CPLP:tuítes – *Corpus* pluricêntrico de tuítes em língua portuguesa

CPLP:tuítes – The pluricentric *corpus* of tweets in Portuguese language

Andressa Rodrigues Gomide[1][0000-0002-1481-4748]

andressa.gomide@fl.uc.pt

[1]CELGA-ILTEC, Faculdade de Letras, Universidade de Coimbra, Portugal.

Resumo. Este trabalho apresenta o processo de coleta, preparação e publicação do *Corpus* Pluricêntrico de Tuítes em Língua Portuguesa (CPLP:tuítes). O CPLP:tuítes é um *corpus* composto de 125.827 tuítes e um total de 2.633.507 tokens. Os tuítes são provenientes de 53 contas de jornais ou fornecedores de notícias em Angola, Brasil, Cabo Verde, Guiné-Bissau, Moçambique, Portugal e São Tomé e Príncipe. Este *corpus* é parte do Banco de Dados em Português (BDP), um repositório que oferecerá acesso livre a *corpora*, bem como os instrumentos utilizados para prepará-los, com conteúdo em português produzidos nos 9 países onde o português é uma língua oficial. A primeira versão do CPLP:tuítes foi lematizada e etiquetada para classes gramaticais e está disponível via CQPweb, um programa de buscas e análises estatísticas em *corpus* que apresenta uma interface amigável e acessível via navegador de internet, sem necessidade de instalação. O artigo apresenta também uma breve discussão sobre decisões a serem feitas quando se prepara um *corpus* de referência para representação de uma língua pluricêntrica em suas diversas variedades.

Palavras-Chave: português língua pluricêntrica, compilação de *corpus*, tweets de notícias.

Abstract. This work presents the process of collecting, preparing and publishing the Pluricentric *Corpus* of Tweets in Portuguese Language (CPLP:tuítes). CPLP:tuítes is a *corpus* composed of 125,827 tweets and a total of 2,633,507 tokens. The tweets come from 53 newspaper accounts or news providers in Angola, Brazil, Cape Verde, Guinea-Bissau, Mozambique, Portugal, and São Tomé and Príncipe. This *corpus* is part of the Portuguese Database (BDP), a repository that will offer free access to *corpora*, as well as the instruments used to prepare them, with content in Portuguese produced in the 11 countries where Portuguese is an official language. The first version of CPLP:tuítes was lemmatized and tagged for grammatical classes and is available via CQPweb, a *corpus* search and statistical analysis program that features a friendly and accessible interface via a web browser, with no installation required. The article also presents a brief discussion on decisions to be made when preparing a reference *corpus* for the representation of a pluricentric language in its many varieties.

Keywords: Portuguese as pluricentric language, *corpus* compilation, news tweets.

1 Introdução

Este artigo apresenta o *Corpus Pluricêntrico de Tuítes em Língua Portuguesa* (CPLP:tuítes), bem como seu processo de compilação.

O CPLP:tuítes é o primeiro lançamento de uma série de *corpora* de textos escritos nas diferentes variedades do português. Essa série, Banco de Dados em Português (BDP), oferecerá acesso livre a *corpora*, bem como os instrumentos utilizados para prepará-los, com conteúdo produzido nos nove países (Angola, Brasil, Cabo Verde, Guiné-Bissau, Guiné Equatorial, Moçambique, Portugal, São Tomé e Príncipe, e Timor-Leste) onde o português é uma língua oficial. Apesar de ser uma língua pluricêntrica, ou seja, uma língua com vários centros nacionais de interação (Clyne, 1992), e falada por mais de 260 milhões de pessoas (Reto et al., 2016), há ainda poucos recursos linguísticos disponíveis para o estudo em conjunto do português e suas variedades.

Espera-se que os recursos presentes no BDP sejam uma fonte confiável e representativa do português e suas variedades e de utilidade para vasta gama de perfis de usuário, preenchendo, de certa forma, esta lacuna. Para ser uma fonte confiável, é importante que o os *corpora* que constituem o BDP sejam criados de forma reproduzível, e que os textos sejam estruturados e disponibilizados em um formato que não exija conhecimento computacional avançado do usuário final. Para contemplar múltiplas áreas dos estudos linguísticos, o BDP incluirá textos de diversos tipos de registros como acadêmico, literário, jornalístico, parlamentar, e textos produzidos e publicados na internet, inspirando-se em, mas não se restringindo à estrutura proposta por Janssen et al. (2018).

As duas próximas seções abordam os elementos básicos da criação de *corpora* geral (2.1) e de tuítes (2.2) e os passos tomados para a criação e publicação do CPLP:tuítes (3 e 4). A seção 5 apresenta o *corpus* e sua estrutura, e é seguida de uma breve discussão conclusiva (seção 6).

2 Criação de *corpora*: estratégias

2.1 Amostragem, representatividade e equilíbrio

Para criar uma representação o mais próxima possível de uma língua em uso, deve-se procurar sempre manter um equilíbrio entre amostras de textos das variedades que o *corpus* deve representar. Para garantir um certo equilíbrio, deve-se levar em consideração variáveis como data de publicação, autores e tipo de registro (ver Leech, 2007). Além de não ser uma tarefa fácil, não há critérios pré-estabelecidos ou instruções explícitas de como cumprir essa tarefa. Por essa e outras razões, um *corpus* não é nunca perfeito, mas é, na melhor das hipóteses, uma representação razoável do repertório completo de texto que se procura estudar (Kilgariff et al., 2006, p. 129). Portanto, um *corpus* é sempre, em maior ou menor medida, tendencioso (Nelson, 2010, p. 57).

O desafio de criar um *corpus* torna-se ainda maior quando buscamos representar uma língua em suas diversas variedades, presentes em países de contextos muito distintos entre si. No contexto do português, alguns dos vários obstáculos para alguns países são: poucos textos disponíveis eletronicamente; alta presença de outras línguas, oficiais ou não, no mesmo discurso; instauração recente do português como língua oficial; ausência de material oficial de referência do português para todos os países.

2.2 *Corpus* de tuítes

Considerando as adversidades mencionadas acima, criar um *corpus* de tuítes mostrou-se uma solução exequível e uma boa forma de pilotar a criação dos outros *corpora* do BDP. Criar um *corpus* de tweets é uma prática crescente e relativamente acessível, especialmente com o surgimento do programa de acesso ao Twitter para pesquisa acadêmica[1]. Esse programa oferece, entre vários benefícios, a busca e coleta gratuita de tuítes sem restrição de data de publicação e com o alto limite de 10 milhões de tuítes por mês.

Entre as vantagens de utilizar o Twitter como fonte para a criação de *corpus*, podemos citar (i) a facilidade de coleta; (ii) a sistematização e padronização dos dados, uma vez que os tuítes possuem a mesma formatação e tamanhos semelhantes; e (iii) a transparência a respeito dos direitos de coleta dos dados. Essas vantagens são particularmente úteis para a criação de um *corpus* piloto. Isso porque os dados são obtidos e preparados de forma automatizada. Assim, uma vez criados os códigos para coleta e formatação dos dados, o processo pode ser repetido com facilidade, mesmo alterando variáveis como contas do Twitter a serem buscadas e data de publicação.

A relativa facilidade de compilar um *corpus* de tweets traz com ela alguns reveses. Algumas dessas adversidades são, felizmente, contornáveis (ver 4.1).

Do ponto de vista da análise linguística, uma das limitações em se usar um *corpus* de tuítes está no curto tamanho dos mesmos, que não permite a realização de certos tipos de análise. Um *corpus* de tweets pode ser uma boa solução por exemplo para identificar neologismos (ex.: Bhosale, 2015; Sang, 2016), ou para explorar tendências no uso da língua (ex.: Sadasivuni & Zhang, 2019). Porém estudos que vão além da estrutura da frase são evidentemente impraticáveis.

[1] https://developer.twitter.com/en/products/twitter-api/academic-research

Ainda na estrutura do tuíte, enfrentamos também a questão da definição de "texto". Para análises quantitativas e estatísticas de *corpora* de postagens curtas, não há uma clara definição de como devemos dispor cada unidade. No caso dos tuítes, podemos agrupá-los por usuários, data de publicação, hashtags, etc. Porém esta decisão pode variar de acordo com o objetivo de quem está a usar o *corpus*.

Uma outra questão que não é clara é o que constitui uma palavra. Casos como a utilização de símbolos para fazer referência a outros usuários (@) e uso de hashtags (#) afetam a lista final de frequência de palavras em um *corpus*. Por exemplo, a palavra "basquetebol" ocorre 202 vezes no *corpus* enquanto a hashtag #basquetebol ocorre 158. Para facilitar a busca no *corpus* por todas as ocorrências de "basquetebol", precedida ou não de #, há algumas soluções de como preparar o *corpus*. Uma possível solução é lematizar (reduzir de forma automática uma palavra à sua forma base) essas ocorrências, eliminando tais símbolos. Uma solução mais simples, e por isso adotada na preparação deste *corpus*, é a de simplesmente adicionar um espaço entre o # e a palavra. Decisões como essas devem ser feitas de forma a satisfazer um número máximo de usuários do *corpus*.

Do ponto de vista técnico e extralinguístico, há decisões a serem tomadas, por exemplo, em relação a tuítes que contêm apenas links[2] ou imagens; a tuítes que são republicação de tuítes de outros usuários (*retweets*); e textos repetidos em vários tuítes sem mudanças significativas (*boilerplates*). A seção 4.1 tratará destas questões.

3 Coleta

Tendo em vista o público-alvo do *corpus* (seção 1), os critérios estabelecidos para seleção das contas foram (i) possuir uma escrita formal ou próxima do formal, ou seja, que uma escrita que procure seguir as normas gramaticais; (ii) não apresentar tópicos potencialmente tendenciosos; (iii) ter um certo prestígio no país de origem. Este último critério foi verificado principalmente através do número de seguidores, postagens publicadas e selo de verificação do Twitter[3].

Esses parâmetros levaram a decisão de criar um *corpus* com tuítes publicados por contas (ou perfis) de jornais ou fornecedores de notícias. Para manter um equilíbrio entre a amostragem, estabeleceu-se que cada variedade do português seria representada por seis contas diferentes de jornais. Para Guiné-Bissau e São Tomé e Príncipe, não foram encontradas contas suficientes com um número considerável de postagens. Por essa razão, esses dois países são representados por mais de seis contas (ver tabela 1). Na ausência de fontes oficiais de notícias, fontes de contas privadas com teor jornalístico foram utilizadas. Todos os casos especiais foram indicados nos metadados e essas informações podem ser facilmente recuperadas ao consultar o *corpus* e utilizadas como filtro.

Para esta versão do *corpus*, Guiné Equatorial e Timor-Leste não foram incluídos por possuírem mais de uma língua oficial. A presença de tuítes em mais de uma língua dificultou o processo de busca e limpeza dos dados. Espera-se, porém, incluir esses dois países em uma próxima versão do *corpus*.

Uma vez estabelecidas as contas, a coleta foi realizada utilizando um código escrito na linguagem de programação Python[4]. O código utiliza a Interface de Programação de Aplicação (API) do Twitter para acessar os conteúdos de forma programática. Com a conexão garantida, o código coleta em ordem cronológica decrescente um máximo de 3500 tuítes da conta escolhida. O resultado enviado pelo Twitter em formato JSON[5] é então convertido para o formato XML[6], preservando o número de identificação e nome de usuário do autor do tuíte, bem como a data e hora de publicação. A conversão do formato para XML é feita por esse ser um formato usual para criação de *corpus* (Hardie, 2014).

A coleta programática via código ajuda a assegurar a replicabilidade da criação do *corpus*. Além disso, a criação e compartilhamento do código permite que novas coletas sejam feitas sem que muito esforço deva ser feito.

---

[2] Ex.: https://twitter.com/CartamzOficial, https://twitter.com/novojornalao, https://twitter.com/santiagoeditora
[3] https://help.twitter.com/pt/managing-your-account/about-twitter-verified-accounts
[4] https://github.com/andressarg/cplp_tweets
[5] *JavaScript Object Notation*
[6] *eXtensible Markup Language*

4 Preparação

4.1 Limpeza

Uma vez coletados os dados, os arquivos XML obtidos após a coleta foram sistematicamente limpos utilizando códigos também escritos em Python. Tuítes constituídos apenas por links ou imagens e *retweets* foram eliminados.

Após a remoção dos tweets indesejados, procedeu-se à limpeza de ruídos no texto. *Boilerplates* frequentemente encontrados em tuítes como “Acabei de ver”, “Clique para ver também”, “NOTICIAS AO MINUTO:” foram eliminados para não enviesar as frequências de palavras.

Por último, todos os links foram encapsulados em uma etiqueta XML[7]. Essa formatação permite que endereços de site (e as palavras que os compõem) possam ser recuperáveis e visualizados no contexto em que ocorrem (ver figura 1) sem que entrem na contagem de palavras do *corpus*.

	Solution 1 to 50	Page 1 / 14
de rega equipados com painéis solares , moto bombas e depósitos de	água	com capacidade de 1000 litros . [https://t.co/QHw3jZQVEL]
da lista de deputados da UNITA [https://t.co/T5FF2aqTIU] Escolas , postos de saúde e	água	entre as prioridades dos munícipes de Nharea [https://t.co/fW
[https://t.co/QHThuNJjwX] Centralidade do Lobito pode “ sair do sufoco ” da falta de	água	[https://t.co/aTh0VbbVPp] Mais 5 mil famílias já beneficiara
ino da língua Kimbundu e a Bibliografia “ distorcida ” [https://t.co/dcQNMVmOfB] A	Água	do Cunene (a salvação da seca severa) [https://t.co/KGqZjbJ
tps://t.co/8iH7FjzPbU] Suíços ajudam Química Verde a acudir mais comunidades com	água	de biofiltros [https://t.co/eVrw8aUOIb] Manuel Gonçalves de
/t.co/zMCd230ALv] PIIM aplica mais de Kz 100 milhões em escola e população pede	água	[https://t.co/wIVeAGVgqK] Familiares da sub-gerente do BI

Figura 1: Linhas de concordância com exibição de links.
Fonte: Elaboração própria.

4.2 Anotação, formatação e publicação

Posteriormente à limpeza dos textos, o *corpus* foi lematizado e etiquetado para classes gramaticais utilizando a rotina *Spacy*[8] com modelo treinado para português. Os textos foram também convertidos para o formato vertical (VRT). Neste formato, cada palavra (ou *token*) é representado em uma linha e suas respectivas anotações (classe gramatical e lema, no caso do CPLP:tuítes) nas colunas subsequentes ao *token* (ver figura 2).

O formato VRT foi escolhido por ser formato utilizado pelo Open *Corpus* Workbench (CWB) (Christ, 1994), um poderoso sistema de busca em textos utilizado por ferramentas amigáveis de consulta de *corpora* como AC/DC (Santos & Bick, 2000), TEITOK (Janssen, 2016) e CQPweb (Hardie, 2012).

Essa primeira versão do CPLP:tuítes está disponível via CQPweb[9] e pode ser consultada livremente. Esta plataforma foi escolhida principalmente pela vasta gama de funções (ex.: distribuição, colocações, contraste de lista de frequências) disponível via interface amigável e acessível via navegador de internet, sem necessidade de instalação. Além disso, o CQPweb é utilizado por pesquisadores em várias línguas, incluindo o português (ex.: Hagemeijer et al., 2022), o que o faz um programa relativamente bem estabelecido e conhecido na linguística de *corpus*.

```
<text id="tweet66414" pt="br" group="other" subgroup="twitter" author="folha" tweetid="1541839858427109376" date="2022-06-28T17:44:02.000Z" year="2022">
CRISE   CRISE   PROPN
ENERGÉTICA  ENERGÉTICA  PROPN
:   :   PUNCT
Governo Governo PROPN
sanciona    sancionar   VERB
lei lei NOUN
que que PRON
devolve devolver    VERB
imposto imposto NOUN
cobrado cobrar  VERB
na  em o    ADP
conta   conta   NOUN
de  de  ADP
luz luz NOUN
<link val="https://t.co/ebZ20PlKHW"/>
</text>
```

Figura 2: Texto do *corpus* anotado e em formatação VRT.
Fonte: Elaboração própria

[7] <link val="www.exemplo.com"/>
[8] https://spacy.io/models/pt
[9] Disponível em https://ola.unito.it/CQPweb32/cplp_tweets/

5 O *corpus*

A versão final do *corpus* após limpeza e formatação contém 125.827 tuítes, e um total de 2.633.507 *tokens* (ver tabela 1). Como descrito na seção três, a proposta inicial era obter cerca de 3500 tuítes para seis contas distintas para cada um dos países.

Porém, para os países Guiné-Bissau e Timor-Leste, apenas três contas para cada país que contemplavam esse critério foram encontradas. Para manter um número semelhante de tuítes para cada país, optou-se por selecionar contas com um discurso semelhante àqueles encontrados em tuítes jornalísticos. Todos esses textos foram marcados como tais, e os usuários podem filtrar a busca de acordo com os critérios desejados (ver figura 3). Esta forma de organização dos dados permite "flexibilizar" os critérios de coleta de textos em favor de obter-se um maior número de palavras sem sacrificar as especificidades do *corpus*, uma vez que o usuário final pode optar por não usar o *corpus* em usa totalidade.

Como demonstrado na tabela 1, apenas as contas brasileiras e portuguesas possuem o selo de verificação do *Twitter*. Isso deve-se provavelmente ao fato de o Twitter não ter grande presença nos outros países. Essa observação corrobora a ideia de que a compilação de um *corpus* pluricêntrico nem sempre pode ter critérios muito específicos para a seleção e coleta de textos, uma vez que os países não compartilham inteiramente os tipos de registros textuais que consomem e produzem. Esse fato enfatiza também a importância da comunicação entre países lusófonos para identificação de fontes ideais a serem utilizadas na criação de *corpora* para o BDP.

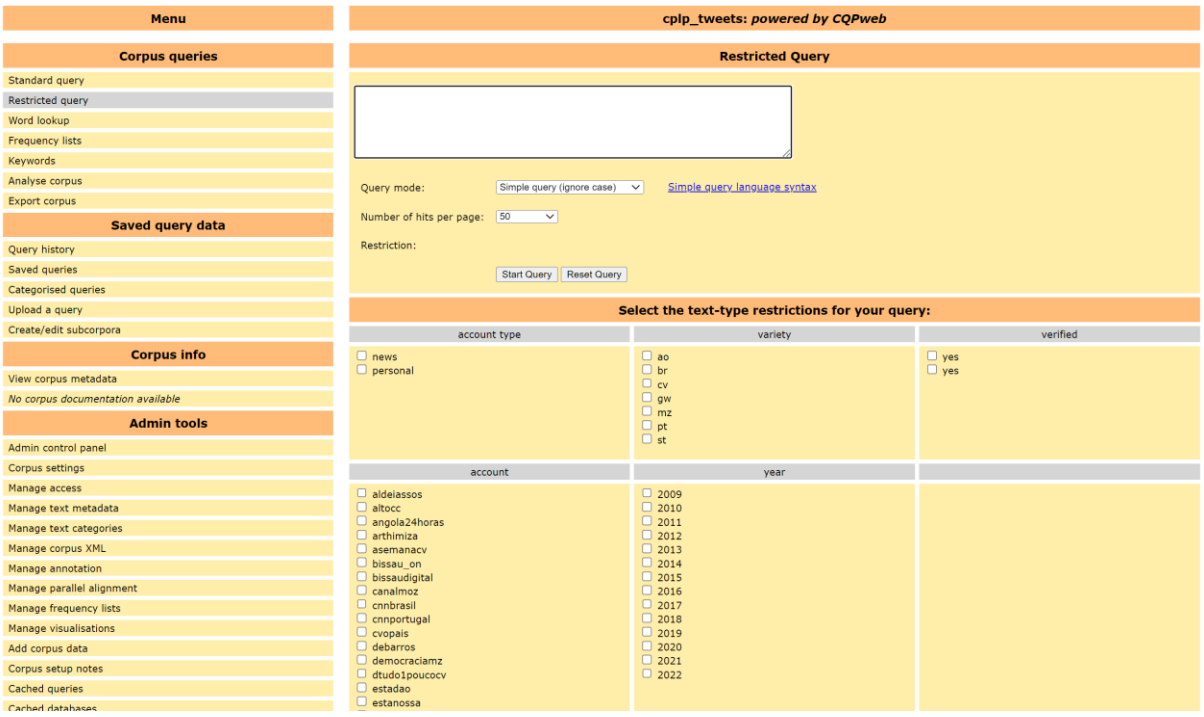

Figura 3: Interface de busca com filtros por metadados.
Fonte: Elaboração própria

Tabela 1: Distribuição de *tokens* e textos no CPLP:tuítes por conta e país.

contas	***tokens***	**textos**	**variedade**	***tokens***	**textos**
angola24horas*	55050	3250			
jornaldeangola*	61855	3172	Angola	407920	19414
jornalfolha8*	95435	3244			

jornalopais*	32959	3250			
portalangop*	37421	3248			
portaldeangola*	125200	3250			
cnnbrasil	79639	2915	Brasil	454948	18447
Estadao	59730	2957			
folha	68503	3141			
g1	67424	3134			
GloboNews	112937	3078			
UOLNoticias	66715	3222			
ASemanaCV*	49292	3250	Cabo Verde	357872	18133
CvOpais*	41323	3250			
dtudo1poucocv	116552	3243			
expressocv	44050	3250			
inforpresscvp	70945	3250			
rtc_caboverde	35710	1890			
aldeiassosgw*º	4510	113	Guiné-Bissau	342073	15600
AltoCCovid19*º	2926	97			
arthimizareal*º	14419	448			
bissau_on*	6874	380			
bissaudigital*	8267	840			
debarros_miguel*º	112774	2781			
EstaNossa*º	8555	262			
guineendade*	50681	3200			
mad95_*º	40391	815			
i_mandjam*º	1249	48			
odemocrata_gb*º	45500	3250			
Odemocratagb*º	40779	3200			
OPalmeirinha*º	1295	38			
UNGuinea_Bissau*º	3853	128			
Canal_Moz*	69247	3200	Moçambique	413048	19031
Democraciamz*	76568	2941			
falamocambique*	55096	3210			
newsmocambique*	36657	3200			
opaisonline*	79598	3250			
verdademz*	95882	3230			
cnnportugal	104238	3250	Portugal	400819	19440
expresso*	111134	3247			
observadorpt	42960	3240			
publico	47623	3204			
rtpnoticias	45516	3249			
sicnoticias	49348	3250			

guia_st*º	5342	349			
jornalst*	20103	2070			
mariolopes_stp *º	41171	2286			
nacoesS*º	25988	699			
oms_stp*º	7076	173	São Tomé	258597	15762
radiostp*	49079	3250			
santola_nation*º	15688	485			
stpdigital*	54183	3200			
telanonstp*	39967	3250			
			Total	2635277	125827

Nota: *, conta sem selo de verificação do Twitter; º, conta não jornalística.

Fonte: Elaboração própria.

6 Discussão

Este artigo apresentou um novo *corpus* do português em sete das suas variedades nacionais, disponível de forma aberta a todos usuários. Espera-se que o uso do *corpus* por diferentes perfis de usuários possa trazer novas discussões e soluções no contexto da criação de *corpora* de referência geral para o português como língua pluricêntrica.

Além do *corpus* per se, este trabalho tem também como produto os códigos e os passos realizados para a criação do *corpus*. A criação de um *corpus* de referência para uma população tão plural como é a dos lusófonos exige níveis tão diversos de conhecimento que as decisões não devem ser tomadas apenas por quem coleta e prepara os textos. Por isso, é de suma importância que a criação dos recursos seja feita de forma contínua e colaborativa. Isso significa que não apenas o *corpus* deve ser disponibilizado para os usuários finais. Os passos para criação de tais recursos também devem ser transparentes, bem documentados, e disponíveis de forma aberta para toda a comunidade.

Referências

Bhosale, M. (2015). *Detecting Neologisms in Twitter*. Dissertação de Mestrado, University of Maryland, Baltimore County, Estados Unidos.

Christ, O. (1994). A modular and flexible architecture for an integrated *corpus* query system. In *Papers in Computational Lexicography* (pp. 22–32), Budapest: Research Institute for Linguistics.

Clyne, M. (2012). Pluricentric Languages – Introduction. In M. Clyne (Ed.), *Pluricentric Languages: Differing Norms in Different Nations* (pp. 1-10). Berlin, Boston: De Gruyter Mouton.

Hagemeijer, T., Mendes, A., Gonçalves, R., Cornejo, C., Madureira, R., & Généreux, M. The PALMA *Corpora* of African Varieties of Portuguese. In *Atas de 13th Conference on Language Resources and Evaluation* (pp. 5047–5053). Marseille: LREC 2022.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a *corpus* analysis tool. *International journal of corpus linguistics*, 17(3), 380-409.

Hardie, A. (2014). Modest XML for *Corpora*: Not a standard, but a suggestion. *ICAME Journal*, *38*(1), 73-103.

Janssen, M., Kuhn, T. Z., Ferreira, J. P., & Correia, M. (2018). The CPLP *Corpus*: A pluricentric *corpus* for the common Portuguese spelling dictionary (VOC). In Čibej, Jaka, Gorjanc, Vojko, Kosem, Iztok & Krek, Simon (eds.) *Atas de 28º EURALEX International Congress* (pp. 835-840). Ljubljana: Ljubljana University Press, Faculty of Arts.

Janssen, M. (2016). TEITOK: Text-faithful annotated *corpora*. In *Atas de Proceedings of the Tenth International Conference on Language Resources and Evaluation,* (pp. 4037-4043). Portorož: LREC 2016.

Kilgariff, A., Rundell, M. and Uí Dhonnchadha, E. (2006). Efficient *Corpus* Development for Lexicography: Building the New *Corpus* for Ireland. *Language Resources and Evaluation* 40, 127–52.

Leech, G. (2007). New resources, or just better old ones? In M. Hundt, N. Nesselhauf & C. Biewer (eds.) *Corpus Linguistics and the Web* (pp. 134–49). Amsterdam: Rodopi.

Nelson, M. (2010) Building a written *corpus*: What are the basics? In A., Keeffe, & M. McCarthy (eds.), *The Routledge handbook of corpus linguistics* (pp. 53:65). London New York, NY: Routledge.

Reto, L., Machado, F.L & Esperança, J.P. (2016). *Novo Atlas da Língua Portuguesa*. Lisboa: Imprensa Nacional - Casa da Moeda

Sadasivuni, S. T., & Zhang, Y. (2019). Analyzing Tweets to Discover Twitter Users' Mental Health Status by a Word-Frequency Method. In *Atas de IEEE International Conference on Intelligent Systems and Green Technology* (pp. 5-53). IEEE.

Sang, E. T. K. (2016). Finding rising and falling words. In E. Hinrichs, M. Hinrichs & T. Trippel (Orgs.), *Atas de Workshop on Language Technology Resources and Tools for Digital Humanities* (pp. 2-9). Osaka: The COLING 2016 Organizing Committee.

Santos, D. & Bick, E. (2010). Providing Internet access to Portuguese *corpora*: the AC/DC project. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.), *Atas de the Second International Conference on Language Resources and Evaluation,* (pp. 205-210). Atenas: LREC 2010.