Anotação de textos não canônicos: um estudo exploratório de *Grande sertão: veredas* pelas dependências universais

André V. L. Coneglian[1][0000-0003-1726-8890], Ana Luisa A. R. Guimarães[1][0000-0001-7278-6856], Thiago Castro Ferreira[1] [0000-0003-0200-3646], Adriana S. Pagano[1][0000-0002-3150-3503]
[1] Universidade Federal de Minas Gerais, Belo Horizonte MG 31270-901, BRA
coneglian@ufmg.br; alarp@ufmg.br; thiagocf05@ufmg.br; apagano@ufmg.br

Abstract. This paper reports on an exploratory study of a sample of 175 sentences retrieved from the renowned Brazilian novel *Grande Sertão: Veredas* [Portuguese for *Great Backlands: Paths*; English translation: *The devil to pay in the backlands*], which were annotated for POS and syntactic relations following the Universal Dependencies guidelines. The study aimed to explore the feasibility of annotating non-canonical text to create treebanks for Brazilian Portuguese. We computed accuracy and precision of the model in order to verify categories annotated more and less successfully. The results show the model performed slightly better for POS than dependency relations and pointed out categories with higher demand for manual revision as being those related to orality phenomena represented by Guimarães Rosa in his novel. The study shows the potential of annotating non-canonical text to enhance existing models with categories less represented in the treebanks.

Keywords: Universal Dependencies, non-canonical text, Brazilian Portuguese.

1 Introdução

A seleção de textos para fins de anotação e treinamento de modelos visando o Processamento de Língua Natural (PLN) tem privilegiado um conjunto limitado de tipos, dentre os quais se destaca o texto jornalístico com expressiva presença nos *treebanks* existentes nas distintas línguas (Silveira et al., 2014; Plank, 2016; Zeldes & Simonson, 2016; Bamman, 2020). Textos técnicos de domínios específicos, como biomedicina, também estão representados nos *corpora* disponíveis e, juntamente com os jornalísticos, constituem “textos canônicos” em PLN. Por outro lado, os chamados textos "não canônicos" vêm se fazendo cada vez mais presentes, e, geralmente, caracterizam-se como de difícil anotação tanto em termos de padrões gramaticais quanto em termos de limitações que esquemas de anotação apresentam para lidar com eles (Hirschmann et al., 2007). São considerados textos "não canônicos" aqueles que são extraídos das chamadas mídias sociais, como *Twitter*, fóruns e *blogs*, bem como narrativas clínicas, textos produzidos por aprendizes de língua e textos literários clássicos ou contemporâneos.

Em português brasileiro, língua considerada com recursos insuficientes para o desenvolvimento de aplicações em PLN, os *treebanks* existentes constituem-se de

textos jornalísticos e, mais recentemente, de textos de mídias sociais, bem como de narrativas clínicas. Não existe ainda um *treebank* de texto literário e, muito menos, de texto literário com uso não convencional da linguagem.

Este estudo apresenta uma anotação exploratória de uma amostra de um texto literário não convencional, o romance *Grande sertão: Veredas,* de João Guimarães Rosa, publicado em 1959, no Brasil, e caracterizado pela crítica literária como inovador tanto no léxico quanto na gramática. Enquanto proposta de uso criativo da linguagem, o romance rosiano também é considerado "não canônico" no escopo do PLN. O objetivo central é explorar a viabilidade de se anotar textos não canônicos, de natureza literária, para desenvolver *treebanks* em português brasiliero pautados pelas diretrizes do projeto Universal Dependencies.

A anotação das relações de dependência que se faz neste estudo vai na direção de explorar as fronteiras entre uma sintaxe da sentença e uma sintaxe do discurso (Kaltenböck et al., 2011; Heine et al., 2013; Heine et al., 2017), uma vez que a interpretação da escrita rosiana desafia as fronteiras desses dois domínios. Assim, para casos limítrofes em que o analista tem de decidir entre uma ou outra função, as decisões de anotação são tomadas considerando-se o co-texto da sentença analisada, isto é, as sentenças anteriores, não apenas o que se apresenta na sentença analisada.

O artigo está organizado em 5 seções, além desta Introdução. Na Seção 2, discute-se o estatuto da canonicidade dos textos em PLN, destacando-se as singularidades de *Grande sertão: Veredas*. Na Seção 3, apresenta-se a Metodologia deste estudo. A Seção 4 apresenta os resultados da anotação, bem como uma caracterização linguística da amostra analisada, apontando-se as categorias nas quais o modelo de anotação automática teve boa performance e aquelas que desafiam o modelo, por estarem menos representadas. Na seção 5, os resultados são interpretados e discutidos com base nas considerações sobre o caráter não canônico do texto analisado neste estudo. A seção 6 recupera e sintetiza os principais argumentos desenvolvidos no estudo.

2 A linguagem de *Grande sertão: Veredas*

2.1 O estatuto "não canônico" dos textos em PLN

No campo do PLN, a canonicidade de um tipo de texto parece ser determinada pelo desvio observado em relação aos padrões mais convencionais de um sistema linguístico ou bem pela dificuldade que seu processamento gera aos modelos existentes (Hirschman et al., 2007). Fenômenos gramaticais geralmente rotulados como fora de um padrão representam casos de fluidez categorial e de multifuncionalidade dos itens linguísticos (Neves, 2012). Sua não canonicidade resolve-se, em última instância, em uma questão de como esses fatos são acomodados teoricamente.

A avaliação que geralmente se faz da linguagem rosiana, especialmente em *Grande sertão: Veredas*, aponta o léxico como requintado e a sintaxe como dificultosa. O léxico de *Grande sertão: Veredas* é resultado de um trabalho criativo, que, mesclando raízes e morfemas de diferentes línguas, resulta em inovações lexicais e fraseológicas, muito bem documentadas lexicograficamente em Martins (2001). O aproveitamento para os *treebanks* é evidente nesse ponto, destacando-se, principalmente, a ampliação lexical dos bancos de dados, por meio da anotação do sistema de classes de palavras, com o auxílio de obras lexicográficas (Martins, 2001). No que diz respeito à sintaxe, a

complexidade do texto de *GS:V* parece ser decorrente de "inversões e elipses" e de "construções carregadas de ênfase" (Martins, 2001, p. XI). Ocorre que a sintaxe de Guimarães Rosa nunca viola as possibilidades sistêmicas da sintaxe portuguesa; antes, o autor encontra espaços de manobra que lhe permitem explorar as estratégias construcionais do português.

2.2 Aspectos lexicogramaticais da obra

Para entender a constituição linguística de *Grande sertão: Veredas* é necessário recorrer à própria configuração da obra. O romance se configura com uma conversa monológica entre o narrador-personagem, Riobaldo, e seu interlocutor, interpelado como "senhor", a quem conta histórias da sua época de jagunço no sertão mineiro, como se mostra em (1). Esse ponto de partida é importante para a consideração da linguagem na obra, porque o que Rosa faz é construir a "linguagem falada" pelo narrador num projeto literário de explorar as possibilidades lexicais e gramaticais em múltiplos registros de um português plurilíngue (Rocha, 2021), como se vê em (2).

(1) O senhor aprova? Me declare tudo, franco — é alta mercê que me faz: e pedir posso, encarecido. (GS:V)
(2) Hem? Hem? Ah. Figuração minha, de pior pra trás, as certas lembranças. (GS:V)

O fato de Rosa configurar o seu romance como se fosse o registro de um relato feito por um jagunço do sertão mineiro implica a incorporação de diversos fenômenos da modalidade falada da língua à modalidade escrita. Assim, as omissões, inversões e elisões que Martins (2001) caracterizou como aspectos dificultosos da sintaxe rosiana são, na verdade e de fato, mecanismos do funcionamento natural da língua falada. Esses mecanismos estão muito bem documentados para o português brasileiro (Castilho, 2002a, 2002b; Ilari, 2002; Castilho e Basílio, 2002; Kato, 2002; Koch, 2002; Neves, 1999; Abaurre e Rodrigues, 2002). Dizem Tarallo, Kato et al (1989, p. 25 – destaque original) que "uma primeira tomada de contato de um *corpus* natural de linguagem oral leva-nos a perceber a quantidade de emissões que, em relação a uma estrutura canônica do tipo *sujeito* + *predicado* (...)" apresentam-se, em alguma medida, fora dessa estrutura. O trecho em (3) ilustra a caracterização de Tarallo, Kato et al (1989).

(3) Se a gente — conforme compadre meu Quelemém é quem diz — se a gente torna a encarnar renovado, eu cismo até que inimigo de morte pode vir como filho do inimigo. (GS:V)

Em (3), à semelhança do que ocorre na língua falada, o narrador começa a formular seu enunciado com "Se a gente", mas logo o interrompe com a inclusão de uma estrutura adverbial, para depois retomar a formulação do seu enunciado com uma reiteração de "se a gente". Uma outra característica da língua falada presente no romance é a topicalização e deslocamento de constituintes para a posição inicial da sentença, como ilustrado por (4) e (5).

(4) Dono dele nem sei quem for. (GS:V)
(5) Solto, por si, cidadão, é que não tem diabo nenhum. (GS:V)

O desafio é, portanto, acomodar esses fenômenos naturais da constituição dos enunciados linguísticos nos aparatos metodológico-descritivos utilizados em PLN,

como é o modelo de Dependências Universais, adotado nos *treebanks* da maioria das línguas atualmente, incluindo o português brasileiro e que contempla estruturas 'canônicas' para basear e exemplificar suas diretrizes. Iniciativas recentes, todavia, vêm ampliando o leque de tipos de textos anotados para construir treebanks visando aplicações de PLN em português basileiro (DiFilippo et al., 2021; Souza et al., 2021). Nessa perspectiva, este estudo explora o potencial de um texto literário não canônico para expandir e diversificar os *treebanks* existentes, como se detalha a seguir.

3 Metodologia

O córpus anotado e analisado consiste numa amostra das primeiras 175 sentenças do romance. As sentenças foram extraídas de um arquivo em formato pdf, convertidas para o formato txt codificação UTF8 e revisadas manualmente para corrigir problemas de conversão. Para a anotação do córpus, foi utilizado o *framework* do projeto *Universal Dependencies* (UDs) v.2 (Nivre et al., 2015), que consiste em 17 etiquetas para anotação de classes gramaticais e 37 etiquetas de relações sintáticas, além de sub-relações. O córpus foi primeiramente anotado de forma automática por meio da ferramenta UDpipe (Straka et al. 2016), com um modelo de língua portuguesa que utiliza o Bosque-UD v. 2.6 de textos jornalísticos (Rademaker et al. 2017). Os arquivos CONLLU foram importados na ferramenta de anotação Arborator-Grew-NILC (https://arborator.icmc.usp), uma versão expandida e aprimorada de Arborator-Grew (Guibon et al., 2020).

A revisão da anotação automática foi realizada por 3 anotadores familiarizados com a abordagem das UDs. Para a anotação, foram utilizadas, além das diretrizes gerais das UDs, os manuais de anotação do ICMC/USP (Duran 2021a,b). Para embasar a interpretação do texto de Guimarães Rosa, foram consultadas obras sobre a linguagem e o estilo do autor (Rocha, 2021; Sant'Anna Martins, 2021). Nos casos de palavras e funções que podem operar localmente na sentença como ADV e 'advmod' ou como CCONJ e 'cc' numa relação de coordenação com uma sentença anterior, as decisões de anotação foram tomadas considerando-se o co-texto da sentença em pauta, isto é, a sentença anterior.

A revisão foi feita de forma independente pelos três anotadores, ao cabo da qual as divergências foram discutidas em grupo até se chegar a uma anotação final consensual.

Concluída a revisão, os arquivos em formato CONLLU foram exportados da ferramenta Arborator-Grew-NILC e processados por script em linguagem Python para contagem das categorias anotadas. Por meio da biblioteca Scikit-learn, foram computados a porcentagem de precisão, *recall*, *F1-score* e *support* (número total de ocorrências da etiqueta na anotação revisada) para cada categoria, juntamente com a acurácia do modelo utilizado.

A análise enfocou o percentual de acerto do modelo de anotação automática comparado com a nossa revisão manual visando verificar quais categorias foram anotadas de forma mais e menos bem sucedida e o que esses resultados poderiam evidenciar sobre o potencial do texto anotado para expandir e diversificar os treebanks em português brasileiro.

4 Resultados

A amostra de texto selecionada para anotação compreendeu 175 sentenças e 2809 *tokens*, incluindo-se sinais de pontuação. A média de *tokens* por sentença foi 16,05, tendo sido verificado um amplo intervalo de variação no tamanho mínimo e máximo das sentenças anotadas, 2 e 83 *tokens*, respectivamente.

No que diz respeito à anotação de classe de palavras, os resultados da anotação automática e sua revisão pelos anotadores estão dispostos na Tabela 1. A Tabela 1 mostra um alto índice de acerto do modelo de anotação automática, com *F1-score* maior que 85 por cento para a maior parte das categorias, com exceção de INTJ (interjeição), PART (partícula) e X (reservada para *tokens* aos quais não se pode atribuir nenhuma das categorias existentes). No caso das duas últimas, trata-se de ocorrências nas quais o modelo classificou de forma inadequada interjeições (como "eh" e "ah") e alguns sinais de pontuação. Por outro lado, a categoria INTJ teve uma alta precisão, mas baixo *recall*, o que pode ser explicado pelo fato de o modelo possuir um número limitado de lemas para a classe interjeição, os quais não contemplam as formas utilizadas por Guimarães Rosa.

Tabela 1. Taxas de precisão, *recall, F1-scor*e e *support* computadas para a anotação automática de POS.

POS	precision (%)	recall (%)	f1-score (%)	support
ADJ	85,44%	86,27%	85,85%	102
ADP	99,29%	99,29%	99,29%	283
ADV	89,30%	95,43%	92,27%	175
AUX	96,10%	100,00%	98,01%	74
CCONJ	100,00%	91,14%	95,36%	79
DET	99,36%	99,04%	99,20%	314
INTJ	100,00%	6,67%	12,50%	15
NOUN	96,32%	95,49%	95,91%	466
NUM	80,00%	100,00%	88,89%	12
PART	0,00%	0,00%	0,00%	0
PRON	98,45%	95,02%	96,71%	201
PROPN	94,92%	91,80%	93,33%	61
PUNCT	100,00%	98,98%	99,49%	587
SCONJ	93,41%	95,51%	94,44%	89
VERB	92,64%	96,87%	94,71%	351
X	0,00%	0,00%	0,00%	0

No que diz respeito às relações de dependência, a Tabela 2 sintetiza os principais resultados obtidos. Dentre as relações que tiveram um *F1-score* inferior a 75% destacam-se, para efeitos da argumentação deste estudo, as relações de *parataxe* (53,54%), *discourse* (30,77%), *orphan*, *reparandum*, *dislocated* e *dep*, estas quatro

últimas com uma única ocorrência na amostra, que o modelo não classificou de forma correta.

Tabela 2. Taxas de precisão, *recall*, *F1-score* e *support* computadas para a anotação automática de relações de dependência.

deprel	precision (%)	recall (%)	f1-score (%)	support
acl	72,73%	70,59%	71,64%	34
acl:relcl	76,32%	96,67%	85,29%	30
advcl	65,00%	61,90%	63,41%	42
advmod	78,53%	92,67%	85,02%	150
amod	81,16%	80,00%	80,58%	70
appos	48,28%	87,50%	62,22%	16
aux	91,67%	100,00%	95,65%	11
aux:pass	33,33%	100,00%	50,00%	1
case	97,83%	97,83%	97,83%	277
cc	100,00%	88,89%	94,12%	81
ccomp	53,66%	88,00%	66,67%	25
compound	0,00%	0,00%	0,00%	0
conj	83,65%	76,99%	80,18%	113
cop	83,33%	91,84%	87,38%	49
csubj	66,67%	66,67%	66,67%	6
dep	0,00%	0,00%	0,00%	1
det	99,35%	98,07%	98,71%	311
discourse	66,67%	20,00%	30,77%	30
dislocated	0,00%	0,00%	0,00%	2
expl	93,33%	56,00%	70,00%	25
fixed	57,14%	40,00%	47,06%	20
flat:name	85,71%	100,00%	92,31%	6
iobj	92,31%	70,59%	80,00%	17
mark	85,06%	94,87%	89,70%	78
nmod	86,61%	83,62%	85,09%	116
nsubj	84,83%	90,96%	87,79%	166
nsubj:pass	0,00%	0,00%	0,00%	1
nummod	91,67%	100,00%	95,65%	11

obj	74,42%	85,71%	79,67%	112
obl	86,73%	74,81%	80,33%	131
orphan	0,00%	0,00%	0,00%	1
parataxis	64,15%	45,95%	53,54%	74
punct	100,00%	98,98%	99,49%	587
reparandum	0,00%	0,00%	0,00%	1
root	86,86%	86,86%	86,86%	175
xcomp	68,09%	82,05%	74,42%	39

A Tabela 3 apresenta os valores médios obtidos para cada conjunto de etiquetas. Nela observamos performance superior do modelo de máquina para POS com acurácia de 92,26% em contraste com 88,22% para relações de dependências. No entanto, ao avaliar o F1-score obtido, a taxa de acerto cai em média 25 por cento para ambas, demonstrando uma performance mediana especialmente para as relações de dependência. Quanto às medidas de precisão e recall, percebe-se uma leve diferença na avaliação das POS, que pode ser justificada pela dificuldade do modelo em reconhecer os tokens que deveriam receber a etiqueta INTJ e pelo reconhecimento errôneo das etiquetas PART e X, como mencionado previamente.

Tabela 3. Média das taxas de precisão, *recall*, *F1-score* e acurácia calculadas para a anotação automática de cada conjunto de etiquetas.

	precisão (%)	**recall (%)**	**F1-score (%)**	**acurácia (%)**
POS	82,83%	78,22%	77,87%	96,26%
deprels	65,42%	67,44%	64,95%	88,22%

As relações que tiveram um *F1-score* inferior a 75% revelam mecanismos típicos da oralidade, como disfluências, interposição e deslocamento de constituintes, bem como relações cuja dependência não pode ser classificada de acordo com nenhuma categoria já estabelecida pelo modelo das UDs. Os exemplos a seguir ilustram algumas dessas formas e os desafios que apresentam para sua anotação.

A Figura 1 mostra um exemplo no qual se verifica falta de alinhamento entre uma classe de palavras e a relação de dependência da qual participa. O advérbio “depois” é utilizado na sentença, não numa relação *advmod,* mas numa relação *obj*, sendo *head e* possuindo um determinante.

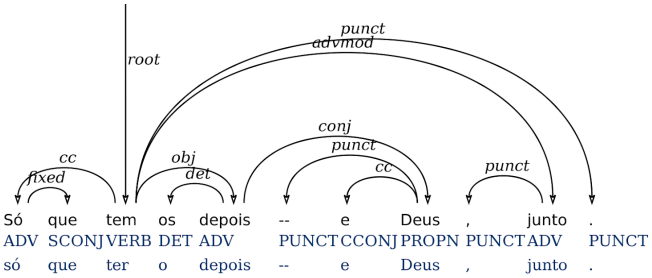

Figura 1. Exemplo de incongruência entre classe e função.

A Figura 2 mostra uma sentença com uma relação *ccomp* envolvendo uma relação na qual um deslocamento de parte de um sintagma nominal demanda o estabelecimento de uma relação não projetiva.

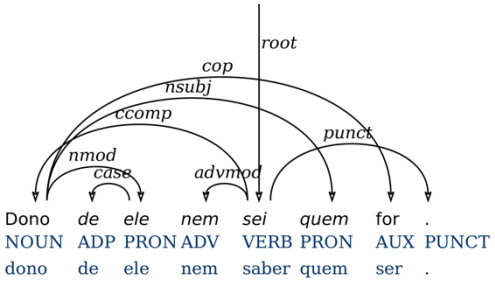

Figura 2. Exemplo de uma anotação com relação não projetiva.

A Figura 3 mostra a anotação de um fragmento de uma sentença, na qual se verifica uma relação de dependência não especificada (*dep*). De acordo com Rocha (2021), "elas se acostumaram a se assim das locas, para papar" envolve o apagamento do verbo "sair". Essa ausência impede atribuir a "se" uma relação específica.

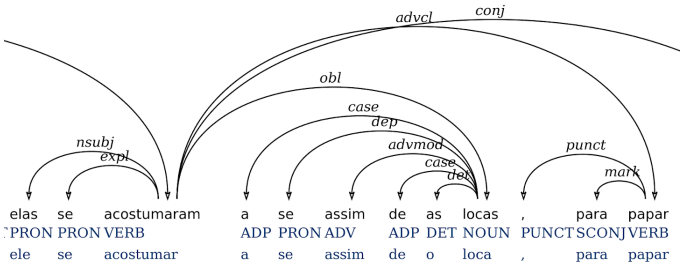

Figura 3. Exemplo da relação *dep*.

5 Discussão

A acurácia do modelo de anotação automática da amostra em questão evidencia, por um lado, o bom desempenho de modelos baseados em textos canônicos, como é o texto jornalístico, para a anotação de textos menos canônicos; por outro, mostra o potencial

do texto não canônico para complementar os modelos existentes em categorias menos representadas nos *treebanks*, evidenciado pelo número pequeno de lemas existentes para INTJ e de instâncias para as relações *discourse, reparandum e dislocated* conforme dados disponibilizados em https://universaldependencies.org/treebanks/pt_bosque/index.html.

Os desafios de anotação, no sentido de total aderência às diretrizes do projeto UD, concentram-se em (i) o não alinhamento entre classe de palavra e função e (ii) configurações sintáticas que demandam anotação não projetiva. Considerando-se que os treebanks disponíveis em português brasileiro que embasam o treinamento do modelo utilizado neste estudo e de outros modelos, em geral, ainda se baseiam em textos canônicos, é possível dizer que os recursos para PLN nesse idioma podem beneficiar-se da anotação de textos não canônicos que contemplem tais desafios, de modo a incrementar a construção de recursos em português brasileiro para PLN.

6 Considerações finais

Este estudo exploratório de *Grande sertão: Veredas* procurou mostrar a pertinência de se incluir textos não-canônicos, dentre eles literários, aos *treebanks* para PLN em português brasileiro. A justificativa para a escolha da amostra é sua representação do registro oral em português, fornecendo dados para expandir o repertório de anotação em português brasileiro pautado pelas Dependências Universais. Longe de representar um desvio, a linguagem rosiana, não canônica do ponto de vista do PLN, tem como matriz o sistema do português brasileiro e muitas das estruturas e do léxico nela presentes permitem incorporar fenômenos da oralidade aos *treebanks*.

Esses fenômenos, tão naturais da língua falada, e, no córpus desta pesquisa, registrado na modalidade escrita, são pervasivos em inúmeros gêneros discursivos em que se verifica essa relativização de fronteiras entre o oral e o escrito (Marcuschi, 2008, 2010; Neves, 2010), como crônicas, postagens em redes sociais e em *blogs*, tipos de textos alvejados em projetos de anotação em PLN.

O corpus de sentenças anotadas está em processo de preparação para sua validação e submissão à comunidade UDs e será também disponibilizado na conta de *github* dos pesquisadores.

Referências

Abaurre, M. B. M., Rodrigues, A. C. S. (orgs.): Gramática do português falado. Vol. 8 – Novos estudos descritivos. Editora Unicamp, Campinas (2002).
Bamman, D.: LitBank: Born-Literary natural language processing.
Castilho, A. T. (org.): Gramática do português falado. Vol. 1 – A ordem. 4a ed. Editora Unicamp, Campinas (2002a).
Castilho, A. T. (org.): Gramática do português falado. Vol. 3 – As abordagens. 3a ed. Editora Unicamp, Campinas (2002b).
Castilho, A. T., Basílio, M. (orgs.): Gramática do português falado. Vol. 4 – Estudos descritivos. 2a ed. Editora Unicamp, Campinas (2002a).
Di Felippo, A. et al.: Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies. In: the Proceedings of the VII Workshop on Portuguese Description (JDP), pp. 335-343. (2021).

Duran, M. S.: Manual de anotação de PoS tags. Relatório Técnico, n. 434. NILC-ICMC/USP, 54p. (2021a) Disponível em: https://sites.google.com/icmc.usp.br/poetisa. Acesso em: 20/09/2021.
Duran, M. S.: Manual de Anotação de Relações de Dependência: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 435. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 79 p. (2021b).
Heine, B., Kaltenböck, G., Kuteva, T., Long, H.: An outline of Discourse Grammar. In: Bischoff, S., Jeny, C. (eds.) Reflections on functionalism in linguistics, pp. 141-157, Mouton de Gruyter, Berlin (2013).
Heine, B., Kaltenböck, G., Kuteva, T., Long, H.: Cooptation as a discourse strategy. Linguistics 55(4), 813-855 (2017)
Hirschmann, H., Doolittle, S., Lüdeling, A.: Syntactic annotation of non-canonical linguistic structure. (2007)
Ilari, R. (org.): Gramática do português falado. Vol. 2 – Níveis de análise. 4a ed. Editora Unicamp, Campinas (2002).
Kaltenböck, G., Heine, B., Kuteva, T.: On thetical grammar. Studies in Language 35(4), 852-897 (2011).
Kato, M. A (org.): Gramática do português falado. Vol. 5 – Convergências. 2a ed. Editora Unicamp, Campinas (2002).
Koch, I. (org.): Gramática do português falado. Vol. 6 – Desenvolvimentos. 2a ed. Editora Unicamp, Campinas (2002).
Marcuschi, L. A.: Produção textual, análise de gêneros e compreensão. Parábola Editorial, São Paulo (2008).
Marcuschi, L. A.: Da fala para a escrita. Atividades de retextualização. 10a ed. Cortez Editora, Campinas (2010).
Martins, N. S.: O léxico de Guimarães Rosa. 3a ed. Edusp, São Paulo (2001).
Marneffe, M. C., Manning, C. D., Nivre, J., Zeman, D. Universal dependencies. Computational Linguistics 47(2), 255-308 (2021).
Neves, M. H. M.: Língua falada e língua escrita. Uma busca da gramática que rege as formulações. In: Neves, M. H. M. Ensino de língua e vivência de linguagem: temas em confronto, p. 151-167, Editora Contexto, São Paulo (2010).
Neves, M. H. M.: As estratégias discursivas e suas implicações na relação entre oralidade e escrita – um estudo do parêntese na crônica. Linguística 27(1), 77-97 (2012).
Neves, M. H. M. (org.): Gramática do português falado. Vol. 7 – Novos estudos. 2a ed. Editora Unicamp, Campinas (1999).
Nivre, J. et al.: Universal Dependencies v2: An ever growing multilingual treebank collection. In: Proceedings of the12th Language Resources and Evaluation Conference, 4034-4043. Marseille, France: European Language Resources Association. (2020)
Plank, B.: What to do about non-standard (or non-canonical) language in NLP. In: Proceedings of the 13th Conference on Natural Language Processing (KOVENS2016), p. 13-20. NLP Association of India, Varanasi (2016).
Rademaker, A. et al.: Universal dependencies for portuguese. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 197–206. (2017).
Rosa, J. G.: Grande sertão: Veredas. 22a ed. São Paulo: Companhia das Letras (2019/1959).
Silveira, N. et al.: A gold standard dependency corpus for English. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, p. 2897-2904. European Language Resources Association, Reykjavik (2014).
Souza, E. et al.: PetroGold – Corpus padrão ouro para o domínio do petróleo. *In*: Anais do 13º Simpósio Braisleiro de Tecnologia da Informação e da Linguagem Humana (STIL), p. 29-38, Porto Alegre: Sociedade Brasileira de Computação (2021).

Straka, M., Hajic, J., Strakov´a, J.: Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4290–4297 (2016).
Tarallo, F., Kato, M. et al.: Rupturas na ordem da adjacência canônicas no português falado. In: Castilho, A. (org.) Gramática do português falado. Vol. 1 – A ordem. 1a ed., pp. 25-52, Editora Unicamp, Campinas (1989).
Zeldes, A., Simonson, D.: Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In: Proceedings of the 10th Linguistic Annotation Workshop, p. 68-78. Association for Computational Linguistics, Berlin (2016).