

Desafios Tipológicos na Construção de Treebanks: O Mapeamento Morfossintático da Língua Kubeo no Padrão UD

Pedro Paulo Freire Oliveira¹, Thiago Blanch Pires²

^{1 2}Instituto de Letras – Universidade de Brasília, Campus Universitário Darcy Ribeiro, Gleba A, Asa Norte – CEP 70910-900 – Brasília – DF – Brasil
pedro.freire@aluno.unb.br, thiagocomaga@gmail.com

Abstract. *The inclusion of Amazonian indigenous languages in the field of Natural Language Processing (NLP) is hindered by the lack of treebanks and structured resources. This paper describes the initial phase of morphosyntactic modeling of the Kubeo language (Eastern Tukanoan family) under the Universal Dependencies (UD) guidelines. Kubeo exhibits high typological complexity, characterized by agglutinative processes, pervasive nominal classifiers, and pragmatic markers that challenge the static grammatical categories of the UD standard. The scope of this stage focuses on the extraction and tagging of native sentences, with a methodological emphasis on measuring the structural distance between the explicit nominal subject and the core verb. Based on data from the Pamiene Mama Buei-Tukubo primer, the study proposes token mapping solutions, utilizing the morphological features layer FEATS to preserve the language's morphosyntactic richness without excessively fragmenting the dependency trees. This pioneering corpus aims to establish a reproducible annotation protocol for future Kubeo computational tools.*

Resumo. *A inserção de línguas indígenas amazônicas no campo do Processamento de Linguagem Natural (PLN) esbarra na ausência de treebanks e recursos estruturados. Este artigo descreve a fase inicial de modelagem morfossintática da língua Kubeo (família Tukano-Oriental) sob as diretrizes das Dependências Universais (UD). O Kubeo apresenta alta complexidade tipológica, caracterizada por processos aglutinantes, classificadores nominais pervasivos e marcadores pragmáticos que desafiam as categorias gramaticais estáticas do padrão UD. O escopo desta etapa concentra-se na extração e etiquetagem de sentenças nativas, com foco metodológico na mensuração da distância estrutural entre o sujeito nominal explícito e o verbo núcleo. A partir de dados da cartilha Pamiene Mama Buei-Tukubo, o estudo propõe soluções de mapeamento de tokens, utilizando a camada de traços morfológicos FEATS para preservar a riqueza morfossintática da língua sem fragmentar excessivamente as árvores de dependência. A criação deste corpus visa estabelecer um protocolo de anotação reprodutível para futuras ferramentas focadas no Kubeo.*

1. Introdução

A construção de recursos computacionais para línguas de baixo recurso constitui um dos principais desafios da Linguística Computacional, especialmente no caso das línguas indígenas brasileiras, para as quais ainda há uma carência crítica de corpora morfossintaticamente anotados e *treebanks*, limitando o desenvolvimento de ferramentas de Processamento de Linguagem Natural (ALENCAR, 2021; ALEXANDRE et al., 2021; CASELI; NUNES, 2022)

Entre essas línguas, o Kubeo, pertencente à família Tukano-Oriental e falado na região do alto rio Uaupés, apresenta características tipológicas que representam desafios relevantes para a anotação computacional, como morfologia aglutinante, evidencialidade

gramaticalizada, classificadores nominais e relativa flexibilidade sintática. Embora a língua conte com uma descrição linguística consolidada, sobretudo nos trabalhos de Chacon (2012), ainda não dispõe de um *treebank* compatível com o padrão *Universal Dependencies* (UD), o que restringe sua integração ao ecossistema internacional de recursos linguísticos.

Este trabalho integra um projeto de pesquisa voltado ao desenvolvimento de recursos computacionais para línguas indígenas brasileiras e utiliza como corpus inicial sentenças extraídas da cartilha *Pamiene Mama Buei-Tukubo* (2023), produzida no âmbito das ações de documentação e ensino da língua Kubeo conduzidas por Thiago Chacon e colaboradores. Nessa perspectiva, o objetivo desta etapa é estabelecer uma metodologia consistente de anotação morfossintática, capaz de subsidiar a expansão do corpus e o desenvolvimento de ferramentas de PLN para a língua.

O processo de anotação adota como referência o modelo *Universal Dependencies* (Nivre et al., 2016; de Marneffe et al., 2021) atualmente um dos principais padrões internacionais para representação sintática de corpora anotados. Entretanto, sua aplicação a línguas tipologicamente distantes das línguas indo-europeias levanta desafios metodológicos conhecidos na literatura brasileira como Problema de Anchieta (SANDALO; GALVES, 2017), relacionados às dificuldades de representar sistemas morfossintáticos tipologicamente distintos por meio de categorias gramaticais universais. No caso do Kubeo, esses desafios concentram-se principalmente na representação de morfemas multifuncionais e na necessidade da utilização extensiva da camada de traços morfológicos FEATS, seguindo as recomendações do próprio projeto UD.

Diante desse cenário, este artigo apresenta a etapa inicial da construção de um *treebank* para a língua Kubeo compatível com UD. São descritos o protocolo inicial de anotação desenvolvido para um conjunto piloto de sentenças, as principais decisões metodológicas adotadas durante o processo de anotação e uma análise preliminar da distância estrutural entre sujeito nominal explícito e verbo núcleo. Ao sistematizar esses procedimentos, o trabalho estabelece uma base metodológica para a ampliação do corpus e para o desenvolvimento de recursos computacionais destinados ao PLN da língua Kubeo.

2. Fundamentação Teórica

Universal Dependencies (UD) é um esquema de anotação morfossintática concebido para representar diferentes línguas por meio de um conjunto padronizado de categorias gramaticais, relações de dependência e traços morfológicos, favorecendo a interoperabilidade entre corpora e o desenvolvimento de ferramentas multilíngues de Processamento de Linguagem Natural (de Marneffe et al., 2021). O modelo organiza as anotações no formato CoNLL-U, estruturado em dez colunas que registram informações como forma, lema, classe gramatical, traços morfológicos e relações sintáticas, permitindo a construção de *treebanks* compatíveis entre diferentes projetos linguísticos.

Nos últimos anos, o UD tem sido progressivamente aplicado a línguas de baixo recurso, incluindo algumas línguas indígenas brasileiras, demonstrando o potencial do modelo para apoiar tanto estudos tipológicos quanto aplicações computacionais. Entretanto, o número de *treebanks* disponíveis ainda é reduzido quando comparado ao de línguas amplamente documentadas, tornando a ampliação desses recursos uma demanda relevante para a Linguística Computacional (Alexandre et al., 2021; Alencar, 2021).

A aplicação de um modelo universal a línguas tipologicamente distintas suscita, contudo, desafios de representação. Sandalo e Galves (2017) denominam esse cenário de Problema de Anchieta, referindo-se às dificuldades de descrever línguas indígenas americanas por meio de categorias gramaticais originalmente formuladas para línguas indo-europeias. No caso do Kubeo, essas dificuldades manifestam-se especialmente na

segmentação de morfemas multifuncionais e na representação de informações morfológicas e discursivas que nem sempre encontram correspondência direta nas categorias universais do UD.

Diante dessas limitações, este trabalho adota a camada de traços morfológicos FEATS prevista na própria especificação do UD para representar informações gramaticais que não podem ser expressas apenas pelas categorias lexicais e pelas relações de dependência. Essa escolha metodológica não altera a arquitetura do modelo, mas explora um de seus mecanismos de extensão para preservar distinções morfossintáticas relevantes da língua Kubeo durante o processo de anotação.

3. Metodologia

O corpus inicial é composto por um conjunto de 20 sentenças extraídas da cartilha *Pamiene Mama Buei-Tukubo* (2023). O material foi selecionado por apresentar construções representativas da língua e por possuir revisão linguística prévia, constituindo uma base adequada para o desenvolvimento do protocolo inicial de anotação. As sentenças foram selecionadas de forma gradual, seguindo um critério de complexidade sintática crescente. Inicialmente, priorizaram-se construções mono-oracionais, expandindo-se posteriormente para estruturas mais complexas à medida que o protocolo de anotação era desenvolvido e as categorias morfossintáticas eram refinadas para a representação da língua Kubeo. Esse processo foi conduzido de forma iterativa, com validações globais e revisões sucessivas das anotações, a fim de assegurar a consistência dos critérios adotados ao longo da construção do corpus.

A anotação foi realizada manualmente no formato *CoNLL-U*, seguindo as diretrizes do UD. Para cada sentença, foram realizadas a segmentação e a lematização, bem como a anotação das categorias morfossintáticas (UPOS), dos traços morfológicos FEATS, das relações sintáticas e da árvore de dependências. Nos casos em que a morfologia do Kubeo não encontrava correspondência direta nas categorias universais, adotaram-se os mecanismos de representação previstos pelo próprio UD, preservando informações relevantes por meio dos traços morfológicos. O protocolo foi refinado iterativamente durante a anotação, de modo a garantir coerência na representação de fenômenos morfossintáticos recorrentes do Kubeo.

Como estudo de caso para avaliação do corpus piloto, analisou-se a distância linear entre sujeitos nominais explícitos e seus respectivos verbos núcleo. Essa métrica foi selecionada por refletir diretamente a estrutura sintática anotada e por permitir verificar a uniformidade das relações de dependência produzidas durante o processo de anotação, além de fornecer um primeiro diagnóstico quantitativo da organização sintática do corpus.

4. Resultados

Como resultado desta etapa, foi produzido um corpus piloto anotado segundo o padrão UD, contemplando segmentação, lematização, categorias morfossintáticas, traços morfológicos e relações de dependência. O processo de anotação permitiu consolidar um protocolo consistente para a representação de fenômenos recorrentes da língua Kubeo, estabelecendo uma base metodológica para a expansão futura do *treebank* e para o desenvolvimento de recursos computacionais voltados ao Processamento de Linguagem Natural.

Como estudo de caso estrutural, analisou-se a distância linear entre sujeitos nominais explícitos e seus respectivos verbos núcleo. O processamento das 20 sentenças do corpus piloto demonstrou a presença do sujeito nominal (nsubj) em 90% da amostra, com forte predominância da ordem Sujeito-Verbo⁵⁸⁵. A distância linear variou de 1 a 4 tokens, com média de 2,5. O distanciamento máximo atingido é advindo da presença de constituintes intervenientes, como pode ser observado na Tabela 1, além de registrar

flexibilidade sintática nos casos em que o verbo precede o sujeito. Esses resultados evidenciam que o protocolo adotado foi capaz de representar adequadamente tanto dependências locais adjacentes quanto dependências de maior extensão.

Tabela 1. Anotação morfossintática no formato CoNLL-U para a sentença 1: "Yawi oko hārāwui āino wore kuibi" (A onça procura comida no dia de chuva).

ID	Forma	Morfemas	UPOS	Feats	DepRel	Head
1	Yawi	yawi (onça)	NOUN	Animacy=AnimlGender=MascINumber=Sing	nsubj	5
2	oko	oko (água/chuva)	NOUN	Animacy=InanINumber=Sing	nmod	3
3	hārāwui	hārāwui (dia) -i (LOC)	NOUN	Animacy=InanICase=Loc	obl	5
4	āino	āino (comida)	NOUN	Animacy=InanICase=AccINumber=Sing	obj	5
5	wore	wo (procurar) -re	VERB	VerbForm=Vnoun	root	0
6	kuibi	kui (estar/AUX) -bi (3M)	AUX	Gender=MascINumber=SingIPerson=3	aux	5

Durante a anotação, verificou-se que determinados fenômenos morfossintáticos do Kubeo demandaram decisões interpretativas que extrapolam uma aplicação estritamente mecânica das diretrizes do UD. A utilização intensiva da camada FEATS permitiu preservar distinções linguísticas relevantes sem comprometer a compatibilidade do corpus com o padrão internacional, evidenciando que, embora o UD forneça uma estrutura robusta para a anotação, a aplicação a línguas tipologicamente distintas daquelas que fundamentaram seu desenvolvimento requer adaptações metodológicas bem documentadas.

Em conjunto, os resultados demonstram a viabilidade da construção incremental de um treebank para a língua Kubeo compatível com UD. Além da produção de um corpus piloto anotado, o trabalho estabeleceu um protocolo reprodutível que poderá orientar a expansão do corpus e subsidiar o desenvolvimento de ferramentas computacionais para a documentação e o processamento automático da língua.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou a primeira etapa da construção de um treebank para a língua Kubeo compatível com o padrão UD, estabelecendo um protocolo de anotação morfossintática desenvolvido a partir de um corpus piloto. Os resultados demonstraram a viabilidade da metodologia proposta e evidenciaram que a utilização dos mecanismos previstos pelo próprio UD, especialmente os traços morfológicos FEATS, permite representar fenômenos relevantes da língua Kubeo preservando a compatibilidade com o padrão internacional.

Como continuidade da pesquisa, pretende-se ampliar o corpus anotado com foco em materiais didáticos e contextos de ensino, incorporando as demais sentenças da cartilha *Pamiene Mama Buei-Tukubo* (2023) e outras fontes desse domínio, a fim de consolidar o treebank da língua Kubeo e disponibilizá-lo como recurso aberto para a comunidade científica. Esse recurso servirá de base para o desenvolvimento de ferramentas de Processamento de Linguagem Natural, como etiquetadores morfossintáticos e analisadores sintáticos, que, após validação supervisionada, poderão apoiar a expansão semiautomática do corpus, tornando-o mais robusto e abrangente. Espera-se que esses produtos contribuam para a documentação digital, a descrição linguística e a preservação da língua Kubeo, ampliando sua representação no ecossistema de recursos computacionais.

Este estudo piloto, de natureza estritamente documental baseada em cartilhas, apresenta limitações de generalização para fenômenos complexos, que serão tratados futuramente, e não envolve intervenção com seres humanos. Declara-se o uso de IA Generativa exclusivamente para revisão gramatical e aprimoramento estilístico do texto final, conforme diretrizes da SBC.

Referências

- ALENCAR, L. F. de. Uma gramática computacional de um fragmento do Nheengatu. *Revista Estudos da Linguagem*, Belo Horizonte, v. 29, n. 3, p. 1717–1777, 2021.
- ALEXANDRE, D. M.; GURGEL, J. L.; ALENCAR, L. F. de. Nheentiquetador: um etiquetador morfossintático para o sintagma nominal do Nheengatu. In: *Anais do XIII Encontro de Linguística de Corpus*. 2021.
- CASELI, H. de M.; NUNES, M. das G. V. (eds.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. São Paulo: SBC, 2022.
- CHACON, T. C. The phonology and morphology of Kubeo: the documentation, theory, and description of an Amazonian language. Tese (Doutorado em Linguística) - University of Hawai'i at Manoa, Honolulu, 2012.
- CHACON, T. C. (Org.). *Pamiene Mama Buei-Tukubo: Primeiro livro para o ensino da língua Kubeo*. Escola Estadual Indígena Huredanaçu Nossa Senhora da Imaculada Conceição; SEDUC-AM, 2023.
- DE MARNEFFE, M.-C.; et al. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255–308, 2021.
- NIVRE, J.; et al. Universal Dependencies v1: A multilingual treebank collection. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association, 2016.
- SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do Brasil: o problema de Anchieta. *Cadernos de Estudos Linguísticos*, Campinas, v. 59, n. 3, p. 631–648, 2017.