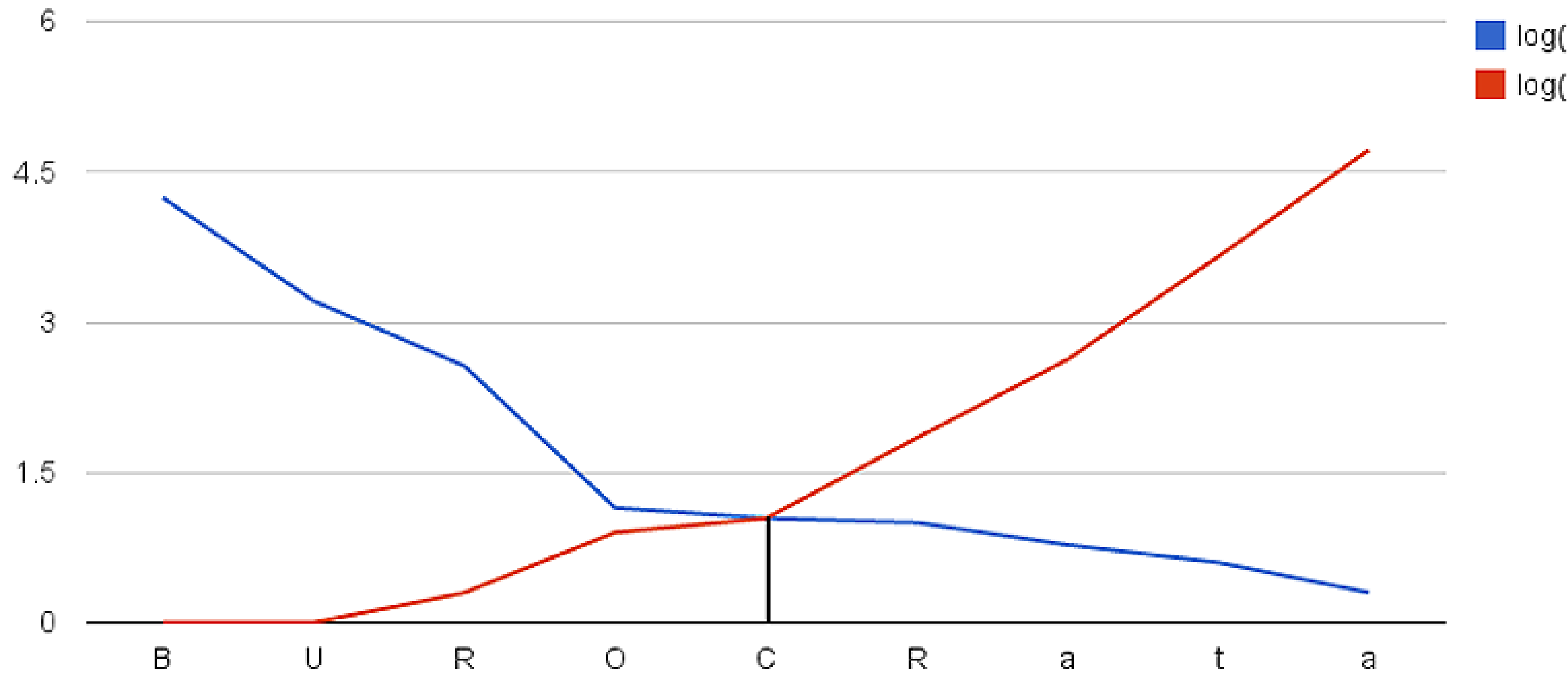
TRUNC:	**C**	**E**	**R**	**V**	e	j	a
SUC-FREQ:	32837	1383	395	42	19	13	7
LOG(SF):	4.51636	3.14082	2.5966	1.62325	1.27875	1.11394	0.8451
LEX-FREQ:	94	6	5	1	1	1	1
PRED-FREQ:	1	2	2	14	111	389	52303
LOG(PF):	0.0	0.30103	0.30103	1.14613	2.04532	2.58995	4.71853

Plotting LOG(SF) and LOG(PF) together gives the intersection, where LOG(SF) > LOG(PF) goes to LOG(SF) < LOG(PF).

- **11 of 20 examples are like *cerveja*, where point of intersection is closest to actual truncation point**:

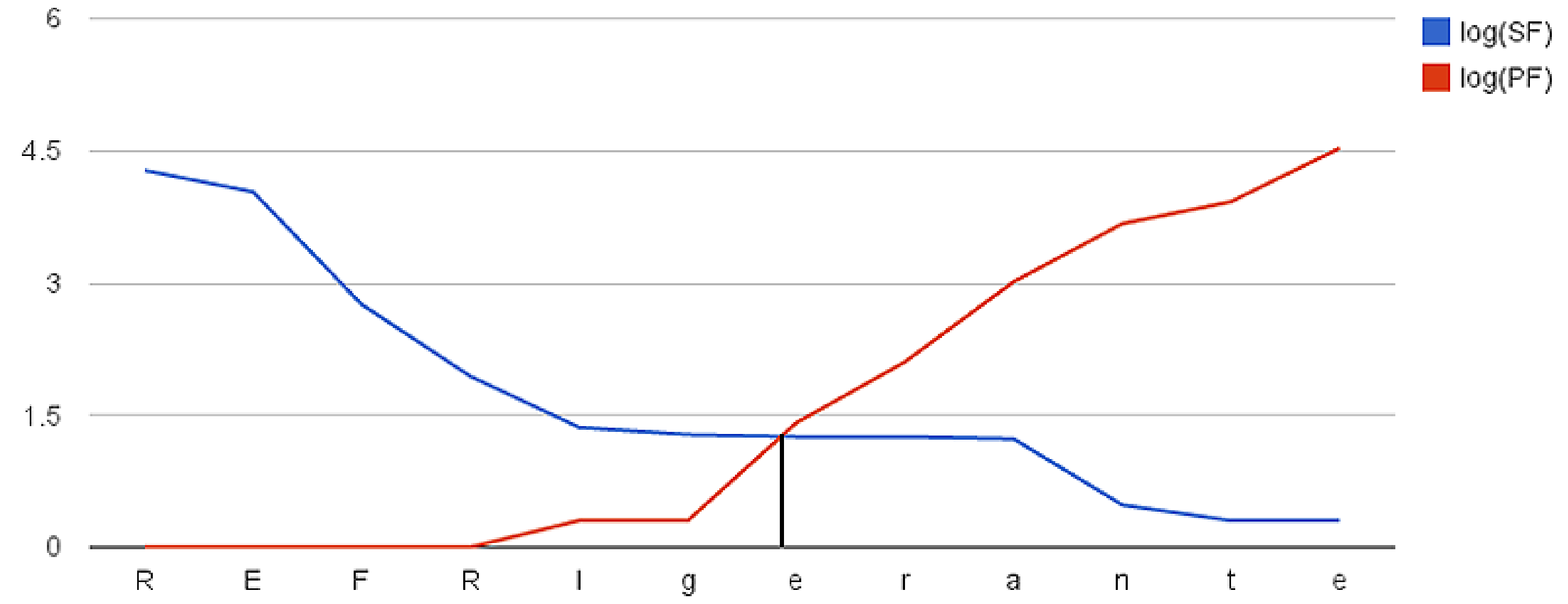


- **8 of 20 examples are like *burocrata*, where point of intersection is closest to segment adjacent to actual truncation point**:



- TS seems to preferably end with consonontal material
- In cases where the actual truncation point is one letter away from the intersection point, the TS gets the entire consonant cluster

- **1 example, *refrigerante*, where point of intersection was not close to actual truncation point**:



Discussion

We find that the intersection point of SUC-FREQ and PRED-FREQ (shown with their log values) is a good predictor of where a word should be truncated in BP to form a TS.

- Preserving material to the right of the intersection is relatively uninformative (high predecessor frequency)
- Deleting material to the left of the intersection introduces relatively many possible reconstructions for word recovery (high successor frequency)
- Intersection of SUC-FREQ and PRED-FREQ is the optimal point for maximal word deletion and maximal probability of word recovery
- *burocrata* shows that consonant clusters behaving as single units; *refrigerante* shows binary foot constraints (Goncalves 2011). This suggests phonological constraints must also be taken into account in predicting TS

Conclusions

- Truncation in BP can be modeled as optimizing deletion of right-edge material and probability of word recovery
- Probability of word recovery is a function of successor frequency and predecessor frequency
- The truncation point is where successor frequency and predecessor frequency intersect

Forthcoming Research

- Determine role of LEX-FREQ
- Examine a larger data set
- Limit successors and predecessors to the identical syntactic category as the original word
- Investigate whether ±1 segment error margin is phonologically conditioned
- Look at truncation cross-linguistically, especially languages where TS's do not seem to be aligned to the left edge of the word
- Pragmatic analysis of evaluative interpretation of TFs

Selected References

Goncalves, C. A. V. (2011) Construcoes truncadas no portugues do Brasil: das abordagens tradicionais a analise por ranking de restricoes. In: Collischonn, Gisela; Battisti, Elisa. (Orgs.). *Lingua e linguagem: perspectivas de investigacao*. Porto Alegre: EDUCAT, p. 293-327.
Hafer, Margaret A. & Stephen F. Weiss (1974), Word segmentation by letter successor varieties, *Information Storage and Retrieval* 10: 371-385.
Harris, Zellig (1955), From phoneme to morpheme, *Language* 31: 190-222.
Scher, Ana Paula (2012), Concatenative affixation in Brazilian Portuguese truncated forms, *The Proceedings of GLOW in Asia IX*.