DOI: 10.20396/cel.v65i00.8673592

ANOTANDO SINTATICAMENTE UMA LÍNGUA ORIGINÁRIA DO BRASIL: O PROBLEMA DE ANCHIETA

SYNTACTICALLY ANNOTATING A LANGUAGE ORIGINATING IN BRAZIL: THE PROBLEM OF ANCHIETA

Maria Filomena Spatti Sandalo[1]
Charlotte Marie Chambelland Galves[2]

Resumo: Neste artigo, apresentamos uma proposta de anotação categorial, morfológica e sintática de uma língua da família Guaikurú, kadiwéu, falada no centro-oeste brasileiro. A anotação gramatical do kadiwéu nos coloca diante da questão da universalidade das categorias linguísticas. O que chamamos de problema de Anchieta é a questão de saber até que ponto é possível encaixar línguas originárias dentro dos esquemas de anotação que funcionam para línguas indo-europeias. Trata-se de uma língua com menos de 500 falantes e sua documentação e fortalecimento são urgentes.
Palavras chave: kadiwéu, família Guaikuru, anotação sintática, corpora anotados

Abstract: In this article we present a proposal for categorical, morphological and syntactic annotation of a language of the Guaikurú family, Kadiwéu, spoken in the Brazilian Midwest. The grammatical annotation of Kadiwéu confronts us with the question of the universality of linguistic categories. What we call the Anchieta problem is the question of how far it is possible to fit indigenous languages within the annotation schemes that work for Indo-European languages. It is a language with less than 500 speakers and its documentation and strengthening are urgent.
Keywords: kadiwéu, Guaikuru family, syntactic annotation, annotated corpora

1. INTRODUÇÃO

As línguas do Brasil, já gravemente ameaçadas antes da pandemia, contaram com a morte de idosos por COVID19, enfrentando, assim, mais uma etapa de enfraquecimento. Este trabalho apresenta nossos esforços na busca de inovação digital inclusiva desenvolvendo uma Plataforma computacional para línguas originárias do Brasil com materiais que são culturalmente e gramaticalmente significativos para as comunidades indígenas. Esses materiais fornecem uma base digital para a educação bilingue aprimorada, conexões intergeracionais, e transmissão de saberes ancestrais. Assim, temos como objetivo poder colaborar com a documentação e preservação nas escolas de línguas nativas do Brasil através de nossa ferramenta para elaboração de corpora linguísticos. Os corpora nessa Plataforma têm ainda anotações automáticas

[1] Professora titular da Universidade Estadual de Campinas Campinas, Unicamp, Campinas, SP, Brasil. sandalo@unicamp.br
Orcid: https://orcid.org/0000-0003-4595-7765
[2] Professora titular da Universidade Estadual de Campinas, Unicamp, Campinas, SP, Brasil. galvesc@unicamp.br
Orcid: https://orcid.org/0000-0002-5326-1568

sintáticas e morfológicas, bem como tradução palavra por palavra e por frase para o português e o inglês. As anotações são feitas de modo a já formar uma gramática das línguas cujos corpora são depositados na Plataforma. Estas gramáticas também servirão como ferramentas de ensino para melhorar o letramento dos falantes locais em suas próprias línguas, bem como a preservação de narrativas culturalmente significativas.

Além disso, um melhor estudo das correspondências, nem sempre óbvias das línguas indígenas e europeias, também melhorará a formação de professores de línguas indígenas para uma educação bilíngue mais eficiente e equilibrada, com resultados positivos para o acesso ao ensino superior. E de um ponto de vista linguístico permitirá novas pesquisas sobre a gramática destas línguas em processo de enfraquecimento e de extinção.

A *Plataforma Tycho Brahe* (https://www.tycho.iel.unicamp.br/home/), desenvolvida na UNICAMP,[3] é um sistema computacional baseado em navegador da web com ferramentas de buscas, visualização e edição de corpora linguísticos, integrado a um mecanismo de etiquetagem e parseamento morfossintáticos que fornecem maneira de analisar narrativas desde o nível da frase até a estrutura interna da palavra.[4]

A primeira língua indígena na *Plataforma Tycho Brahe* é o kadiwéu, uma língua da família Guaikurú falada no Mato Grosso do Sul, cujos falantes atualmente não passam de 500 pessoas. O material já anotado do kadiwéu mostra nosso engajamento e força para alcançar corpora digitais maiores de línguas que estão rapidamente enfraquecendo em seus usos e transmissão para as próximas gerações.

É importante ressaltar que este trabalho se baseia em uma concepção/filosofia de construção de corpus diferente do *mainstream* de abordagens computacionais que partem de uma enorme quantidade de dados (*big data*), como perfeitamente exposto no seguinte trecho de um artigo intitulado "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", publicado em 2008 na revista *Wired Magazine*:

> "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology.
> The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world."[5]
> (Chris Anderson, Wired Magazine 16:7, 2008)

Assumimos uma posição diametralmente oposta que contesta a obrigatoriedade de 'big data' para constituir corpora de línguas utilizáveis na pesquisa. Para línguas em perigo de extinção, em particular, grandes corpora não podem ser constituídos, e isso por si só as excluiria desse tipo de projetos. Defendemos que a anotação acrescida aos textos é que os torna relevantes para pesquisas e aplicações para fins educativos e sociais. Afirmamos que grandes corpora são inúteis se não contiverem informações extras que permitam aos pesquisadores recuperar dados para responder às suas perguntas sobre a linguagem e sua dinâmica. Essas informações extras são adicionadas por meio de

[3] A Plataforma Tycho Brahe foi idealizada e tem sido desenvolvida pelo cientista da computação e linguista Luiz Veronesi no âmbito dos projeto temáticos FAPESP 2012/06078-9 e 2022/09158-5, e do edital universal CNPq 436209/2018-7.

[4] A Plataforma conquistou o segundo lugar no prêmio ABRALIN em Tecnologia e Inovação em Pesquisas Linguísticas em 2021.

[5] Este é um mundo onde grandes quantidades de dados e matemática aplicada substituem todos os outros instrumentos que podem ser usados. Fora com todas as teorias sobre o comportamento humano, da linguística à sociologia. A nova disponibilidade de grandes quantidades de dados, juntamente com as ferramentas estatísticas para processar esses números, oferece uma nova maneira de entender o mundo (tradução dos autores).

anotação e a anotação é baseada em modelos linguísticos. O que precisamos é desenvolver métodos que permitam a anotação de textos de forma rápida e confiável. Isso implica interdisciplinaridade, em particular, cooperação com cientistas da computação.

A *Plataforma Tycho Brahe* junta e complementa esforços como o *ANNIS* (corpus-tools.org/annis), desenvolvido na Universidade Humboldt, em Berlim, e aplicado a uma variedade de idiomas de alto recurso, como alemão, árabe e outros. O *ANNIS* também é uma arquitetura de pesquisa e visualização baseada em navegador para corpora linguísticos multicamadas com diversos tipos de anotação. Uma vantagem da nossa plataforma são suas ferramentas integradas de etiquetagem e anotação de vários níveis atuando em nível de morfemas, palavras e sintagmas, que a tornam particularmente adequada para lidar com a estrutura morfológica altamente complexa das línguas aglutinantes da América do Sul.

A *Plataforma Tycho Brahe* é pioneira em sua aplicação para as línguas originárias da América do Sul, e sua interface baseada na web permite a disseminação imediata para as comunidades indígenas, além de permitir a criação de corpora significativamente maiores do que era possível anteriormente.

A anotação morfossintática do kadiwéu, língua com pouco tradição gramatical, nos confronta à questão da universalidade das categorias linguísticas, tanto a nível das palavras quanto a nível das orações. O que chamamos de problema de Anchieta é a questão de saber até que ponto é possível encaixar línguas desse tipo dentro dos esquemas de anotação funcionando para línguas indo-europeias como o português e o inglês. Anchieta, bem como os jesuítas que escreviam gramáticas para as línguas do novo mundo, usavam, por exemplo, o sistema de casos do latim, ou os conceitos de tempos, modos, e aspectos de línguas como o português para organizar os paradigmas nominais e verbais encontrados nelas. Da mesma maneira, procuramos usar noções como orações completivas, relativas, clivadas, relevantes para o português e o inglês, para expressar distinções gramaticais entre tipos de orações do kadiwéu. Em certos casos, essas categorias se mostram eficientes na descrição, em outros, elas se revelam de pouca serventia para estabelecer as distinções necessárias entre diferentes construções. Enquanto gerativistas, acreditamos, contudo, que as línguas são todas o produto da mesma máquina cognitiva, a faculdade de linguagem, e que, no nível sintático, as diferenças não devem impedir o uso de categorias de descrição as mais semelhantes, ou comparáveis, possível. No sistema de anotação proposto neste artigo, tentamos assim um equilíbrio entre os dois polos opostos da semelhança e da diferença.

Partindo dessas considerações, o artigo está organizado da seguinte maneira. Na seção 2, apresentamos nossa proposta de anotação para o kadiwéu, começando pela anotação de palavras (2.1), e propondo em seguida um nível suplementar de anotação morfêmica devido à natureza polissintética da língua, que a diferencia fortemente de línguas como o português e o inglês (2.2). Este nível é o lugar em que a especificidade do kadiwéu está maximamente representada, acarretando etiquetas totalmente diferentes daquelas usadas até agora em corpora sintaticamente anotados no modelo dos *Penn Parsed Corpora of Historical English* (Kroch et al. 2000; Kroch et al. 2004; Kroch, et al. 2016) como o *Corpus anotado do português histórico Tycho Brahe*. Finalmente, em 2.3, propomos uma primeira versão do sistema de anotação sintática, que se inspira fortemente desse modelo. A seção 3 apresenta casos que desafiam a mera transposição de categorias emprestadas de línguas europeias, numa abordagem "anchietista", e sugere modificações no sentido de uma apreensão mais aproximada do funcionamento da língua, mesmo que isso nos obrigue a nos atermos, pelo menos num primeiro passo, a uma anotação mais genérica, dado o caráter ainda parcial da nossa compreensão dos detalhes da gramática do kadiwéu.

2. O SISTEMA DE ANOTAÇÃO

2.1 A anotação categorial (POS) e morfológica

O sistema de marcação dos PPCHE (cf. Santorini 2020) já havia sido expandido para acomodar a rica morfologia flexional do português (cf. Galves et al. 2017). No sistema do português, cada palavra é marcada com sua categoria principal e, opcionalmente, com uma ou mais etiquetas (*tags)* secundárias codificando propriedades morfológicas (ver Britto et al. 2002). Este sistema não foi suficientemente rico, entretanto, para a morfologia do kadiwéu, exigindo mais camadas de rotulagem. Portanto, adicionamos uma camada puramente de rótulos morfológicos. E o sistema de etiquetagem foi ampliado para permitir uma maior variedade de *tags* de palavras e morfemas.

O processo de marcação que está sendo testado atualmente consiste em dois níveis: primeiro, o etiquetador (*tagger*) é executado no nível da frase, atribuindo uma etiqueta POS (Part of Speech) a cada palavra. Segundo, o processo é executado dentro de cada palavra, atribuindo etiquetas morfológicas aos constituintes internos de palavras. Na Figura 1 é possível observar os dois níveis de anotações:[6]

[6] As abreviações usadas neste trabalho são:
POS: VB = verbo, VBAPL = verbo marcado por um aplicativo, N = nome, NAPL = nome marcado por um aplicativo, N$ = nome possuído, D = categoria funcional modificadora de um nome, DNUM = classificador numeral e numeral, Q = quantificador, WPRO = partícula interrogativa de argumentos, WADV = partícula interrogativa de adjuntos, PRO = pronome pessoal, PRO$ = pronome possessivo, EV = evidencial, T = tempo, C = complementador, CT = complementador marcado por tempo, CNeg = complementador negativo, c = núcleo do complementador, NPR = nome próprio, NEG = partícula negativa, ADVAPL = adjetivo marcado por um aplicativo

Rótulos Morfológicos: Plu 'plural', Ant 'antipassivo', Imp 'impessoal', Erg 'ergativo', Abs 'absolutivo', Inv 'inverso', Hit 'hither', v 'raiz verbal', n 'raiz nominal', Gen 'genitivo', Cla 'classificador nominal', Ncl = classificador numeral, Der 'morfema derivacional', Dim 'diminutivo', ADJ = adjetivo, Anf = anafórico, Gnr = gender, Num = núcleo do numeral, 1sg = primeira pessoa singular, 2sg = segunda pessoa singular, 3sg = terceira pessoa singular, 1pl = primeira pessoa plural, 2pl = segunda pessoa plural, 3pl = terceira pessoa plural, Prosp = futuro prospectivo, Compl = aspecto completivo.

Rótulos Frasais/Sintáticos: IP-MAT = sentença matriz, NP-SBJ = sintagma nominal sujeito, NP-OB1 = sintagma nominal objeto direto, NP-OB2 = sintagma nominal argumento de verbos seriais, NP-APL = sintagma nominal objeto de verbos aplicativos, NP-LOC = sintagmas nominais locativos, NP-ADV = sintagmas nominais adverbiais, NP-LFD = sintagmas nominais deslocados, NP-PRD = sintagmas nominais predicativos, IP-SUB = oração subordinada, CP-THT = oração completiva com complementador, CP-REL = oração relativa, ADVP = sintagma adjetival, CP-QUE = oração interrogativa, IP-ADV = oração adjunta adverbial, IP-IMP= oração imperativa, CP-ADV = oração adjunta adverbial com complementador, CP-ME = oração cujo núcleo é a partícula *me*.

Rótulos para categorias vazias: *pro* = sujeitos e objetos nulos interpretáveis como pronomes lexicais, *T* = vestígios de movimento-QU, nas orações relativas e interrogativas, *ICH* = categorias vazias associadas à topicalização de sintagmas, WNP * = operadores nulos, quando somente o complementador é realizado lexicalmente em construções QU, T* = para Tempo não realizado lexicalmente, EXT* = existencial vazio, VB* = verbo leve vazio.

Figura 1: Níveis de Anotação do Corpus kadiwéu na Plataforma Tycho Brahe (https://www.tycho.iel.unicamp.br/browser/catalog/C12)

atone yoe lotiidi ica iwaalo

original	atone		yoe		lotiidi		ica		iwaalo
POS tag	CNeg+EV		VB		N$		D		N
gloss-br									
gloss	se diz que não		faz		leite		a		mulher
morphemes	ade	one	y	oe	l	otiidi	i	ca	
tag	c	ev	Erg	v	Gen	n	Gnr	Ncl	
gloss-br									
gloss									

'Se a mulher não fizer leite (materno)

Abaixo apresentamos a etiquetagem categorial usada na plataforma, ou seja, as categorias do primeiro nível. Esta anotação é crucial para a elaboração de regras sintáticas que anotam automaticamente os dados em uma análise sintática. [7]

2.1.1 A Anotação Categorial

A anotação de rótulos categoriais (POS),[8] a primeira camada da anotação, é uma adaptação do sistema usado em português e inglês, uma vez que as categorias sintáticas são, em sua maioria, universais. Há, entretanto, exceções. Por exemplo, o kadiwéu não tem adposições e os argumentos internos indiretos são licenciados por aplicativos conforme a seguinte estrutura sintática:[9]

[7] Sobre a marcação morfológica, a segunda camada de etiquetagem, a tradição da linguística indígena, no Brasil e fora do Brasil, usa o sistema de anotação de Leipzig. O sistema de anotação de Leipzig foi desenvolvido pelo Departamento de Linguística do Instituto Max Planck de Antropologia Evolutiva (Bernard Comrie, Martin Haspelmath) e pelo Departamento de Linguística da Universidade de Leipzig (Balthasar Bickel). Usaremos, entretanto, nossa própria anotação, que não difere radicalmente daquela de Leipzig.

[8] Entendemos por categorias POS os nós terminais de árvores sintáticas. É, portanto, uma noção só em parte equivalente a classes de palavras. Na anotação simplificada que adotamos, porém, não anotamos certos elementos funcionais como nós independentes quando eles nunca aparecem como morfemas livres. Por exemplo, é fato que aplicativos são núcleos sintáticos independentes na teoria (cf. Pylkkanen 2002), mas como nunca aparecem como morfemas livres no kadiwéu, anotamos de modo simplificado VBAPL, DAPL e NAPL. Por outro lado, a nível POS, anotamos os classificadores como D, um núcleo funcional modificador do nome na estrutura arbórea, deixando o rótulo CL (classificador) para o nível morfológico. Para defesa de que classificadores ocupam um núcleo sintático, veja, por exemplo, Rothstein (2011).

[9] Abaixo seguem exemplos de núcleos funcionais não comuns entre línguas europeias:

Aplicativos:

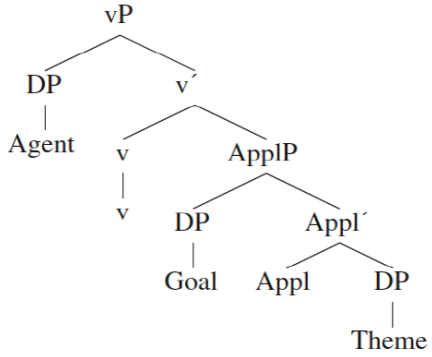
CHAGA:

N-óa-óý-lyì-í-à m- kà k-élyá
FOCO-1SG-T-comer-APPL-FV 1-esposa 3-comida
'Ele está comendo comida por sua esposa'
(Bresnan and Moshi 1993: 49-50 via Pylkkanen 2002)

CHICHEWA:

Mavuto a-na-umb-ir-a mpeni mtsuko
Mavuto SP-PAST-moldar-APPL-ASP faca pote de água

Figura 2: Estrutura sintática de argumentos internos no kadiwéu (Nevins & Sandalo 2011)



Há dois tipos de morfemas aplicativos (que consideramos como o núcleo de um ApplP): aqueles que coocorrem com concordância e aqueles que não coocorrem com concordância. Os primeiros licenciam argumentos indiretos, segundo a estrutura acima, e os segundos sintagmas nominais adjuntos. Neste caso, a literatura das línguas Guaikurú tem rotulado tais morfemas como direcionais. Embora na primeira camada rotulamos todos como APL, na segunda diferenciamos os segundos como direcionais (Dir). Os aplicativos dos dois tipos são sempre morfemas inseparáveis e ocorrem mais frequentemente com o verbo, portanto aparecem na camada POS como VBAPL. Mas podem estar ligados a nomes e classificadores numerais quando o verbo é copular (fonologicamente nulo), ou ainda, em adjuntos nominais, e serão rotulados de NAPL e DAPL.[10]

(1) naigitece idalaGata emokaye
NAPL VBAPL N
Pelo-caminho falo-sobre bocaiuva
'Pelo caminho, falo sobre bocaiuva'

Tempo e aspecto são categorias funcionais independentes em orações matrizes e são rotulados como T:

'Mavuto moldou o pote de água com uma faca'
(Baker 1988: 354 via Pylkkanen 2002)

Evidenciais

JAPONÊS:

Kazuko wa kinoo Tokyo e ikimashita yo
Kazuko Top ontem Toquio para foi EV
'Ontem Kazuko foi para Tokyo (estou dizendo para você)
(Tenny 2006)

Classificadores numerais

MANDARIM

Liǎng zhī bǐ
2 CL caneta
'Duas canetas'
(Doetjes 2017)

[10] Sobre a estrutura morfológica do kadiwéu, veja mais adiante (seção 2.1.2).

(2) ja diote
T VB
Compl dormiu
'ele dormiu'

(3) domaGa joletibige ica me ele
T VB D C ADJ
Prosp procurarei o que bom
'Eu vou procurar o que é bom'

O kadiwéu tem dois marcadores aspectuais: *jaG* 'perfectivo' (Perf), e *baanaGa* 'imperfectivo' (Imperf), e uma marca de tempo futuro, *domaGa* 'futuro prospectivo' (Prosp). Embora sejam palavras independentes, em fala rápida podem se cliticizar em categorias lexicais adjacentes e, neste caso, marcamos a juntura com o sinal +, usado na anotação do português para os pronomes e determinantes cliticizados a verbos e preposições.[11]

(4) jama
T+VB
'Acabou'

(5) jonaGa yeleo
T+EV VB
'Diz que já morreu'

Nas orações subordinadas, as marcas de tempo aparecem junto ao complementador obrigatoriamente. Em orações subordinadas adverbiais, Tempo é fundido ao complementador, isto é, não é possível separar o complementador e Tempo e, neste caso, marcamos como CT. O complementador não marcado por tempo é rotulado C. Os exemplos abaixo são de Sandalo (2023).

(6) alawini naGa dopitedice
VB CT VBAPL
Preste-atenção que (passado) voltou
'Preste atenção que ele retornou'

(7) jigaalatece nigaanigipi nige jiwidatiogi
VBAPL N CT VBAPL
Sigo crianças que (futuro) alcançarei-as
'Eu sigo as crianças para alcança-las'

Todas as orações adverbiais são introduzidas pelos CTs *nige* and *naGa*. Os complementadores *nige* e *naGa* são marcados por tempo, como já mencionado: *naGa* é passado e *nige* é futuro, como pode ser destacado no par mínimo abaixo de Sandalo (1997):

[11] Sobre evidencialidade (EV) veja mais adiante para uma discussão mais detalhada.

(8) Pedro yatemati Ecode naGa yoe diimigi
NPR VBAPL NPR CT VB N
Pedro conta-para Ecode quando-que construiu casa
‘Pedro contou para Ecode quando ele construiu as/a casa/s’

(9) Pedro yatemati Ecode nige yoe diimigi
NPR VBAPL NPR CT VB N
Pedri contou-para Ecode quando-que construirá casa
‘Pedro contou para Ecode quando ele vai construir as/a casa/s’

As orações complementos são marcadas pelo complementador *me*:

(10) Ana yemaa me Maria dabaqenaGa
NPR VB C NPR VB
Ana quer que Maria lave-roupa
'Ana quer que Maria lave roupa'

O complementador *me* é também usado em orações controladas (Sandalo 1997):

(11) oyatita napalite me oilojoGo
VBAPL N$ C VB
Usaram machado para amassá-la/lo
‘Eles usaram um martelo para amassá-la/lo’

(12) Pedro eeta Ecode me dinoojete domoojia
NPR VBAPL NPR CVB N
Pedro disse-para Ecode para comprar carro
‘Pedro disse para Ecode para comprar um carro’

Entre as categorias funcionais, o kadiwéu tem dois tipos de marcas de negação. A primeira é a*G-*, que é um clítico que marca negação na oração matriz. A segunda é a palavra funcional *daGa*, que segue um complementador e tem escopo na oração subordinada (Sandalo 1997):

(13) Pedro ayemaa me dawii
NPR NEG+VB C VB
Pedro não-quer que cace
‘Pedro não quer que ele cace’’

(14) Pedro eeta Ecode me daGa dinoojeteta domoojia
NPR VBAPL NPR C NEG VBAPL N
Pedro disse-para Ecode que não compre-para ele carro
‘Pedro disse para Ecode não comprar um carro para ele’

Para negar a matriz e a subordinada, ambas marcas de negação precisam ser usadas (Sandalo 1997):

(15) Pedro aGeeta Ecode me daGa dinojeteta domoojia
NPR NEG+VBAPL NPR C NEG VBAPL N
Pedro não-disse-para Ecode que não compre-para elecarro
'Pedro não disse para Ecode não comprar um carro para ele'

O kadiwéu, entre suas categorias funcionais, também tem uma marca de evidencialidade reportativa, anotada EV. O evidencial reportativo *one* expressa que o que o falante está trazendo é uma informação adquirida oralmente (Sandalo 2023).

(16) one doita ica ejiwajegi daGa dilaike
EV VBAPL D N C VB
Diz-que teme o kadiwéu que (condicional) grisalhar
'Diz que o povo Kadiwéu tem medo de ter cabelo branco'

O evidencial pode se fundir com as categorias de aspecto e tempo *ja* e *naGa*.

(17) jona leegitece noGone yotaGaneGe
T+EV ADVAPL CT+EV VB
Completivo+diz-que longe-de quando-que+diz que falou
'Diz que ela estava longe quando falou'

O kadiwéu tem ainda um morfema que introduz orações condicionais negativas e outro que introduz sentenças condicionais não negativas, *ade* e *daGa* respectivamente. Como introduzem sentenças subordinadas, eles também são rotulados categoricamente como C (Sandalo 2023).

(18) Goniotagodi GodacawaneGegi ade oko jaaGa
N$ N$ C PRO VB
Senhor nos-ajude se não nós morremos
'Senhor nos ajude, se não nós morreremos'

Quanto ao sintagma nominal, o kadiwéu é uma língua de classificadores numerais (Sandalo & Michelioudakis 2016). Marcamos os classificadores como D na camada categorial, mas a camada morfológica deixa claro de que se trata de um classificador numeral e não de artigos. Sandalo & Michelioudakis (2016) mostram que os nomes nus no kadiwéu não são necessariamente entendidos como singulares e que a língua conta com classificadores para individualizar os nomes, como no chinês (Chierchia 1998). Não há diferença entre definido e indefinido em kadiwéu, o classificador apenas garante individualização. E, como em outras línguas da tipologia de classificadores numerais, um numeral não pode aparecer desacompanhado de um classificador em kadiwéu, sejam os nomes nucleares do sintagma nominal massivos ou não massivos.

Segundo Sandalo & Michelioudakis (2016), há, entretanto, uma diferença entre o kadiwéu e línguas mais conhecidas de classificadores numerais: o fato de que os classificadores são incorporados ao numeral kadiwéu dada a natureza polissintética da língua. Abaixo (exemplo 19), segue um dado com um numeral. O numeral 2 não pode aparecer como uma palavra independente: o numeral não tem acento primário próprio e não pode ser separado do classificador. E a

sentença é agramatical sem o classificador incorporado ao número 2, *ta:le*, como mostra (19). A anotação nesse caso é DNUM.[12]

(19)

João	y-a:	i-n:i-wa-ta:le	apolikaGanaGadi
NPR	VB	DNUM	N
João	compra	dois	cavalos

'João compra dois cavalos

Os pronomes pessoais são anotados PRO, como em português e inglês. Há também um pronome relativo *ane*, anotado como WPRO. Palavras interrogativas são anotadas como WADV, quando remetem a um adjunto, e WPRO+D quando significam 'o que' (cf. Seção 3, ex. 43, 44, 45, respectivamente). Os pronomes possessivos, como em português e em inglês, são anotados PRO$:

(20)

inebi	Ganebi	nebi	Gonebi
PRO$	PRO$	PRO$	PRO$
'meu'	'teu'	'seu'	'nosso'

Como já mencionamos, os verbos e verbos aplicativos são marcados como VB e VBAPL respectivamente.

Os nomes têm dois tipos de marcação, N e N$. Isso ajuda a deixar claro a estrutura de sintagma nominal genitivo do kadiwéu que é recursiva como do inglês. A diferença é de que, em kadiwéu, o possuído é marcado com o morfema genitivo e há concordância com pessoa. N marca nomes não possuíveis, e, portanto, que não estão em uma estrutura nominal genitiva. Todos os nomes inalienáveis são marcados por morfemas genitivos e são, portanto, N$.

Os adjetivos (ADJ) são raros, e, o que é adjetivo em português, é frequentemente ou um verbo ou um nome em uma estrutura de posse (estrutura genitiva). Os exemplos abaixo ilustram respectivamente os três casos.

(21) Adjetivo

jiGini	ane	ele	lowoodi
T+D	WPRO	ADJ	N$
Compl+Ncl	que	boa	roupa-dele

'Aquele cujas roupas são boas'

(22) Verbo

ajo	liwatece	ja	iwagadi
D	N$	T	VB
A	canoa-dele		Compl estar-pesada

'A canoa dele estava pesada'

[12] Os numerais tradicionais do kadiwéu, isto é, de origem Guaikurú, são aqueles de 1 a 3 e seus derivados apenas (veja Griffiths 1975 para uma descrição dos numerais em kadiwéu). A língua emprestou numerais do português para preencher as lacunas dos numerais não existentes ao compararmos kadiwéu com o português. E isso tem um impacto em sua gramática. Uma diferença entre os numerais nativos e os emprestados é que os classificadores não são jamais incorporados aos numerais emprestados do português, ao contrário do que acontece com os numerais nativos. Neste caso, os classificadores aparecem como palavras independentes e são opcionais em leituras de contagem.

(23) Nome

ica	liwigo	libinienigi
D	N$	N$
A	comida-dele	beleza-dela

‘A sua comida dele/dela é gostosa ‘ (lit.: A gostosura de comida dele/dela.’)

Nomes e verbos são reconhecidos pelas suas diferentes morfologias. Verbos e nomes em estruturas genitivas são marcados por diferentes prefixos de pessoa. Adjetivos são monomorfêmicos e, se derivados, um sufixo adjetivador aparecerá.

Há ainda advérbios:

(24)

Oda	aGaGa	yakadi
ADV	ADV	VB

‘Então também pode’

E quantificadores (Q):

(25)

eliodi	oko	oyowoGodi
Q	N	VB
Muita/muito	gente	sabem

‘Muita gente sabe.’

Há ainda uma estrutura em kadiwéu usada em formação de grupo, e que contém dois classificadores numerais (D). Anotamos o primeiro como Q. A sentença abaixo significa que entre as pessoas da aldeia, havia um certo casal.

(26)

one	idi	ica	wadonadi
EV	Q	D	VB

‘Diz que havia um casal

Sandalo & Michelioudakis (2016) observam também que dois Ds (Q e D) aparecem no sintagma nominal quando há leitura de medida. Segundo os autores, em (27), “vemos que há um primeiro iniciando o sintagma nominal e outro incorporado ao numeral: [**id:i**wa [[**ito**wata:le [waka loti:di]]]]. O primeiro atomiza a frase [itowata:le waka loti:di] ‘dois leites de vaca’ servindo como uma unidade de medida. E o outro atomiza [waka loti:di] ‘leite de vaca’, já que se trata de uma língua de classificador numeral”:

(27)

id:iwa	in:iwata:le	waka	l-otidi
Q	D-NUM	N	N$

‘Há dois litros de leite de vaca (no bolo).

2.1.2 A anotação morfológica

O verbo é sempre marcado por pessoa em qualquer estrutura sintática. O verbo concorda com o sujeito se for inergativo e transitivo cujo objeto é de terceira pessoa. Mas concorda com o objeto em verbos inacusativos e transitivos cujos objetos são de primeira ou segunda pessoa. Quando os argumentos são de primeira e segunda pessoa, o verbo concorda com a segunda pessoa (hierarquia 2>1>3, veja Nevis & Sandalo 2011 e Sandalo

2023). Neste último caso há ainda um morfema rotulado de inverso. Verbos reflexivos, médios, antipassivos e alguns inacusativos são marcados por uma marca de sujeito intransitivo. Sena (2016) mostra que certos verbos psicológicos são marcados pelo morfema inverso.

Figura 3: Concordância de pessoa (Sandalo 2009)

	SUBJEITO (transitivo)	SUBJEITO (intransitivo)	OBJETO
1sg	j-	i-…	i-
2sg	a-…-i	a-…-i	Ga-
3sg	y- ~ w-	∅ ~ n-	∅
1pl	j-…-Ga	i-…-Ga	Go
2pl	a-…-i	a-…-i	Ga
3pl	o-y-	∅…-Ga	∅

O nome possuído é marcado pelos seguintes morfemas genitivos:

Figura 4: Concordância nominal (possessivos) (Sandalo 1997)

1sg	i-
2sg/pl	Gad:-
3sg/pl	l-
1pl	God:-
indefinido	n-

Recursividade é comum em estruturas genitivas:

(27)	Ganebi	wa:ka	libol:e	libinyenig:i
	PRO$	N	N$	N$
	Gad:-nebi	*wa:ka*	*l-bol:e*	*l-binye-nig:i*
	Gen-pro$	n	Gen-n	Gen-n
	2Gen-prossessivo	vaca	3Gen-leite	3Gen-beleza

‘Sua bela carne de vaca’ (Lit.: A beleza da sua carne de vaca)

No *Corpus kadiwéu*, seguindo Galves et al. (2017), marcamos a morfologia como nas tabelas abaixo. O detalhamento dos morfemas é dado nas glosas como ilustrado no exemplo (28).

Figura 5: Morfologia do kadiwéu (Galves et al. 2017, pp. 636, Table 2)

POS TAGS	Morpheme Tags		Examples	
VB	Plu	plural	o-y-a:lGe Plu-Erg-v	'They kidnap him.
	Imp	impersonal	eti-Ga-d:-d:egi Imp-Abs-Inv-v	'Someone brought you.'
	Erg	ergative agreement	j-awi: Erg-v	'I hunt it.'
	Abs	absolutive agreement	i-d:-abi-d Abs-Inv-v-Asp	'I'm standing up.'
	Inv	inverse voice	Go-d:-ili: Abs-Inv-v	'We grow.'
	Ant	antipassive	n-ema-ta Ant-v-Obl	'She/he loves him/her in distance.'
	Hit	hither	n-ad:e:gi Hit-v	'He brings it.'
	v	verbal root		
	Val	valence change morpheme	j-otaGan-Gen:-aGa Erg-v-Val-Plu	'We talk to him.'
	Asp	aspect	o-y-aqage-di Plu-Erg-v-Asp	'They cut it.'
	Obl	oblique argument agreement	me-ta v-Obl	'He says to him.'
	Dir	directional morpheme	ji-l:o-ko-tigi Erg-v-Val-Dir	'I look up at something.'
	Mot	motion	ji-n-otiqo-tijo Erg-Hit-v-Mot	'I come wistling.'
	Apl	aplicative	j-ao-tGa-domi Erg-v-Obl-Apl	'I make it for you.'

POS TAGS	Morpheme Tags		Examples	
N	Gen	genitive agreement	l-okaGe-te-di Gen-n-Cla-Plu	'his friends'
	Ant	antipassive	n-gato-je Ant-n-Cla	'a bullet'
	n	root	dom:o:jya n	'car'
	Cla	classifier	apaqa-co-di n-Cla-Plu	'rheas'
	Der	derivation	n-dele-Gikajo Ant-v-Der	'warrior'
	Dim	diminutive	l-atope-nig:i Gen-n-Dim	'his gun'
	Plu	number	Gonel:egi-wa-tedi n-Cla-Plu	'groups of man'
D				
	Anf	anaphoric	nG-i-jo nG-Gnr-Ncl	This/the/ An one_ mentioned before
	Gnr	gender	i-di Gen-Ncl	This/the/An one
	Ncl	numeral classifier	i-di Gen-Ncl	This/the/An one
	Plu	number	i-di-wa Gnr-Ncl-Plu	These/the ones
NUM	Num	numeral	i-ni-wa-ta:le Gnr-Ncl-Plu-Num	two
Q				
	Qnt	quantifier	oni-ni-te-k-beke Num-Gnr-Ncl-Obl-Apl-Qnt	each
WPRO	Int	interrogative pro-noun	am-i:-na Int-Gnr-Ncl	'who'
WADV	Whs	Wh-support	ig-ame Whs-Int	'why'
PRO	Pro	pronoun	aqa:m:-i Pro-Plu	'you'

2.2. A anotação sintática

A anotação sintática do kadiwéu ainda está em fase de construção e por isso não está implementada na plataforma. O que apresentamos aqui é o sistema que foi elaborado a partir da anotação manual de 50 frases de um dos textos que compõem o corpus do kadiwéu, *nigedioli* "a mulher onça". Esta anotação, além de constituir a base do manual que será disponibilizado para os usuários do corpus, servirá também de base para o anotador automático (*parser*) baseado em regras, construído no modelo do parser para o português (Magro 2017)[13], que faz uso da função de revisão da linguagem de busca *Corpus Search*,[14] para construir as árvores sintáticas associadas às orações (cf. Faria et al. a sair). A anotação sintática do kadiwéu toma como ponto de partida o sistema de anotação inicialmente elaborado para anotar o inglês histórico, e adaptado mais tarde ao português. (cf. respectivamente https://www.ling.upenn.edu/hist-corpora/annotation/index.html e https://alfclul.clul.ul.pt/portuguesesyntacticannotation). Contudo, como já preliminarmente discutido em Galves et al. (2017, pp. 642-643), modificações envolvendo tanto supressões como acréscimos tiveram que ser efetuadas. Por exemplo, a ausência já mencionada de P na língua prescinde a categoria PP. A nível das orações, uma vez que a finitude não é expressa, as categorias correspondendo a orações infinitivas, gerundivas e participiais também não são necessárias. Dois fatos aliás sugerem fortemente que a oposição finito/infinitivo não existe em kadiwéu (cf. nota 19 para uma problematização). O primeiro é que não há distinção morfológica suportando a oposição finito/não finito. O segundo é que todas as construções de subordinação têm complementador, menos no caso dos verbos seriais. Examinaremos esses casos mais em detalhe na seção 2.2.2. Algumas inovações são também necessárias, como mostraremos no esboço de manual de anotação sintático que apresentamos a seguir. A questão que não pode deixar se ser colocada é se o projeto como um todo não deve esbarrar na grande diferença tipológica entre kadiwéu por um lado e inglês e português por outro lado. Nesta seção, mostraremos que, numa primeira abordagem, pelo contrário, são múltiplas as semelhanças sintáticas, e que os sistemas já elaborados para outras línguas são perfeitamente utilizáveis. Na seção 3, mostraremos os limites dessas semelhanças e as decisões práticas que isso nos levará a adotar para evitar os percalços devidos ao problema de Anchieta.

Antes de apresentar o sistema de anotação para o kadiwéu, vale lembrar algumas das propriedades gerais da abordagem que seguimos, que independem das línguas consideradas. Trata-se de uma anotação de tipo "Treebank", em que estão anotados os sintagmas categoriais e suas funções na oração. O quadro teórico de referência é a gramática gerativa. Contudo, é preciso ressaltar que o sistema de anotação não tem como objetivo propor uma análise detalhada dos enunciados, em perfeita coerência com tal ou tal modelo, mas permitir a recuperação de aspectos fundamentais da estrutura sintática das frases a partir de grandes corpora digitais. Isso legitima alguns desvios notáveis em relação a princípios da teoria gerativa. O mais importante é o fato de que a teoria X-barra não é sistematicamente aplicada de maneira a que todo núcleo seja projetado num sintagma de mesma categoria. Isso vale em particular para o verbo e a ausência de VP. Os sujeitos e objetos diretos são assim ambos irmãos do verbo. A razão disso é a dificuldade de localizar os limites do VP em línguas em que o verbo se move para posições mais altas da oração. Outras categorias nunca projetam, como a Negação (NEG), e algumas projetam somente em certos contextos (por exemplo dentro dos sintagmas nominais os adjetivos só projetam se tiverem modificadores ou complementos). O

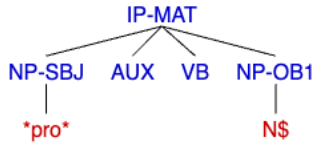
[13] A partir de uma ideia inicialmente implementada por Beatrice Santorini para o francês.
[14] Cf. https://corpussearch.sourceforge.net/

resultado é uma estrutura mais achatada, com ramificações múltiplas. As informações perdidas no nível estrutural são compensadas por subetiquetas sintáticas que expressam as funções dos sintagmas, em particular, dos sintagmas nominais (cf. 2.2.1).
Na continuação da seção, apresentaremos sucessivamente a anotação dos sintagmas nominais (2.2.1), dos sintagmas adjetivais e adverbiais (2.2.2), e das orações (2.2.3).[15]

2.2.1. Sintagmas nominais

No que diz respeito aos sintagmas nominais, as funções anotadas são, o sujeito (NP-SBJ), o objeto de um só verbo (NP-OB1), o objeto aplicativo de verbos seriais (NP-OB2)[16], o objeto de verbos aplicativos (NP-APL), os NPs locativos (NP-LOC), os NPs adverbiais (NP-ADV), os NPs deslocados (NP-LFD), e os NPs predicativos (NP-PRD). Essas etiquetas são ilustradas nas frases a seguir. Note-se que os sujeitos e objetos diretos são frequentemente nulos, nesse caso eles dominam a categoria vazia *pro* (cf. nota 9), e estão anotados, por convenção, no início do IP.

(29) Objeto direto (NP-OB1)
ejigo jiwi ionigi
AUX VB N$
'Eu vou ver meu filho'



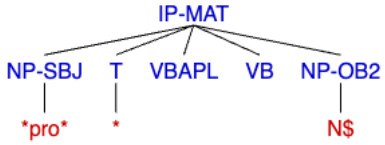
[15] Usamos também as seguintes categorias vazias, que são uma das marcas desse tipo de anotação:
1. para as categorias sintagmáticas:
- *pro* para os sujeitos e objetos nulos interpretáveis como pronomes lexicais;
- *T* para os vestígios de movimento-QU, nas orações relativas e interrogativas;
- *ICH* para as categorias vazias associadas à topicalização de sintagmas;
- WNP * para os operadores nulos, quando somente o complementador é realizado lexicalmente em construções QU.
T e *ICH* estão coindexados com o sintagma deslocado
2. para os núcleos:
- T* para Tempo não realizado lexicalmente;
- EXT* existencial vazio
- VB* verbo leve vazio
A questão, levantada por um revisor anônimo, de saber quando se anota um elemento ausente como categoria vazia, e quando não, tem parte da sua resposta ligada à teoria – por exemplo as construções relativas e interrogativas são anotadas como derivando do movimento de um operador -QU que deixa um vestígio coindexado com ele e expressando sua função na oração. No que diz respeito aos sujeitos e objetos nulos, a sua existência se justifica pelo fato de alternarem com sujeitos e objetos preenchidos. Isso explica por que não anotamos sujeitos de orações infinitivas controladas em línguas como o inglês ou o português. Em kadiwéu argumentamos que não há orações infinitivas, portanto, a questão não se coloca (mas ver nota 17). Esse raciocínio vale para o Tempo, anotado vazio quando não é realizado morfologicamente. Os existenciais vazios e verbos leves vazios são mais problemáticos uma vez que nunca têm realização, bem como a cópula vazia. Nesse caso, é uma decisão tomada para fins de busca. Nas orações copulativas, não é necessário anotar a cópula, uma vez que os NPs predicativos têm uma etiqueta especial.

[16] NP-OB1 e NP-OB2 são usados na anotação do inglês, respectivamente para objetos diretos e indiretos. Para saber mais sobre serialização em kadiwéu, ver Sandalo (1997).

(30) Serialização verbal (NP-OB2)

joleGatibige	ejinaGa	nemaGa
VBAPL	VB	N$

Procuramos falamos os mortos (= procuramos e reportamos ao achar os mortos)



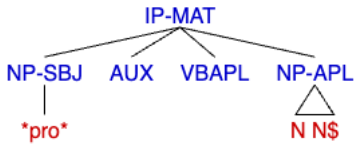
Observe que marcamos como T* o tempo vazio, uma vez que o presente não tem realização lexical em kadiwéu.

Os NPs objetos de verbos aplicativos são anotados NP-APL e os NPs locativos, NP-LOC:

(31) NP aplicativo (NP-APL)

ejigo	joletigi	exate	lamodi
AUX	VBAPL	N	N$

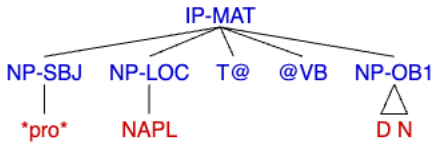
'Eu vou procurar folha de bacuri.'



(32) NP locativo (NP-LOC)[17]

naigitece	jonoGonadi	ica	epakagigo
NP-LOC	T+VB	D	N

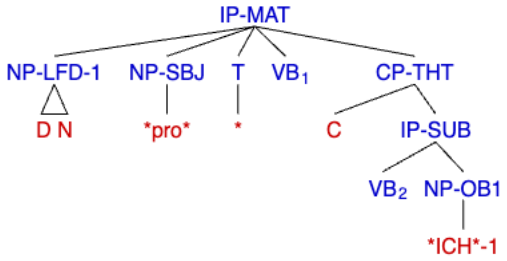
'Pelo caminho, viram uma ema'



(33) NP deslocado (NP-LFD) [18]

naGaniemokaya	jakadi	me	jolaGa
D	N	VB	C	VB
Essas	bocaiuvas	consigo	que	cozinha-las

'Essas bocaiuvas, eu posso cozinha-las'

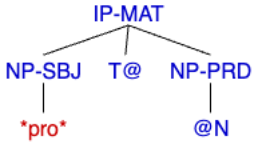


17 As categorias de palavras marcadas com + na anotação POS são separadas na sintaxe e marcadas com o sinal @.

18 Discutiremos as orações contidas nesse exemplo, bem como no próximo, na seção 2.2.3

(34) NP predicativo (NP-PRD)

jeGeijeeGagi
T+N
'Ja era animal'

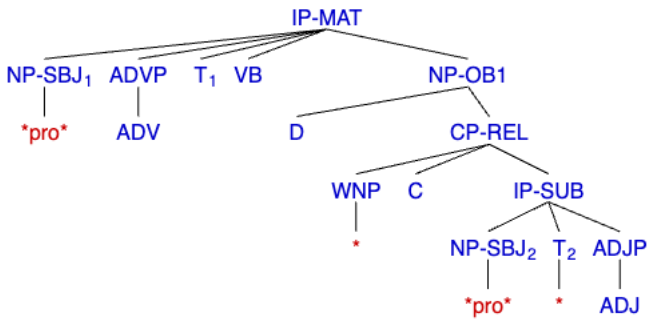


2.2.2 Outras categorias

Há ainda ADJP e ADVP como pode ser observado no exemplo abaixo:[19]

(35)	natigide	domaGa	joletibige	ica	me	ele
	ADV	T	VB	D	C	ADJ
	Agora	Prosp	procurarei	o	que	bom

'Agora eu vou procurar o que é bom'



2.2.3 Orações

O sistema de anotação das orações se faz em dois níveis distintos, IP (Inflectional Phrase) e CP (Complementizer phrase). O primeiro não envolve subordinador nem elemento que justifique a presença de um complementador (C).

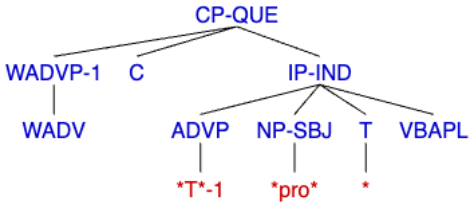
Como no português e no inglês, as orações matrizes, em que nenhum elemento da categoria C é realizado na periferia esquerda são anotadas IP-MAT, como em (29)-(35).

IP-SUB é o rótulo das orações subordinadas dominadas por CPs subordinados, como em (34). Os IPs dominados por CPs raízes são por sua vez etiquetados IP-IND (36).[20]

(36)	igaa	me	adopiliti
	WADV	C	VBAPL
	Por que	que	você voltou?

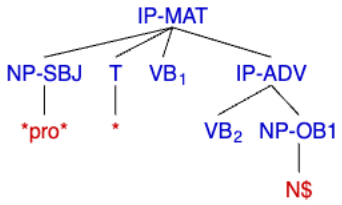
Por que que você voltou?

[19] Sobre os detalhes da estrutura (35), ver mais abaixo a anotação das orações relativas.
[20] Cf. mais abaixo a anotação das orações interrogativas.



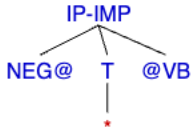
Consideramos mais um tipo de oração subordinada, que corresponde ao grupo dos IPs, uma vez que não tem complementador. Trata-se do IP-ADV, que se diferencia de construções seriais por não ter complemento compartilhado, e que é geralmente traduzido em português pelos falantes de kadiwéu como gerúndio. Quando temos IP-ADV, embora tenhamos dois VBs em sequência, como no caso de verbos seriais, o segundo VB é o núcleo de uma sentença adjunta, e projeta um IP independente, como ilustrado na árvore em (37):

(37) yelowadi inoke lotoinaGadi
VB VB N$
'Matou quebrando seu pescoço'



Há também orações imperativas, anotadas IP-IMP:

(38) adotaGaneGeni
NEG+VB
'Não fale'

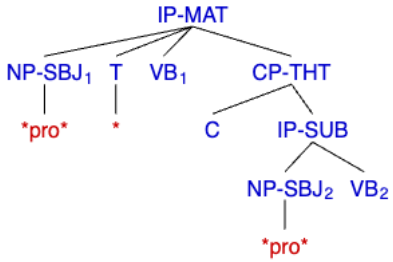


Quanto às orações subordinadas introduzidas por complementador, existem diversos tipos correspondendo ao que se encontra em línguas como o português e o inglês:

- orações completivas (CP-THT)
Estas orações são introduzidas pela conjunção *me*.[21]

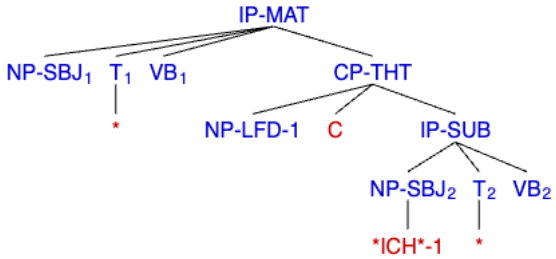
(39) ayema me diote
NEG+VB C VB
'Não quer dormir'

[21] As orações controladas nunca apresentam tempo. Não anotamos, portanto, Tempo vazio. Isso seria uma semelhança com as línguas como o português, em contradição com o que dissemos acima sobre a ausência do contraste oração finita/oração infinitiva. Deixamos essa questão para futuras discussões.



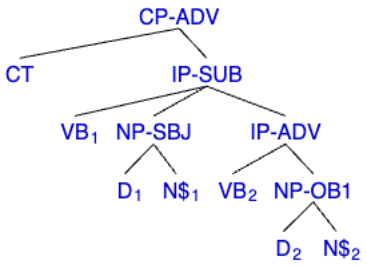
Note que orações subordinadas introduzidas por *me* sempre é possível deslocar o sujeito para antes deste complementador. Nesse caso, como anotado em (40) anota-se um sujeito nulo *ICH* na oração subordinada.

(40) Maria yema Pedro me diote
NPR VB NPR C VB
'Maria quer que Pedro durma'



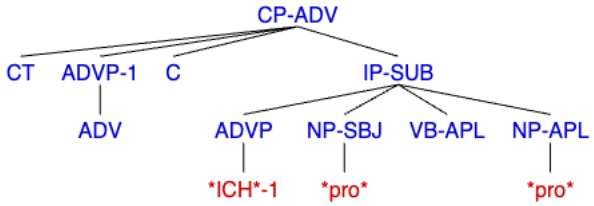
- Orações adverbiais (CP-ADV)
As orações adverbiais são introduzidas por subordinadores que carregam Tempo, como apresentado na seção 2.1.

(41) naGa
naGa didele icoa lioneGa ibake ica lodajo
CT VB D N$ VB D N$
'Quando lutou o jovem usando a sua faca'



O complementador *me* pode também aparecer em orações adverbiais quando há topicalização de algum sintagma da oração. Nesse caso, a categoria vazia *ICH* carrega a função do elemento topicalizado, com o qual está coindexado, dentro de IP.

(42) naGa owidijegi me ixomagatedijo
CT ADV C VBAPL
'Quando o colocou dentro dele pela última vez ....'

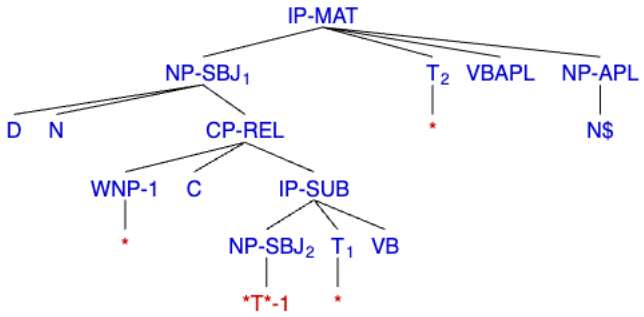


Esse tipo de sentença nos dá uma informação preciosa sobre a periferia esquerda do kadiwéu, evidenciando a posição baixa de *me* em relação aos outros complementadores.

- Orações relativas (CP-REL)
O kadiwéu também tem orações relativas introduzidas pela complementador *me* ou pelo pronome *ane* (o qual/a qual). O primeiro tem o valor de relativa restritiva, e o segundo se assemelha mais a uma relativa explicativa.[22]

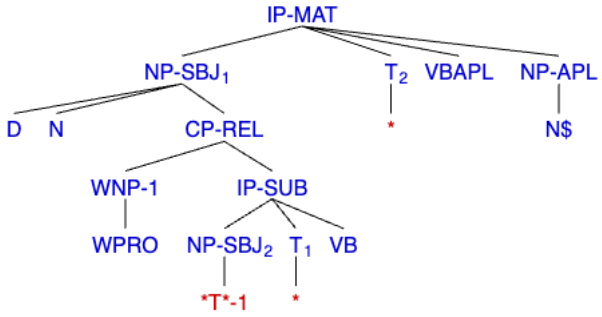
(43)	ani	wetiGa	me	iwaGadi	eniteloco	iGonagi
	D	N	C	VB	VBAPL	N$
	a	pedra	que	pesa	caiu em	meu pé

‘a pedra que é pesada caiu no meu pé’



(44)	ani	wetiGa	ane	iwaGadi	eniteloco	iGonagi
	D	N	WPRO	VB	VBAPL	N$
	a	pedra	a qual	pesa	caiu em	meu pé

‘a pedra, que é pesada, caiu no meu pé’



Note que, seguindo o sistema de anotação do português e do inglês, a diferença entre os dois tipos de relativas é que no primeiro, o elemento-QU é nulo, havendo somente o

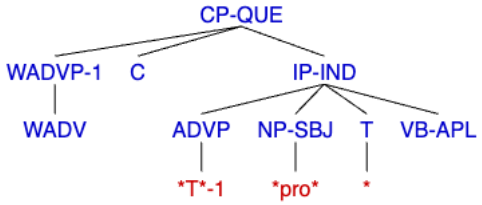
[22] Agradecemos a Vanda Pires por nos ter esclarecido a diferença entre *me* e *ane* nas orações relativas. Note-se que essa distinção lembra aquela observada em inglês entre *that* e *which*. Griffiths (1987) discute as relativas em kadiwéu com *ane* apenas.

complementador preenchido, enquanto no segundo caso, há um pronome relativo que projeta o sintagma relativo WNP. Nos dois casos, WNP é coindexado com o vestígio que carrega a função sintática (aqui o sujeito).

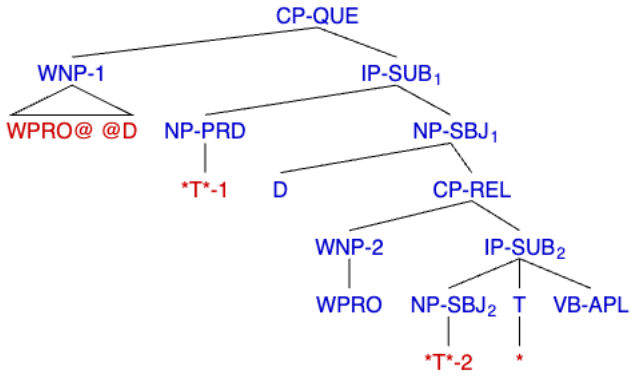
- Orações interrogativas (CP-QUE)

Para terminar, ainda se devem mencionar as orações interrogativas, em que o sintagma-QU se desloca para o início da oração. Como no inglês e no português, essas orações são rotuladas *CP-QUE*. Vale notar que *me* ocorre também depois do elemento-QU, quando se trata de um adjunto. Quando o sintagma interrogativo é sujeito ou objeto, a construção envolve oração relativa, introduzida por *ane*.

(45) igaa me adopiliti
WADV C VBAPL
Por que que você voltou?
Por que que você voltou?



(46) amiijo ica ane dopiliti
WPRO+D D WPROVBAPL
'Quem é que voltou?



Mostramos até aqui que o kadiwéu pode ser, em grande parte, anotado sintaticamente de maneira semelhante ao inglês e ao português. Na próxima seção, porém, evidenciaremos a existência, na língua, de construções que resistem a esse enquadramento, sendo necessária mais pesquisa para esclarecer a sua estrutura sintática.

3. DESAFIOS PARA A ANOTAÇÃO

Várias dúvidas, listadas a seguir, surgem quanto à anotação de construções com a conjunção *me*. Como já observamos, *me* encabeça vários tipos de estruturas com diferentes valores. Algumas correspondem claramente a construções encontradas em outras línguas, e não colocam problemas de anotação. Mas em certos casos, é difícil identificar a estrutura, bem como decidir se se trata de subordinações a sintagmas nominais (ou seja, orações relativas) ou de orações dominadas por outros IPs, uma vez

que, como vimos, *me* aparece com essas duas funções. No segundo caso encontramos construções que se assemelham a clivadas (47) e outras, a comparativas elípticas, em que a tradução de *me* em português poderia ser *como* (48). A marca de aplicativo em (48) sugere a presença de um argumento nulo, e a possível interpretação da oração encabeçada por *me* como uma relativa.

(47) oda aGaleGacowa me dabiditedi
CONJ NEG+ADV+D **C** **VB**
Então não mais ela que se levantou
'Então não foi mais ela/essa que se levantou'

(48) dinanatigi me negediogo
VBAPL **C** **N**
'Ela virava em (algo) que era onça'

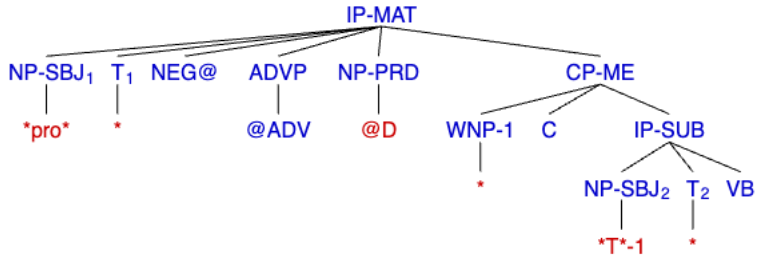
(49) jaGadowa me iwalo
T+D **C** **N**
'Ela era como mulher ...'

Os CPs com *me* podem ainda ser recursivos, como mostramos que ocorria no DP no caso de estruturas genitivas.

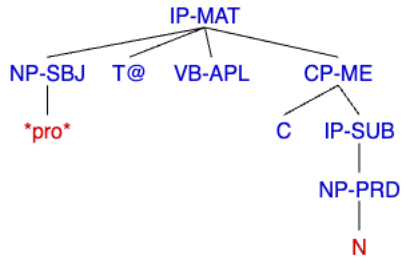
(50) ica me ele ica me idi ica Gonibole
D C ADJ D C D D N$
Essa que boa essa que ela a nossa-carne
'Essa é a nossa carne e ela é boa.'

Dada a variedade de sentenças introduzidas por *me*, e para evitar tropeçar no problema de Anchieta, nossa proposta é anotar, num primeiro momento, todas as sentenças encabeçadas por *me*, inclusive aquelas que não parecem oferecer dúvidas, como CP-ME. Isso facilitará a tarefa da ferramenta de anotação, sem perder informação a nível da busca uma vez que os diferentes tipos de CP-ME podem ser recuperados pelo contexto sintático: os CP-MEs completivos são irmãos de verbos (cf. 39-40), os CP-MEs relativos são irmãos de N (Cf. 43). Por contraste, os CP-MEs clivados e comparativos não são irmãos de verbos nem de nomes (são orações copulares), e os segundos só dominam um sintagma nominal. No caso de ambiguidade (como em 47), o default será a dominância imediata por IP. Isso é representado em 47'-49'.

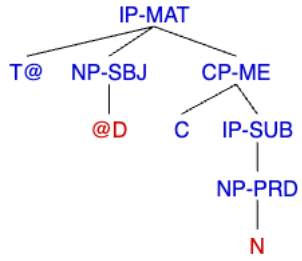
(47')



(48’)



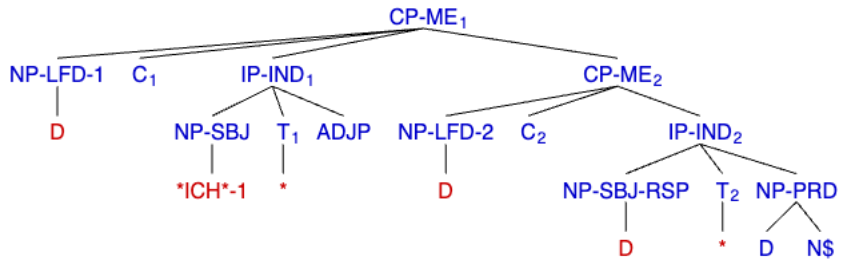
(49’)



Note que em (47’) anotamos no CP-ME um operador nulo ligando um vestígio sujeito para representar o fato de que interpretamos o sujeito dessa oração como correferente do predicado da oração principal. Isso é verdade em qualquer interpretação, clivada ou relativa. Já em (48’) e (49’), por razões de simplicidade, só o sintagma nominal é anotado dentro do CP-ME.

Finalmente, como ilustrado em (50), a sequência *Tópico me* ocorre em orações independentes que se assemelham ao que chamamos CP-D em português, por terem um valor discursivo diferenciado em relação às matrizes simples. Nessas orações, novamente, um tópico (anotado NP-LFD) precede o complementador. Manteremos a simplificação proposta acima, que consiste em também rotular essas sentenças como CP-ME, uma vez que elas são facilmente recuperáveis por serem raízes. Anotamos assim (50) com a estrutura (50’):[23]

(50’)



3. CONSIDERAÇÕES FINAIS

Neste artigo, propomos uma anotação morfossintática completa para o kadiwéu inspirada do sistema criado para o inglês dos *Penn Parsed Corpora of Historical English*, e adaptado ao português do *Corpus anotado do português histórico Tycho Brahe*. Parte dessa anotação, o que diz respeito à rica morfologia da língua e às classes de palavras já

[23] No segundo CP-ME, parece haver um elemento lexical resuntivo (idi), por isso, acrescentamos a sub-etiqueta -RSP no sujeito.

foi implementada no Corpus kadiwéu da *Plataforma Tycho Brahe*, e está disponível para buscas.

Ao longo da apresentação, discutimos a seguinte questão que intitulamos "Problema de Anchieta": é possível enquadrar em sistemas concebidos para línguas indo-europeias uma língua polissintética da família Guaikuru falada na América do Sul? A nossa resposta, empiricamente baseada no que foi realizado até agora, bem como teoricamente inspirada no pressuposto de que as línguas são todas o produto da Gramática Universal, é positiva, resguardado o necessário cuidado com as diferenças. A primeira dessas diferenças é, no que diz respeito à língua kadiwéu, uma morfologia que não cabe no sistema estabelecido para o português, de sub-etiquetas morfológicas afixadas às etiquetas de classe de palavras. Uma terceira camada de anotação foi, portanto, criada para anotar os morfemas que compõem as palavras. A segunda dessas diferenças são construções sintáticas que, ao contrário de muitas outras, não cabem claramente nas categorias estabelecidas para o inglês e o português, pelo menos no estado do nosso conhecimento da língua. Não querendo fazer caber à força uma língua nas outras, decidimos aplicar uma anotação mais genérica para todas as orações encabeçadas pela conjunção *me*, que aparece em numerosos contextos, alguns familiares, outros opacos. Podemos aguardar assim a inclusão de muito mais dados na base sintaticamente anotada, para voltar, se possível, a um sistema mais explícito empiricamente confiável.

Achamos que esse desafio vale a pena ser enfrentado, por duas razões essenciais. A primeira que a elaboração de um sistema de anotação para o kadiwéu permite que essa língua entre no mundo digital, apesar da paucidade atual dos documentos compondo o *Corpus kadiwéu*. Como dissemos na introdução a anotação multiplica o valor do corpus, mesmo com uma quantidade de palavras limitada. A inserção no mundo digital assegura à língua uma forma robusta de preservação da sua cultura e sua língua. A segunda razão é que o processo de anotação de uma língua com tradição gramatical limitada tem um valor heurístico muito forte. Com efeito, nos desafia a analisar cada sentença dos textos presentes no corpus, de maneira consistente com o sistema de anotação e não somente aquelas que são escolhidas para representar tal ou tal fenômeno. Ressaltamos de novo que não se trata de propor uma análise definitiva, mas fazer com que os diferentes fenômenos produzidos pela gramática da língua possam ser recuperados no corpus, de maneira exaustiva, para posterior análise. Para que isso seja possível, esse trabalho também servirá de base à construção de um analisador baseado em regras que permitirá fazer buscas sintáticas, no modelo daquele que usamos para o português, usando a linguagem de *Corpus Search*. Esse trabalho também não prescinde da participação dos falantes nativos, cuja intuição sobre a língua permite esclarecer muitas dúvidas sobre a anotação. Enfim, ele constitui a base de gramáticas pedagógicas, elaboradas também em colaboração com os falantes.

REFERÊNCIAS

BRITTO, H., Finger, M., GALVES, C. (2002) Computational and linguistic aspects of the construction of the Tycho Brahe Parsed Corpus of Historical Portuguese. Romance Corpus Linguistics - Corpora and Spoken language. Tubingen: Narr.

FARIA, P., GALVES, C., MAGRO, C. (a sair) Syntactic annotation for Portuguese corpora: standards, parsers, and search interfaces. *Language Resources and Evaluation, Especial Issue on Computational Approaches to Portuguese*.

FINGER, M. (2000) Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe. Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR2000).

CHIERCHIA, G. (1998). 'Plurality of mass nouns and the notion of "semantic parameter"'. In 'Events and Grammar', 53–103. Springer Netherlands.
DOETJES, J. (2017). Measure words and classifiers. Letras 96: 291-308.
GALVES, C., SANDALO, F., SENA, T.A., VERONESI, L. (2017) Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos da Linguagem*, (59.3), pp. 631-648.
GRIFFITHS, G. (1987). *Relative Clause Formation and other Word Parameters in Kadiwéu*. Reading University master thesis.
GRIFFITHS, G. (2002). *Dicionário da Língua Kadiwéu*. SIL ms. https://www.sil.org/system/files/reapdata/74/06/08/74060839706011162756896570533590209458/KDDict.pdf.
KROCH, A., TAYLOR, A., SANTORINI, B. 2000-. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4).
KROCH, A., SANTORINI, B., DELFS, L. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3).
KROCH, A., SANTORINI, B., DIERTANI. A. 2016. The Penn Parsed Corpus of Modern British English (PPCMBE2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1).
MAGRO, C., GALVES, C. (2019) Portuguese Syntactic Annotation Manual. http://alfclul.clul.ul.pt/portuguesesyntacticannotation/
NEVINS, A., and SANDALO, F. (2011). Markedness and morphotactics in Kadiwéu. [+participant] agreement. *Morphology* 21(2): 351-378.
PYLKKANEN, L. (2002). Introducing Arguments. MIT PhD Dissertation.
SANDALO, F. (1997). *A Grammar of Kadiwéu with Special Reference to the Polysynthesis Parameter*. MIT Occasional Papers in Linguistics 11.
SANDALO, F. (2009). Person hierarchy and inverse voice in Kadiwéu. *LIAMES* 9: 27-40.
SANDALO, F. (2020). Individuation, counting, and measuring in the grammar of Kadiwéu. *Linguistic Variation* 20(2): 239-254.
SANDALO, F., and Michelioudakis, D. (2016). Classifiers and Plurality: evidence from a deictic classifier language. Baltic International Yearbook of Cognition, Logic and Communication, 11: 1-40.
SANDALO, F. (2023). Evidencialidade Reportativa, Tempo e Negação em kadiwéu. Liames 23(1) https://doi.org/10.20396/liames.v23i00.8671197
SANDALO, F. (2023b). On the Guaikuruan inverse system: interpreting Kadiwéu and Mocoví person hierarchies. International Journal of American Linguistics 89(1). https://doi.org/10.1086/722239
SANTORINI, B. (2022) Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence https://www.ling.upenn.edu/hist-corpora/annotation/index.html
SENA, T. A. (2016). *Obviação em Kadiwéu*. University of Campinas Master Thesis, Brazil.
ROTHSTEIN, S. (2011). 'Counting, measuring, and the semantics of classifiers'. The Baltic International Journal of Cognition, Logic, and Communication 6: 1-42.
TENNY, C. (2016). Evidentiality, Experiencers, and the Syntax of sentence in Japanese. Journal of East Asian Linguistics (2006) 15: 245-288 Springer 2006 DOI 10.1007/s10831-006-0002-x

Recebido: 26/5/2023
Aceito: 18/8/2023
Publicado: 18/9/2023