Contrasting Deverbal Nouns in BP and EP Corpora:
Does it really matter?

1. The issue

Under a descriptive point of view, failure in systematically treating nominal suffixation contrasts between Brazilian and European Portuguese (BP and EP) may create a one-dimensional evaluation of Portuguese morphology. This directly affects applied linguistics domains, such as Portuguese as a Second Language teaching, translation, not to count other multidisciplinary fields, such as Natural Language Processing (NLP), particularly applications such as Information Retrieval (IR). In this later domain, it is largely known that nouns have a critical say in generating information, having a distinctive status if compared to other parts of speech (POS); not surprisingly, they are called *search terms* or merely *terms.* Hence, if one bears in mind that quite a few cases of nominalization in BP and EP undergo a different suffixation process (such as the noun "dirt": *sujeira* (BP) and *sujidade* (EP)), this becomes an issue. Therefore, contrastive nominal suffixation has a major impact not only to Portuguese morphology per se but to lexicon-dependable domains.

2. Theoretical upshot and methodological accounts

This investigation considers the phenomenon of suffixed deverbal nouns with a clear emphasis on empirical data available in both BP and EP. Thus, we rely on corpora evidences to sustain what is a recurrent and preferred suffix in both variations. For this preliminary stage, we used the corpus NILC/São Carlos (Aires & Aluísio, 2001), which contains 42 million words of BP with journalistic texts from *Folha de São Paulo*, but also commercial letters and didactic texts. For EP, we used corpus *CETEMPúblico* (Corpus de Extractos de Textos Electrónicos MCT/Público) (Rocha & Santos, 2000), containing 180 million words.

For that purpose, we specially took into consideration the case study of nouns-deriving suffixes _AGEM and _DA, which form nouns from movement verbs as *virar* ("to turn") and *parar* ("to stop"). Through BP and EP corpora analysis, using Linguateca database lookup, we could perceive a clear morphosemantic and syntactic difference regarding suffixes *_da* and *_agem* adjoined to the verbs *virar* and *parar* (namely "to turn" and "to stop"*)*. More precisely, we could observe that deverbal suffixation *_da* in BP when adjoined to these verbs of movement forms both adjectives and nouns (*para**da*** and *vira**da***). In EP, however, suffix *_da* tends to form stricly adjectives when attached to these verbs, while suffix *_agem* is used to form nouns (*para**gem** and vira**gem***). Thus, according to corpora-based data, EP reveals a steeper morpho-semantic restriction regarding derivation based on verbs of movement such as *virar* and *parar*, in which the word formation rule (WFR), inspired by Aronoff (1976), $[V_{mov}]$*_da* would form adjectives and $[V_{mov}]$*_agem* would form nouns.

If one takes into consideration computational lexicography of the Portuguese language, we could consider that the collocation "*bus stop"* in BP would be *para**da** de ônibus* and in EP, *para**gem** de autocarro*. This preference reveals that lexicographic contrasts go beyond an accidentally inherited lexicon (such as *ônibus* and *autocarro*) but involves derivational morphology itself. For instance, a BP native speaker will be likely to use the sentence: "*A escova**ção** e o fio dent**al** garantem que seus dentes fiquem livres de suj**eira**",* while an EP native speaker might choose: "*A escova**gem** e o fio dent**ário** garantem que seus dentes fiquem livres de suj**idade**",*(which could be translated as "Regular brushing and dental flossing

guarantees dirt-free teeth"). This is clearly relevant to Portuguese contrastive suffixation description, a somewhat neglected issue, as well as to NLP and to Portuguese morphology.

3. Preliminary results

The data obtained in corpora analysis confirms the hypothesis that, indeed, suffix *_da* in BP has a more productive categorial function if compared to EP, deriving both adjectives and nouns. Only 5 cases of *viragem* and 2 of *paragem* were spotted in BP corpus. Yet, through empirical evidences, we could confirm that this figure would represent EP native speakers'discourse in a BP corpus. As a result, for verbs of movement, we could assume a pattern of WFR, which in EP would be $[V_{mov}]$_da $\rightarrow$ Adj and $[V_{mov}]$_agem $\rightarrow$ N. In BP, however, the WFR $[V_{mov}]$_da would form both adjectives and nouns. Nevertheless, this rule does not seem to be as productive in EP. In other words, regarding other verbs of movement such as *correr* ("to run"), *caminhar* ("to walk"), *descer* ("to go down"), *subir* ("to go up"), the WRF in EP would be kept the same as in BP: $[V_{mov}]$_da $\rightarrow \frac{\text{Adj}}{\text{N}}$

4. And now what? Discussion and future work

After this pilot study we could confirm that the implementation of the assumed rule described for verbs *virar* and *parar* is not applicable for other nominalizations of verbs of movement in EP. On the other hand, we did identified that EP reveals a more flexible suffix production regarding i) other deverbal nouns (e.g. "retirement": *aposenta**ção*** /*aposenta**doria*** and "movement": *desloca**ção/desloca**mento***); ii) adjectives (e.g. "promising": *prom**etedor*** /*promi**ssor***) and iii) verbs (e.g. "to optimize": *potencia**r/ potencia**lizar***). BP, however, reveals a clear-cut inflexible suffix production (*aposenta**doria***, *desloca**mento,*** *promi**ssor*** and *potencia**lizar***). In this stage of research, what we can state is that BP derivational morphology does not entirely mirror EP derivational morphology (even though this usually goes unspoken). This is critical not only for lexicography as a whole but also to NLP purposes. For that reason, we are able to answer the question posed in this work's title. And, yes, it does matter.

5. References

Aires, R. & Aluísio, S. (2001). *Criação de um Corpus com 1000.000 de Palavras Etiquetado Morfossintaticamente.* Technical Report NILC-TR-01-8, NILC, Campinas.

Aronoff, M. *Word Formation in Generative Grammar*. Cambridge, Massachussetts: The MIT Press, 1976.

Basílio, M. . *Formação e Uso da Nominalização Deverbal Sufixal no Português Falado.* In: Castilho e Basilio. (Org.). Gramática do Português Falado v. IV. Campinas: UNICAMP/FAPESP, 1996, v. 4, p. 23-30.

Rocha, P & Santos D. (2000) CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa, in Volpe Nunes (ed.), Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000) Atibaia, São Paulo.