Rev. Estud. Ling., Belo Horizonte, v. 29, n. 3, p. 1717-1777, 2021

Uma gramática computacional de um fragmento do nheengatu

A computational grammar for a fragment of Nheengatu

Leonel Figueiredo de Alencar
Universidade Federal do Ceará (UFC), Fortaleza, Ceará / Brasil
leonel.de.alencar@ufc.br
https://orcid.org/0000-0001-8148-6994

Resumo: A disponibilidade de recursos para o processamento computacional constitui um dos fatores de sobrevivência de uma língua. O objetivo deste trabalho foi implementar um fragmento do nheengatu no formalismo *Grammatical Framework*, especialmente projetado para o desenvolvimento de aplicações multilíngues. Outrora mais falado que o português na Amazônia, o nheengatu está ameaçado de extinção, embora ainda conte com estimados 14000 falantes. O fragmento restringe-se a orações que expressam estados contingentes e não-contingentes, mas inclui fenômenos gramaticais estruturalmente complexos típicos da família tupi-guarani, os quais contrastam fortemente com as construções equivalentes em português e inglês. Constitui um dos módulos da GrammYEP, uma gramática computacional multilíngue que integra módulos análogos do inglês e do português. A implementação tomou como ponto de partida as descrições gramaticais não formalizadas de Navarro (2011) e Cruz (2011). A formalização revelou lacunas e inconsistências nessas abordagens, em parte sanados por meio de uma reanálise dos dados. A GrammYEP alcançou resultados bastantes satisfatórios na tradução do e para o nheengatu. Traduziu para o português e o inglês a totalidade de um conjunto-teste de 142 sentenças dessa língua. Inversamente, verteu para o nheengatu 98,18% e 84,11% dos conjuntos-teste correspondentes em português e inglês. Por outro lado, analisou apenas dois exemplos de um conjunto-teste negativo com 171 construções agramaticais em nheengatu. Desta avaliação resultou um *treebank* com 243 sentenças do nheengatu, emparelhadas com as sentenças equivalentes em português e inglês.

Palavras-chave: língua geral amazônica (LGA); tupi moderno; predicação qualificativa; construção possessiva; tradução automática; linguística computacional; processamento de linguagem natural.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.3.1717-1777

Abstract: The availability of resources for computational processing is one of the survival factors of a language. The goal of this work was to implement a fragment of Nheengatu in the Grammatical Framework formalism, specially designed for the development of multilingual applications. Once more widely spoken than Portuguese in the Amazon region, Nheengatu is threatened with extinction, although it still has an estimated number of 14,000 speakers. The fragment is restricted to sentences that express contingent and non-contingent states, but includes structurally complex grammatical phenomena typical of the Tupí-Guaraní family, which strongly contrast with the equivalent constructions in Portuguese and English. It constitutes one of the modules of GrammYEP, a multilingual computational grammar comprising equivalent English and Portuguese modules. The starting point of the implementation was the non-formalized grammatical descriptions of Navarro (2011) and Cruz (2011). The formalization revealed gaps and inconsistencies in these approaches, which were partly remedied through a reanalysis of the data. GrammYEP achieved quite satisfactory results in the translation from and to Nheengatu. It translated into Portuguese and English all examples from a test set with 142 Nheengatu sentences. Conversely, 98.18% and 84.11% of the corresponding Portuguese and English test sets were rendered into Nheengatu. On the other hand, it parsed only two examples from a negative test set with 171 ungrammatical constructions in Nheengatu. This evaluation resulted in a treebank with 243 Nheengatu sentences, paired with the equivalent sentences in Portuguese and English.

Keywords: Amazonian Lingua Franca; Modern Tupí; qualifying predication; possessive construction; machine translation; computational linguistics; natural language processing.

Recebido em 27 de agosto de 2020

Aceito em 04 de novembro de 2020

1 Introdução

Uma gramática computacional é uma descrição formal das estruturas sintáticas e lexicais de uma língua capaz de ser utilizada pelo computador para análise ou geração de sentenças (DUCHIER; PARMENTIER, 2015). Esse tipo de recurso integra a arquitetura de diversas tecnologias de linguagem natural, como tradutores automáticos, sistemas de diálogo etc. (JURAFSKY; MARTIN, 2009). Por outro lado, constitui valioso instrumento para a linguística teórica, ao permitir a verificação automática da consistência interna e da validade empírica de uma determinada modelagem de fenômenos gramaticais (BENDER, 2008, 2010; MÜLLER, 2015).

Um dos fatores de sobrevivência de uma língua nos dias de hoje é possuir recursos para o processamento computacional (PIRINEN *et al*., 2017). Entre os mais urgentes figuram gramáticas computacionais, por permitirem a construção de vários outros recursos, como *treebanks* (HAJIČOVÁ, 2010).

Neste artigo, apresentamos um fragmento de gramática do nheengatu no formalismo computacional *Grammatical Framework* (RANTA, 2011), doravante GF, especialmente projetado para facilitar o desenvolvimento de aplicações multilíngues, incluindo tradutores automáticos, sistemas de diálogo e ferramentas de aprendizagem de línguas mediada por computador. Esse fragmento integra a GrammYEP, uma gramática computacional multilíngue da qual fazem parte fragmentos equivalentes do inglês e do português. Nessa sigla, as três línguas do sistema estão identificadas pela primeira letra dos respectivos códigos no padrão ISO 639-3, respectivamente, *yrl*, *eng* e *por* (EBERHARD; SIMONS; FENNIG, 2020). O primeiro deriva de *yeral*, designativo em espanhol do nheengatu, também conhecido como geral, língua geral amazônica ou tupi moderno, entre outros termos (RODRIGUES, 1996; EBERHARD; SIMONS; FENNIG, 2020).

Descendente do tupinambá, língua tupi-guarani que era falada no norte do Brasil até o século XVIII (FREIRE, 2011; RODRIGUES, 1996), o nheengatu é a L1 ou L2 de estimadas 6000 pessoas no município de São Gabriel da Cachoeira, na região do alto rio Negro, e 8000 em território colombiano, pertencentes a etnias originalmente falantes de línguas não tupis, quase todas extintas ou moribundas (CRUZ, 2011; EBERHARD; SIMONS; FENNIG, 2020). Praticamente desaparecido na Venezuela e ameaçado de extinção tanto na Colômbia quanto no Brasil, ainda é usado neste último pela população em idade reprodutiva, porém, a transmissão às crianças está sendo interrompida (EBERHARD; SIMONS; FENNIG, 2020). Felizmente, nos últimos anos, tem passado por um processo de revitalização, impulsionado por traduções literárias (NAVARRO; ÁVILA; TREVISAN, 2017) e cursos em universidades brasileiras, despertando o interesse também do público não indígena.

Vários outros fatores concorreram para a escolha do nheengatu como protótipo inicial de um projeto maior de implementação de gramáticas computacionais de línguas indígenas brasileiras. Em primeiro lugar, pesou a sua importância histórica, por ter sido, durante dois séculos e meio, "a principal língua da Amazônia", posição que perderia para o

português apenas na segunda metade do século XIX, conforme Freire (2011, p. 16-17), certamente o mais abrangente levantamento da história social da língua geral amazônica da sua origem no século XVII ao século XXI. Afirma Aryon Dall'Igna Rodrigues no prefácio desse livro:

> A história da Língua Geral Amazônica é, sem dúvida, uma das mais interessantes nas Américas e deve ser conhecida não só pelos estudiosos da História do Brasil, mas também por todos os que estudam as Ciências Sociais, as Letras e a Linguística neste país. (RODRIGUES, 2011, p. 13)

O segundo fator decisivo foi a existência de uma descrição normativa didática de acesso fácil e gratuito em formato de PDF editável (NAVARRO, 2011). Finalmente, o GF não havia sido aplicado antes a qualquer língua ameríndia. A implementação de fenômenos gramaticais de uma língua do tronco tupi oferece a oportunidade de testar a abrangência do formalismo, possibilitando detectar eventual aspecto de difícil ou impossível implementação.

A GrammYEP é um projeto em andamento. Cobre, na versão atual, cerca de um quinto do conteúdo gramatical de Navarro (2011), incluindo diversos fenômenos de considerável complexidade gramatical típicos da família tupi-guarani, exemplificados em (1)-(17).[1]

(1) kambi s-aku u-iku
leite 3s.INACT-quente 3s.ACT-estar
'o leite está quente'

(2) kuá meiú-itá ta r-aku u-iku
DEM.PROX beiju-PL 3p.INACT N3s-quente[2] 3p.ACT-estar
'estes beijus estão quentes'

[1] Na glosas interlineares, seguimos as *Leipzig Glossing Rules* (LGR), complementado-as com símbolos propostos por Gynan (2017) para determinados morfemas do guarani, a saber: ACT=ativo, INACT=inativo, IMPR=impessoal (*impersonal* em inglês) e RLL= relacionalizador.

[2] A abreviatura N3s segue convenção das LGR, onde N- corresponde ao prefixo *non-* 'não'. Marca concordância com todas as pessoas, exceto a 3ª do singular.

(3) nhaã t-imbiú sepiasu retana
DEM.DIST IMPR-comida caro muito
'aquela comida é muito cara'

(4) kuá-itá pirá piranga sepiasuíma u-iku
DEM.PROX-PL peixe vermelho barato 3p.ACT-estar
'estes peixes vermelhos estão baratos'

(5) s-endaua puranga
3s.INACT-comunidade bonito
'a comunidade dele é bonita'

(6) s-uka s-uri
3s.INACT-casa 3s.INACT-alegre
'a casa dele é alegre'

(7) i igara i pusé u-iku
3s.INACT canoa 3s.INACT pesado 3s.ACT-estar
'a canoa dele está pesada'

(8) (ixé)[3] se pusé
(eu) 1s.INACT pesado
'(eu) sou pesado'

(9) (ixé) se r-uri
(eu) 1s N3s-alegre
'(eu) estou alegre'

(10) (aé) s-uri[4]
(ele) 3s.INACT-alegre
'(ele) é alegre'

---

[3] Nos exemplos deste artigo, constituintes entre parênteses são opcionalmente realizáveis.

[4] O nheengatu não marca gênero nos pronomes e prefixos flexionais pessoais. Nas traduções para o português, traduzimos esses elementos indistintamente pelo masculino ou feminino.

(11) (ixé) a-iku iké
(eu) 1s.ACT-estar aqui
‘(eu) estou aqui’

(12) (iandé) ti ia-iku t-endaua upé
(nós) NEG 1p.ACT-estar IMPR-comunidade em
‘(nós) não estamos na comunidade’

(13) (indé) ti ne r-uri re-iku
(tu) NEG 2s.INACT N3s-alegre 2s.ACT-estar
‘(tu) não estás alegre’

(14) Maria r-uka ti s-uaki
Maria RLL-casa NEG 3s.INACT-perto.de
‘a casa da Maria não é perto dele’

(15) kurumĩ r-endaua pe taua r-uaki
menino RLL-comunidade 2p.INACT cidade RLL-perto.de
‘a comunidade do menino é perto da cidade de vocês’

(16) nhaã x-imiriku kisé ti u-iku uka pupé
DEM.DIST 3s.INACT-esposa faca NEG 3s.ACT-estar casa dentro.de
‘aquela faca da esposa dele não está dentro da casa’

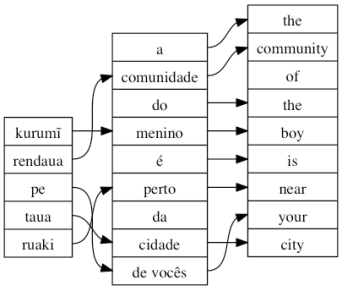
(17) nhaã Pedro pindá u-iku igara kuara upé
DEM.DIST Pedro anzol 3s.ACT-estar canoa buraco em
‘aquele anzol do Pedro está dentro da canoa’

Destaquemos três desses fenômenos. O primeiro é a distinção entre verbos ativos e inativos, flexionados por duas séries diferentes de prefixos de concordância. O segundo é a realização do argumento interno de substantivos e posposições pelos prefixos da série inativa. Finalmente, temos a divisão de verbos, substantivos e posposições em uniformes e multiformes. Enquanto em todas as traduções a mesma forma básica dos substantivos *comunidade* e *casa* é utilizada, nos exemplos correspondentes do nheengatu em (5)-(16) ocorrem três formas distintas

de cada lexema. Analogamente, os verbos inativos traduzidos pelos adjetivos *quente* e *alegre* e as posposições manifestam duas formas diferentes, que compartilham os mesmos prefixos *s-* e *r-* desses dois substantivos.

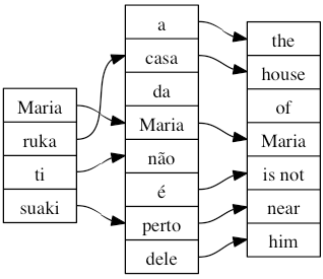
Características tipológicas constituem uma das principais dificuldades para a tradução automática (JURAFSKY; MARTIN, 2009). Um sistema como o que propomos precisa lidar não apenas com diferenças significativas entre o português e o inglês, de ramos distintos do tronco indo-europeu, mas também com discrepâncias estruturais ainda maiores em relação ao tronco tupi, como se pode constatar em (1)-(17). Não obstante isso, a GrammYEP possibilita a tradução automática entre as três línguas de forma satisfatória, como exemplificam as Figuras 1 e 2.

FIGURA 1 – Tradução automática de (15) com alinhamento das palavras



Fonte: Elaboração própria.

FIGURA 2 – Tradução automática de (14) com alinhamento das palavras[5]



Fonte: Elaboração própria.

[5] Na versão em inglês, a cópula e a negação ocupam uma mesma célula porque foram amalgamadas num único *token* para simplificar a implementação da contração *isn't*.

Na próxima seção, especificamos o fragmento do nheengatu abrangido pela GrammYEP. A seção seguinte trata do formalismo GF, utilizado na modelação computacional desses fenômenos. Na seção 4, descrevemos os aspectos mais importante da implementação. A seção 5 apresenta os resultados da aplicação do sistema na análise e tradução automáticas de sentenças. A última seção traz as conclusões, problemas remanescentes e perspectivas para a expansão da gramática.

2 Cobertura do fragmento gramatical

A implementação de uma gramática computacional de qualquer língua natural é uma tarefa extremamente complexa, que impõe a necessidade de segmentar o desenvolvimento em sucessivas fases de crescente nível de complexidade e abrangência (FRANCEZ; WINTNER, 2012). De fato, não se limita a converter, para um dado formalismo, as descrições linguísticas disponíveis, mas implica também dirimir inconsistências e preencher lacunas destas, decorrentes da formalização insuficiente. Nesta seção, apresentamos o recorte do nheengatu do qual a GrammYEP, atualmente, implementa a maior parte. A quantificação do DP e construções possessivas com múltiplos núcleos nominais encaixados ficaram para uma próxima versão.

Na especificação desse fragmento, utilizamos a gramática independente de contexto (doravante CFG, do inglês *context-free grammar*), formalismo computacional que, pela sua simplicidade conceitual, é mais acessível a um público não especializado do que os formalismos mais complexos como a Gramática Léxico-Funcional (BRESNAN, 2001) e o GF. A limitação da CFG é dificultar a implementação de fenômenos que envolvem propriedades de subclasses de categorias sintáticas, como a concordância e a valência, tratados de forma elegante nesses outros modelos por meio de mecanismos mais expressivos (SAG; WASOW; BENDER, 2003). Desse modo, o fragmento apresentado nesta seção gera também muitos exemplos agramaticais, problema que será enfrentado com a modelação em GF descrita na seção 4.

Utilizamos como fontes de conhecimento sobre a língua, principalmente, duas das exposições gramaticais mais recentes e livremente disponíveis na Internet em formato de PDF editável, a saber, Cruz (2011) e Navarro (2011). Para dirimir dúvidas, consultamos Casasnovas (2006), exposição gramatical sucinta com muitos exemplos

traduzidos seguida de uma coletânea de lendas e glossários, que tem sido utilizada em cursos para falantes.[6] Constitui importante documentação da língua falada no rio Negro, mas as explicações gramaticais não são suficientemente detalhadas para não falantes. Por exemplo, não há qualquer menção à flexão de verbos inativos, da qual se apresentam apenas exemplos, como na p. 26, para ilustrar o tópico "Verbo SER, ESTAR (IKÚ)". Apresenta-se apenas a conjugação dos verbos ativos.

Cruz (2011) é uma tese de doutorado sobre a língua falada no alto rio Negro. A transcrição dos inúmeros exemplos, extraídos sobretudo de um corpus oral compilado a partir de textos autênticos, procura geralmente refletir a pronúncia da língua falada.

Navarro (2011, p. 7) segue uma orientação oposta. Propõe uma gramática normativa da língua com base "nos seus vários autores, mas respeitando os fatos linguísticos da língua geral falada hoje em dia, principalmente nos centros urbanos do médio e alto rio Negro." Consiste de treze lições que introduzem o vocabulário e as estruturas gramaticais de forma progressiva por meio de diálogos, narrativas, canções etc., acompanhados de exercícios. Esse manual é utilizado em cursos ministrados pelo autor na Universidade de São Paulo.[7]

Todos esses trabalhos são de inestimável valor e, ao nosso ver, complementam-se. Por um lado, Cruz (2011) permite dirimir lacunas de Navarro (2011), embora também suscite outras dúvidas, como veremos mais adiante. No caso de Navarro (2011), essas questões resultam, provavelmente, de uma preocupação pedagógica em evitar o jargão linguístico, atendo-se à nomenclatura tradicional da gramática do português. Casasnovas (2006), por sua vez, constitui um rico repositório de dados para cotejo com esses dois trabalhos.

---

[6] Conforme um dos avaliadores anônimos, "a gramática mais utilizada pelos falantes de Nheengatu". Infelizmente, quando da implementação da GrammYEP, não encontramos na Internet o arquivo completo desse manual, que é citado tanto por Navarro (2011) quanto por Cruz (2011). Agradecemos o envio de link para a imagem de um exemplar físico da segunda edição (CASASNOVAS, 2006). Esse arquivo, porém, apresenta alguns problemas. Faltam as páginas 94 e 95. Anotações manuscritas e trechos apagados dificultam a conversão para o formato de PDF editável e, consequentemente, a compilação de um corpus para buscas automáticas de exemplos. Uma edição de 2014, indicada pelo parecerista, não foi encontrada no site da editora.

[7] Disponível em: http://tupi.fflch.usp.br/curso-de-lingua-nheengatu-e-de-cultura-amazonica. Acesso em: 27 ago. 2020.

Por outro lado, acreditamos que o estabelecimento de uma norma padrão, sem inibir a riqueza da variação intralinguística, constitui fator de vitalidade de uma língua. Em primeiro lugar, permite o uso da língua em contextos não coloquiais, por exemplo, em textos legais, científicos etc. Em segundo lugar, promove o intercâmbio entre falantes de diferentes variedades, incluindo o diálogo intergeracional propiciado pela literatura de épocas passadas, da qual o nheengatu é particularmente rico.[8] Em terceiro lugar, favorece o desenvolvimento de aplicações de processamento de linguagem natural, para o qual a variação linguística constitui um entrave, ainda mais no caso de uma língua minoritária como o nheengatu. De fato, a relação custo-benefício de aplicações customizadas para determinadas variedades seria bastante desfavorável, dado o pequeno número de falantes de cada uma. Finalmente, facilita o aprendizado do idioma como língua estrangeira (LE) ou como segunda língua (L2), sobretudo por pessoas sem acesso a falantes nativos. Esse último aspecto é particularmente relevante no caso do nheengatu, que parece ocupar uma posição singular no quadro das línguas indígenas brasileiras pela expressiva quantidade de textos produzidos por usuários da língua como LE. Nesse repertório, sobressaem as traduções para o nheengatu de clássicos da literatura (NAVARRO; ÁVILA; TREVISAN, 2017).

Há muita divergência na representação ortográfica do nheengatu. Cruz (2011), Navarro (2011) e Ávila (2016), por exemplo, discrepam no uso do acento agudo. Diferentemente da primeira, os dois últimos, continuando tradição que remonta pelo menos a Sympson (1877), separam das respectivas bases os prefixos do Quadro 1, ver (7) e (18).[9] Casasnovas (2006, p. 26), por sua vez, grafa *peyumasí* (‘vocês são famintos’ na nossa tradução), amalgamando o prefixo de 2ª pessoa do plural *pe* ao verbo inativo *yumasí* ‘ser faminto’. Em *pe igara* ‘canoa de vocês’, porém, o mesmo prefixo é grafado separado (CASANOVAS, 2006, p. 34).

(18)	a-iku	i	irũmu
	1s.ACT-estar	3s.INACT	com
	‘estou com ela’		

[8] Ver levantamentos de Freire (2011) e Ávila (2016).

[9] A segunda edição de Navarro (2011), datada de 2016, atualmente indisponível, adota convenções ortográficas parecidas com as de Ávila (2016).

QUADRO 1 – Prefixos inativos silábicos conforme Navarro (2011)

1s.INACT	2s.INACT	3s.INACT	1p.INACT	2p.INACT	3p.INACT
se	ne	i	iané	pe	aintá (ta)

Fonte: Elaboração própria.

Os prefixos do Quadro 1 integram a série estativa ou série II, distinguindo verbos estativos ou inativos de verbos dinâmicos ou ativos (CRUZ, 2011; PRAÇA; MAGALHÃES; CRUZ, 2017). Desempenham três funções em nheengatu, conforme o tipo de base a que se adjungem. Quando a base é um verbo inativo flexionável, como na segunda ocorrência de *i* em (7), marcam a concordância com o sujeito da sentença. No caso de posposições e substantivos, realizam o argumento interno, como em (18) e na primeira ocorrência de *i* em (7).

A GrammYEP objetiva inicialmente auxiliar o autoestudo da língua a partir de Navarro (2011) por meio da implementação de um sistema de tradução automática e da compilação de *treebank*. Desse modo, preferimos seguir tanto suas convenções ortográficas quanto sua orientação normativa na definição do fragmento a ser modelado computacionalmente. Certamente, o estabelecimento de uma ortografia unificada e a adoção de uma norma padrão para uma língua indígena devem resultar de decisão por parte das comunidades de falantes.[10] A GrammYEP, porém, constitui software livre e de código aberto, podendo vir a ser adaptada por qualquer um com conhecimentos de programação em GF para refletir outras convenções gramaticais e ortográficas.

Dois critérios orientaram a delimitação do fragmento. Por um lado, tomamos como ponto de partida a minigramática bilíngue do inglês e do italiano com que Ranta (2011) introduz as noções fundamentais do GF, formalismo utilizado na especificação computacional objeto da seção 3. Essa minigramática (doravante ITALENG) é capaz de analisar e traduzir comentários sobre alimentos nessas duas línguas análogos aos exemplos (1)-(4). Por outro lado, seguimos, em linhas gerais, a progressão gramatical de Navarro (2011), de que destacamos, no Quadro 2, apenas os pontos que implementamos. Os títulos dos tópicos foram adaptados à terminologia adotada neste artigo. Entre parênteses, indicamos exemplos de cada fenômeno. A terceira coluna indica o percentual aproximado implementado de cada lição.

[10] Agradecemos a um dos pareceristas por essa observação.

QUADRO 2 – Tópicos gramaticais das lições de Navarro (2011) implementados

Lição	Conteúdo gramatical	Cobertura
1	I: flexão de verbos ativos (11)-(13) II: pronomes pessoais livres e prefixos inativos (7)-(13) III: modificadores nominais adjetivais (4) e predicação com verbos inativos (1) (5) (8) IV: substantivo como genitivo adnominal (14)-(16)	100%
2	I: posposições (12) (14)-(18) II: negação sentencial (12)-(14) IV: marcação de plural no domínio nominal (4)	60%
3	I: prefixos inativos na função de genitivo adnominal (5)-(7) II: prefixos inativos como complementos de posposições (14) (18) VI: determinantes demonstrativos (2)-(4)	40%
4	II – substantivos multiformes (5) (12) (15)	20%
5	V – verbos inativos multiformes (1) (6) (9)	10%
6	IV – posposições multiformes (14) (15)	20%
9	III – posposição *pupé* e locução posposicional *kuara upé* (17) (16)	10%

Fonte: Elaboração própria.

Coincidentemente, abstraindo das diferenças sintáticas superficiais entre inglês, italiano e nheengatu, as duas introduções, Ranta (2011) e Navarro (2011), esta à língua objeto, aquela à metalinguagem, focam inicialmente o mesmo tipo fundamental de construção linguística: a predicação qualificativa (*qualifying predication*) do tipo copulativo, que, conforme Mathesius (1975, p. 114), consiste na atribuição de uma qualidade (*qualificans*) a uma entidade (*qualificandum*), expressas, respectivamente, pelo predicado e pelo sujeito, ligados por um verbo do tipo de *ser* (inexistente em nheengatu) ou *estar*.

Preferindo, como Casasnovas (2006), uma nomenclatura mais próxima da portuguesa, Navarro (2011) diverge de Cruz (2011) no tratamento dos pronomes pessoais, verbos inativos e prefixos do Quadro 1, exigidos por parte desses verbos. Aos verbos inativos não flexionáveis denomina adjetivos de primeira classe e classifica os inativos flexionáveis em dois grupos, a saber, adjetivos e verbos de segunda classe, segundo correspondam em português a um adjetivo, como em (1), (2) e (6)-(10), ou a um verbo, como no caso de *kérpi* ‘sonhar’. Trata os prefixos flexionais

desses dois grupos, bem como na função de complemento de posposições, como pronomes pessoais de segunda classe, mas os denomina pronomes adjetivos possessivos quando adjungidos a substantivos. Os pronomes pessoais livres, na acepção de Cruz (2011), como *ixé* 'eu' em (8), são para Navarro (2011) pronomes de primeira classe.

Analogamente a *ser* e *estar* em português, mas diferentemente de *be* em inglês, o nheengatu lexicaliza, em orações predicativas do tipo de (1)-(18), a distinção entre predicado de nível de indivíduo (*individual level predicate*) e predicado de nível de fase *(stage level predicate*),[11] ou, conforme Cruz (2011, p. 475), entre estados não-contingentes e contingentes. Estes são marcados pelo auxiliar *iku* 'estar', enquanto aqueles não são marcados, compare-se (7) com (8) e (4) com (3).

Utilizando a nomenclatura de Navarro (2011), definimos em (19) um fragmento de gramática do nheengatu capaz de gerar sentenças como (1)-(10), que transcendem o recorte da ITALENG, uma vez que incluem pronome de primeira classe (Pron1) como *qualificandum* na função de sujeito da sentença e pronome possessivo adnominal (Poss). Pron2 refere-se aos pronomes de segunda classe.

(19) S→(NP) VP
VP→AP (V)
NP→(Det) (Poss) (AP) N (AP)
NP→Pron1
AP→(Pron2) A (Adv)
V→"aiku"|"uiku"
N→"igara"|"kambi"|"timbiú"|"sendaua"|"suka"|"meiú-itá"| "pirá"|"pirá-itá"
A→"pusé"| "saku"| "raku"| "ruri"| "suri"| "sepiasu"| "sepiasuíma"|"piranga"
Det→"kuá"| "nhaã"|"kuá-itá"
Poss→"i"|"se"|"ta"
Pron1→"ixé" | "aé"
Pron2→"i" | "se" | "ta"
Adv→"retana"

[11] Segundo Bentley (2017, p. 343), essa é apenas uma entre outras explicações para a alternância entre as cópulas ESSE e STARE em construções predicativas copulativas das línguas românicas, questão cujo aprofundamento ultrapassaria o escopo deste artigo.

Nos formalismos gramaticais derivados da CFG, como a LFG, a estrutura gramatical de uma língua natural é modelada em duas dimensões distintas (SAG; WASOW; BENDER, 2003). O nível sintagmático modela a combinação de categorias lexicais e funcionais para formar sintagmas, enquanto estruturas de traços implementam restrições relacionadas à concordância e à valência, bloqueando combinações que as violam. Como veremos na próxima seção, o GF faz distinção análoga.

O fragmento (19) limita-se à primeira dimensão. Desse modo, hipergera, produzindo muitas sentenças agramaticais como (20)-(23). Em (20), a flexão do verbo auxiliar conflita com a do principal, ver correções em (25) e (26). O exemplo (21) também admite duas correções alternativas: ou adjungimos um Pron2 a *pusé* para obter (26) ou tratamos esse constituinte como modificador do NP nucleado por *igara*, satisfazendo as exigências valenciais de *iku* por meio de um outro núcleo A, ver (27). O exemplo (22), corrigido em (28), é agramatical porque o substantivo *timbiú* e o verbo inativo *raku* são multiformes,[12] incompatíveis com o prefixo *i*, exigindo, em vez disso, o alomorfe assilábico *s*-, ver (1), realizado como *x* antes de *i*.[13] Finalmente, (23) e (24) não constituem sentenças por carecerem de um sujeito (CRUZ, 2011, p. 142), comparem-se (12) e (29).

(20) *se pusé u-iku
1s.INACT pesado 3s.ACT-estar

(21) *igara pusé u-iku
canoa pesado 3s.ACT-estar

(22) *i timbiú i raku
3s.INACT comida 3s.INACT quente

(23) *t-tendaua upé
IMPR-comunidade em

[12] A esse respeito, seguimos Cruz (2011, p. 137-138). Navarro (2011) não considera *timbiú* multiforme.

[13] Para Cruz (2011, p. 137), os alomorfes *s*- e *x*- estão em distribuição complementar. Ao contrário, Navarro (2011) apresenta *ximiriku* 'esposa dele' como variante de *simiriku*, forma que não ocorre em nenhum exemplo de Cruz (2011).

(24) *puranga
bonito

(25) se pusé a-iku
1s.INACT pesado 1s.ACT-estar
'estou pesado'

(26) igara i pusé u-iku
canoa 1s.INACT pesado 3s.ACT-estar
'a canoa está pesada'

(27) igara pusé sepiasuíma u-iku
canoa pesado barato 3s.ACT-estar
'a canoa está barata'

(28) x-imbiú[14] s-aku
3s.INACT-comida 3s.INACT-quente
'a comida dele é quente'

(29) aé puranga
ela bonito
'ela é bonita'

Seguindo a progressão do Quadro 2, expandimos o fragmento de (19) nas duas dimensões referidas. Na primeira, modificamos as regras sintagmáticas de modo a também gerar exemplos do tipo de (12)-(18), que envolvem os tópicos 1.IV, 2.I, 2.II e 9.III, e exemplos análogos do tipo de (11), em que o complemento locativo é expresso por um advérbio.

Desses quatro tópicos, o primeiro envolve um grau considerável de complexidade, mas os demais são de implementação bastante simples, abstraindo da questão das posposições multiformes, que abordaremos no final desta seção. A negação sentencial é feita em nheengatu por meio do clítico *ti*, que se adjunge à primeira posição do rema (CRUZ, 2011,

[14] Ou *s-imbiú*, se considerarmos livre a variação entre <s> e <x>, como sugere Navarro (2011), ver (175).

p. 406), ocupada por um verbo ou por um sintagma posposicional em (13)-(16).[15]

Tanto prefixos inativos quanto NPs plenos podem realizar o argumento interno de posposições, ver (12)-(18). Para gerar esse tipo de sintagma, incluímos (30) e (31) na gramática de (19), onde P designa tanto uma posposição quanto uma locução posposicional, compare (16) e (17). Essas regras produzem também estruturas agramaticais, uma vez que não modelam a checagem de caso e a alomorfia entre *i* e *s*- na realização de 3s.INACT, que abordaremos mais adiante.

(30) PP → NP P

(31) NP → Pron2

A construção possessiva expressa uma relação de posse *lato sensu* entre dois participantes, um possuidor (PSOR, do inglês *Possessor*) e uma entidade possuída (PSUM, do inglês *Possessum*). Essa relação abstrata envolve uma série de relações mais específicas como posse (*ownership*), parte-todo, parentesco etc. (KARVOVSKAYA, 2018).

Analogamente a outras línguas ameríndias, o nheengatu distingue entre nomes relativos e nomes autônomos (CRUZ, 2011). Enquanto estes são intrinsecamente monovalentes, licenciando, em caráter facultativo, um complemento genitivo, aqueles são intrinsecamente divalentes, sendo obrigatória a realização do argumento interno, mesmo quando recuperável contextualmente, como em (32), extraído de Cruz (2011, p. 155).[16]

(32) u-riku i paia
3s.ACT-ter 3s.INACT pai
'Ele tem pai.'

[15] Navarro (2011, p. 18) consigna também a forma *niti*, que, numa contagem por meio da ferramenta *grep*, ocorre 98 vezes no texto, contra 28 da forma clítica *ti*. Um dos pareceristas observa que *niti* ocorre em documentos do século XIX, mas não nos dados de Cruz (2011). No nheengatu do rio Negro, conforme Casasnovas (2006, p. 49), a negação do verbo é feita apenas por meio de *ti* (*te* no imperativo). No entanto, implementamos também a forma *niti* para possibilitar a análise de textos mais antigos.

[16] Em todos os exemplos extraídos da literatura, adaptamos a ortografia e as glosas às utilizadas neste artigo. As traduções são dos autores citados, salvo indicação contrária.

A valência dos nomes em nheengatu constitui um ponto que mereceria um aprofundamento, diante de contradições internas das abordagens de Cruz (2011) e Navarro (2011). Embora aquela classifique todos os nomes multiformes como relativos, apresenta diversos exemplos de *uka* 'casa' e *timbiú* 'comida' na forma absoluta, sem complemento genitivo, ver exemplos análogos em (16) e (3). O segundo autor, por sua vez, trata os nomes de parentesco e as partes do corpo como necessariamente possuíveis, portanto, incompatíveis com a forma absoluta. No entanto, apresenta, de diversos desses nomes que são multiformes, uma forma absoluta, como *tamunha* 'avô'.

Uma comparação superficial entre os NPs na posição de sujeito em (14) e (15) com os exemplos equivalentes em (33) e (34) sugere que a construção possessiva do nheengatu organiza-se da mesma forma que a construção com genitivo saxão do inglês, ressalvada a inexistência, naquela língua, de marca morfológica correspondente:

(33) the boy's community

(34) Mary's house

Há, porém, diferenças fundamentais na realização dos participantes nos dois tipos de construção. Para mostrar isso, recorremos à hipótese DP, proposta em sintaxe gerativa no quadro da teoria X-barra. Na construção do inglês, o PSOR é um DP completo, gerado na posição de especificador do DP nucleado pelo determinante *'s*, enquanto o PSUM realiza-se como NP complemento desse núcleo (CARNIE, 2002, p. 145-146), conforme (35). Desse modo, o PSUM não admite determinação por meio de um demonstrativo como no exemplo do nheengatu em (17), pelo que os exemplos análogos (36) e (37) são agramaticais. Na construção genitiva normanda, porém, não há essa restrição, ver (38).[17]

(35) [DP [DP the boy] [D' [D 's] [NP community]]]

(36) *Peter's that hook

(37) *that the boy's hook

(38) that hook of the boy

[17] Sobre os dois genitivos do inglês, ver Comrie (1983, p. 85).

Inversamente ao que se observa no genitivo saxão, na construção genitiva do nheengatu o PSUM é um DP completo, ao passo que o PSOR corresponde a um NP, pelo que (40) não constitui tradução de (39), mas de (38).

(39) that boy's hook

(40) nhaã kurumĩ pindá
DEM.DIST menino anzol
'aquele anzol do menino'

Antes de detalhar essa análise, formalizamos, por meio do fragmento de gramática (41)-(48), o esquema que Cruz (2011, p. 282), no quadro da abordagem tradicional do sintagma nominal, propõe para esse constituinte em nheengatu. Três tipos de constituintes podem anteceder um núcleo nominal NOME na ordem especificada: (i) quantificadores (Quant), (ii) determinantes (Det), incluindo demonstrativos (Dem), indefinidos (Indf) e numerais (Num), e (iii) nomes (N) ou prefixos inativos (Pron2) na função de complemento nominal (CN), ver (7) e (14). NOME, por sua vez, segundo Cruz (2011, p. 282), é um "nome dêitico", i.e., um Pron1, ou um "nome substantivo", ou seja, um N.

(41) NP → (Quant) (Det) (CN) NOME

(42) Quant → "panhẽ"…

(43) Det → Dem | Indf | Num

(44) Dem → "nhaã"…

(45) Num → …

(46) CN → N | Pron2

(47) N → "igara"|"kambi"…

(48) NOME → N | Pron1

Complementada com regras de inserção de Pron1 e Pron2, essa minigramática gera qualquer dos sintagmas nominais dos exemplos anteriores, assim como estruturas mais complexas como (49), em que se realizam todas as categorias facultativas de (41). No entanto, gera também estruturas agramaticais como (50) e (51), em que um NOME realizado como pronome pessoal rege um CN.[18]

(49) [$_{Quant}$ panhẽ] [$_{Det}$ nhaã] [$_{CN}$ kurumĩ] [$_{NOME}$ igara-itá]
todo DEM.DIST menino canoa-PL
'todas aquelas canoas do menino'

(50) *[$_{CN}$ se] [$_{NOME}$ indé]
1s.INACT tu

(51) *[$_{CN}$ kurumĩ] [$_{NOME}$ penhẽ]
menino vós

Por outro lado, (41)-(48) não geram os exemplos (52)-(54) de Cruz (2011, p. 258, 184, 176), que contrariam a sua própria abordagem, como evidenciamos por meio da análise estrutural dos dois primeiros. Em ambos os exemplos, o CN de *nheenga* constitui-se de um nome que, por sua vez, possui um prefixo inativo como CN, ao passo que, no segundo, o próprio núcleo *nheenga* recebe um desses prefixos, configurações não contempladas pelo esquema estrutural formulado por Cruz (2011). Retomaremos (54) mais adiante.

(52) [$_{CN}$ [$_{CN}$ ne] [$_{NOME}$ mena]] [$_{NOME}$ nheenga]]
2s.INACT marido língua
'língua do teu marido'

(53) [$_{CN}$ [$_{CN}$ ne] [$_{NOME}$ kiuíra]] [$_{CN}$ ta [$_{NOME}$ nheenga]]
2s.INACT irmão 3p.INACT fala
'o conselho dos teus irmãos'[19]

---

[18] Karvovskaya (2018, p. 3) chama atenção para o fato de que o PSUM "geralmente não é um elemento pronominal: **John's she*." Esse tipo de construção somente é possível com substantivos homônimos de pronomes pessoais: *O eu do João*.

[19] Uma tradução mais próxima da estrutura original seria talvez 'o conselho deles de irmão teu'.

(54) aitenhaã paa sukuriu pedasu-itá kuera
DEM.DIST QUOT sucuri pedaço-PL resquício
'Diz que aquilo eram pedaços de sucuri.'

No quadro do debate sobre recursividade nas línguas ameríndias, Leandro e Amaral (2014) defendem a existência de genitivos recursivos em wapichana, da família arawak. O exemplo (54) com dois genitivos nominais, extraído de Cruz (2011), sugere que esse fenômeno também ocorre em nheengatu. Desse modo, reformulamos (41)-(48), com base na hipótese DP, inicialmente como (55)-(59). Conforme (56), o núcleo D pode não se realizar foneticamente, pelo que o DP assume, via de regra, uma interpretação definida.[20]

(55) DP→(QP) D'

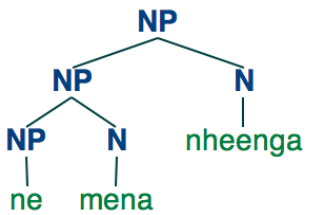
(56) D'→(D) NP

(57) NP→(NP) N

(58) N→"igara"|"kambi"|"nheenga"|"mena"|…

(59) NP→"se"|"ne"|"i"|…

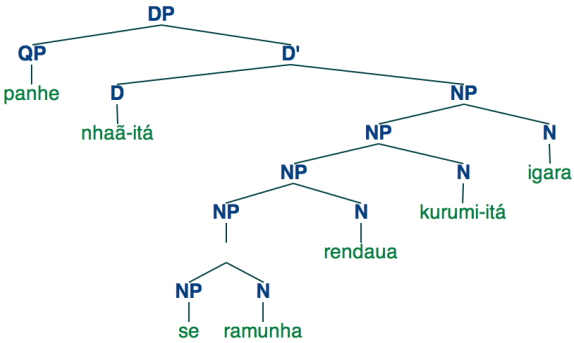
FIGURA 3 – Representação arbórea do NP de (52) conforme (55)-(59)



Fonte: Elaboração própria.

[20] Cruz (2011, p. 265) mostra que, em narrativas, DPs com D zero assumem interpretação indefinida quando não se referem ao tópico discursivo, marcado, apenas na primeira menção, pelo D indefinido *yepé* 'um'.

FIGURA 4 – Representação arbórea do NP de (60) conforme (55)-(59)[21]



Fonte: Elaboração própria.

Além de (52), essa gramática gera, entre infinitos outros, exemplos como (60), produzindo representações recursivas como as da Figura 3 e da Figura 4, onde NPs funcionam como complementos nominais de outros NPs.

(60) panhẽ nhaã-itá se r-amunha r-endaua kurumi-itá igara
todo DEM.DIST-PL 1s.INACT RLL-avô RLL-comunidade menino-PL canoa
'todas aquelas canoas dos meninos da comunidade do meu avô'

A gramática de (55)-(59), contudo, não gera (53), pois analisa o PSUM *ta nheenga* como NP, não admitindo, portanto, o complemento *ne kiuíra*. Substituímos, então, (58) e (59) por (61)-(63). Nessa reformulação, nomes com prefixos inativos são formados no léxico, em consonância com o Princípio da Integridade Lexical (BRESNAN, 2001). Conforme (61), um N é formado pela concatenação opcional de um prefixo inativo a um radical nominal. Esse prefixo satura o lugar vazio de complemento dos nomes relativos. Nomes em que o prefixo de 3ª pessoa não é *i*, mas *s*-, compare (5)-(7), parecem corroborar essa abordagem.

(61) N →(Prf) N_{Stem}

(62) Prf→"se" |"ne" |"i" |"s"…

(63) N_{Stem}→"nheenga"|"kiuíra"|"mena"…

[21] O til sobre *e* e *i* não é suportado pelo programa que utilizamos para gerar esse gráfico.

Uma série de questões, contudo, resta investigar por meio de *corpora* mais extensos ou de experimento nos moldes do proposto por Leandro e Amaral (2014). Primeiro, se a construção de (54) é produtiva em nheengatu, dada a gramaticalização de *kuera* (CRUZ, 2011). Segundo, se um número maior de encaixes é licenciado. Finalmente, se um outro tipo de construção é utilizado para expressar relações de posse mais complexas, por exemplo, orações relativas ou sintagmas posposicionais.[22]

Independentemente dessas questões, a Figura 4 permite visualizar as características fundamentais da construção possessiva genitiva do nheengatu: (i) quantificadores e determinantes têm escopo sobre o NP mais alto, que representa o PSUM; (ii) o PSOR, realizado pelo NP complemento (genitivo) do N núcleo do PSUM, não é passível de quantificação nem determinação; (iii) admite, porém, adjunção da partícula de plural *itá*, como evidencia o exemplo (64) de Cruz (2011, p. 209).

(64) aitenhaã ia-seruka kariua-itá nheenga rupi Martim Pescador
DEM.DIST 1p.ACT-chamar não.indígena-PL palavra PERL Martim Pescador
'Aquele lá, chamamos pela língua dos brancos de Martim Pescador.'

As propriedades (i) e (ii) são reconstruções da análise de Cruz (2011), dirimindo uma aparente incoerência dessa abordagem. De fato, por um lado, analisa o complemento nominal genitivo em exemplos do tipo de (49) como N. Por outro, afirma que a partícula de plural incide sobre o NP com núcleo contável (CRUZ, 2011, p. 164). Logo, se um N funcionando como complemento nominal pode estar sob o escopo dessa partícula, deve constituir um NP, exatamente como propomos. A análise do PSOR como NP prevê, também, que possa ser modificado por um AP, o que de fato se verifica em (65).

(65) piranga tuí kue(ra) nhaã i iuru=pe (CRUZ, 2011, p. 139)
vermelho sangue resquício DEM.DIST 3s.INACT boca=LOC
'O resquício de sangue vermelho ficou na boca dele.'

[22] Na variedade descrita por Sympson (1877, p. 66), ocorre, ao lado do genitivo pré-nominal do tipo de (14), construção com possuidor pós-nominal introduzido por posposição: *kisaua Mandu ressé* 'rede de Manuel'.

Seguindo Navarro (2011, p. 12), para quem "adjetivos qualificativos", i.e., verbos inativos na classificação de Cruz (2011), podem anteceder ou subseguir o núcleo nominal, formalizamos a modificação adjetival do NP por meio de (66), reformulação da terceira regra de (19).

(66) NP→(AP) N (AP)

Ao contrário, Cruz (2011, p. 194) nega não só a existência de adjetivos em nheengatu,[23] como também a possibilidade de verbos inativos, sem a intermediação do relativizador *waa*, funcionarem como modificadores nominais, contrariamente ao que (65) evidencia e Navarro (2011) preconiza, no que é corroborado por exemplos como (67) de Casasnovas (2006, p. 53) (adaptamos a ortografia e incluímos as glosas, mas mantivemos a tradução original).

(67) A-putari apukuitaua miri mirá tauá suiuara.
1s.ACT-querer remo pequeno pau amarelo de
'Quero um remo pequeno feito de pau amarelo.'

Voltando à comparação entre a construção possessiva com *'s* do inglês e a construção genitiva do nheengatu. Na primeira, conforme a análise no âmbito da hipótese DP, tanto o PSOR quanto o PSUM são DPs, sendo que esse último tem *'s* como núcleo D. Na segunda língua, o PSUM é um DP que admite como núcleo qualquer membro da categoria funcional D (respeitadas restrições semânticas), ao passo que o PSOR é um NP. Há uma diferença fundamental entre DPs e NPs: apenas os primeiros são referenciais (BERNSTEIN, 2003). Isso se reflete no contraste entre os modificadores genitivos de (68) e (69). No primeiro exemplo, a categoria PP domina um DP que introduz um participante da situação referida pelo nome modificado, ao passo que, no segundo, domina um NP, não introduzindo participante, mas funcionando apenas como classificador do nome modificado, à semelhança de um adjetivo denotativo de qualidade (FÁBREGAS, 2017).

[23] Como salienta um dos pareceristas anônimos, Cruz (2011, p. 130) classifica palavras como *puku* 'comprido' como verbos estativos, e não adjetivos, porque, analogamente aos verbos dinâmicos, recebem um sufixo nominalizador para funcionarem como argumentos.

(68) este retrato da criança

(69) este retrato de criança

Dada a interpretação do complemento genitivo em (49) como uma expressão referencial, não haveria um núcleo D foneticamente vazio dominando esse NP? Ao nosso ver, essa interpretação não é determinada na sintaxe, mas no nível discursivo. De fato, como evidenciam exemplos como (70) de Cruz (2011, p. 260), em consonância com exemplos análogos de Navarro (2011), noutros contextos o genitivo não introduz participante, funcionando apenas como classificador. Dada a sua natureza de NP, o genitivo não é especificado quanto à determinação, pelo que (71) traduz-se como (68) ou (69), dependendo da situação comunicativa concreta.

(70) mirá rakanga
árvore galho
'galho de árvore'

(71) kuá taína r-angaua
DEM.PROX criança RLL-retrato

Na segunda dimensão da formalização gramatical, modelamos as relações de compatibilidade sintagmática entre os diferentes itens lexicais, de modo a excluir exemplos agramaticais tais como exemplificados em (20)-(22). Para tanto, implementamos restrições em diferentes domínios: (i) concordância verbo-nominal, (ii) marcação de plural no DP, (ii) checagem de caso e (iii) seleção dos alomorfes de substantivos, verbos inativos e posposições multiformes.

Vejamos em detalhe cada um desses fenômenos no âmbito do recorte definido. Em nheengatu, os verbos subdividem-se em dois grupos principais: flexionáveis e não-flexionáveis. Os primeiros classificam-se em ativos e inativos, recebendo prefixos das séries ativa ou inativa, respectivamente. Esses prefixos expressam a concordância número-pessoal com o sujeito, tanto sob a forma de um DP pleno ou pronome livre quanto de um pronome nulo. Verbos não-flexionáveis do mesmo modo que construções sem verbo não admitem sujeito nulo, licenciado pela flexão verbal, ver (24) e (23).[24]

[24] Utilizamos a noção de pronome nulo, simbolizado como *pro*, e sujeito nulo da gramática gerativa (CARNIE, 2002, p. 273).

Na construção de verbo inativo flexionável com o auxiliar *iku*, ambos se flexionam, marcando a concordância com o sujeito. Na 3ª pessoa da série ativa, porém, não há, propriamente, concordância de número na variedade descrita por Navarro (2011): o prefixo é *u-* tanto no singular quanto no plural, como em Casasnovas (2006, p. 27-28). Exemplos de Casasnovas (2006, p. 34):

(72) kuá apigá u-puraki puranga
DEM.PROX homem 3s.ACT-trabalhar bom
'este homem trabalha bem'

(73) musapíri kumurĩ-itá u-musarai u-iku
três menino-PL 3p.ACT-brincar 3p.ACT-estar
'três meninos estão brincando'

Essa era a situação no século XIX, conforme Cruz (2015), que mostra que a concordância de número na 3ª pessoa da série ativa constitui um fenômeno mais recente. Para Cruz (2011, p. 133), ao prefixo *u-* no singular opõe-se *ta(u)-* ~ *tu-* no plural. Segundo Cruz (2015, p. 430-31), no nheengatu do século XXI, a combinação do clítico *ta=* (pronome de 3ª pessoal do plural) e do prefixo *u-* (3ª pessoa da série ativa) é característica da fala dos mais velhos, cedendo lugar, na fala dos mais jovens, a *tu-* no dialeto do Xié e a *ta-* no dialeto do Negro e do Içana. Tanto Navarro (2011, p. 89) quanto Casasnovas (2006, p. 28) contemplam a marcação da 3ª pessoa do plural por meio de *ta=u-*.

O plural do DP é marcado formalmente por meio da partícula *itá*, sufixada ao núcleo nominal ou ao núcleo D, compare (2) e (4). Embora Navarro (2011, p. 25-26), na exposição dos demonstrativos (tópico L3.VI do Quadro 2), se limite ao segundo caso, ele abona o primeiro com o exemplo (74) (NAVARRO, 2011, p. 48).

(74) Re-rasu kuá maniaka-itá memũitendaua kiti.
2s.ACT-levar DEM.PROX mandioca-PL cozinha para
'Leve estas mandiocas para a cozinha.'

Para Navarro (2011, p. 19), somente se usa a partícula *-itá* quando absolutamente indispensável, como em (75), onde, devido à ambiguidade da flexão verbal, a sua omissão resultaria na leitura de (76). Segundo ele, a partícula é omitida quando outro elemento indica pluralidade, como em

mukũi apigaua ‘dois homens’. Na língua falada no alto rio Negro, ocorre, com quantificadores discretos, tanto a marcação redundante quanto a não redundante, como nos exemplos (77) e (78) extraídos de Cruz (2011, p. 269-270). Casasnovas (2006, p. 34-35) apresenta exemplos análogos.

(75) kunhã-itá puranga u-iku
mulher-PL bonito 3p.ACT-estar
‘as mulheres estão bonitas’

(76) kunhã puranga u-iku
mulher-SG bonito 3s.ACT-estar
‘a mulher está bonita’

(77) ia-uasému mukũi pesoa-itá
1p.ACT-encontrar dois pessoa-PL
‘Encontramos duas pessoas.’

(78) aikué musapíri pesoa u-iku uaá ape
há dois pessoa 3s.ACT-estar[25] REL lá
‘Havia três pessoas, que estavam lá.’

Não há, portanto, concordância de número entre os constituintes do DP em nheengatu: um quantificador plural admite tanto um NP com *-itá* quanto um NP sem essa partícula, como evidenciam, respectivamente, os exemplos (77) e (78) de Cruz (2011). A GrammYEP ainda não contempla DPs com quantificadores nem numerais como nesses dois exemplos, apenas com núcleo D demonstrativo. Nesse caso, estipulamos que um DP é plural se o núcleo D (demonstrativo) ou o núcleo N é marcado com *itá*. Admitindo que a marcação do plural é um fenômeno lexical, núcleos D demonstrativos e substantivos constituem paradigmas que variam conforme o número, a forma de plural diferindo da de singular por meio do sufixo *itá*. Na sintaxe, um demonstrativo no plural seleciona um substantivo no singular, ao passo que um demonstrativo no singular não impõe restrição sobre o número do NP complemento.

[25] Cruz (2011) analisa essa forma como 3ª pessoa do singular, pois, para ela, como vimos acima, a flexão de 3ª do plural é *ta(u)-*.

Essa análise da partícula *itá* é uma reconstrução das abordagens de Cruz (2011) e Navarro (2011), falseável por dados de corpora ou julgamentos de aceitabilidade corroborando exemplos do tipo de (79), que, por ora, consideramos agramaticais. De fato, não encontramos, nesses dois trabalhos, DP com iteração da partícula *-itá* como em (79). Cruz (2011, p. 377) parece negar a possibilidade desse tipo de construção ao caracterizar *-itá* como "partícula independente" com escopo sobre o sintagma nominal (DP na nossa abordagem) que pode adjungir-se tanto ao núcleo D ou ao núcleo N.

(79) *kuá-itá kunhã-itá

Aplicando a análise do DP e do NP formalizada em (55)-(57) e (61) a exemplos como (5)-(12) e (80)-(84), as seguintes generalizações emergem sobre a distribuição de pronomes pessoais livres (Pron1) e prefixos inativos (Pron2): os primeiros ocupam a posição de sujeito de verbos ativos e inativos e complemento (objeto direto) de verbos ativos, enquanto os segundos funcionam como marcas de concordância de verbos inativos flexionáveis e complementos (i.e., argumentos internos) de nomes e posposições.

(80) aé i pusé
ele 1s.INACT pesado
'ela é pesada'

(81) (*kunhã) s-uka taua upé
mulher 3s.INACT-casa cidade em

(82) (*kunhã) i igara paranã upé
mulher 3s.INACT canoa rio em

(83) ixé a-iku (*kunhã) s-uaki

(84) *a-iku aé irũmu
1s.ACT-estar ele.ACC com

(85) aé s-aku
ele 1s-INACT quente
'ela é quente'

(86) ixé a-maã aé
eu.NOM 1s.ACT-ver ela.ACC
'eu a vejo'

O Quadro 3 evidencia que, em posições argumentais de V ou P, os itens das colunas 2 e 3 estão em distribuição complementar. Os elementos da coluna 4, por sua vez, como vimos em (61), saturam, no nível sublexical, a posição de complemento de N_{Stem}. Estendendo essa análise aos elementos da coluna 3, reformulamos (30) e (31) como (87). Tanto no caso de N_{Stem} quanto de P_{Stem}, a saturação da posição de complemento no nível sublexical impede que um NP ou DP pleno venha a ocupar essa mesma posição na sintaxe, ver (81) e (82).[26] As marcas de concordância de sujeito, ao contrário, coocorrem com DP pleno ou Pron1, ver, por exemplo, (8), (80) e (85). Desse modo, em relação aos itens das colunas 3 e 4, as flexões verbais inativas constituem um paradigma à parte de formas homônimas.

(87) P→(Prf) P_{Stem}
PP→(DP) P

QUADRO 3 – Distribuição de Pron1 e Pron2 no singular em posições argumentais

Pessoa	**DP argumento externo ou interno de V ou externo de P (Pron1)**	**Argumento interno de P_{Stem} (Pron2)**	**Argumento interno de N_{Stem} (Pron2)**
1s	ixé	se	se
2s	indé	ne	ne
3s	aé	i|s-	i|s-

Fonte: Elaboração própria.

Para modelar a distribuição dos elementos do Quadro 3, recorremos à noção de caso: Pron1 satura posições argumentais associadas aos casos nominativo e acusativo, ao passo que Pron2 realiza o complemento genitivo de nomes e posposições, compare (7), (18) e

[26] Em (53), o primeiro NP não ocupa a mesma posição que o prefixo do segundo, dada a discrepância de número entre esses constituintes.

(86). Pron1 manifesta, portanto, sincretismo sistemático entre nominativo e acusativo, fenômeno bastante comum nas línguas do mundo (ZOMPÌ, 2017). DPs e NPs plenos não exibem caso morfológico.

Tanto os nomes relativos quanto os autônomos admitem um complemento genitivo, assim como as posposições de modo geral. No nosso fragmento, limitamo-nos às posposições locativas *upé* 'em', *pupé* 'dentro de' e *ruaki* (*suaki*) 'perto de', à comitativa *irũmu* e à locução pospositiva *kuara upé* 'dentro de'. Segundo Cruz (2011, p. 199), a primeira é incompatível com prefixos inativos, analogamente à posposição *arama* ou *(a)rã* do dativo intralocutivo, que exige como complemento um pronome livre de 1ª ou 2ª pessoa, ver (88). A posposição *supé* do dativo extralocutivo, porém, formalmente relacionada a *upé*, admite prefixo inativo de 3ª pessoa (CRUZ, 2011, p. 201).[27] Ao contrário, para Navarro (2011), *arama* é a única posposição que não admite pronome pessoal de segunda classe. Apesar de Cruz e Navarro não apresentarem qualquer exemplo de *upé* regendo pronome (seja de primeira, seja de segunda classe), seguimos a exposição desse último, não tratando essa posposição como excepcional no que tange à regência de caso. Desse modo, consideramos (89) como gramatical e (90) como agramatical.

(88) re-rúri t-imbiú ixé arama (NAVARRO, 2011, p. 23)
2s.ACT.IMP-trazer IMPR-comida eu.ACC para
'traz-me comida'

(89) tukandira-itá u-iku i upé
tocandira-PL 3s.ACT-estar 3s.INACT em
'as tocandiras estavam nele'

(90) *tukandira-itá uiku aé upé
tocandira-PL 3s.ACT-estar ele em

[27] A abordagem de Cruz (2011) parece contraditória nesse aspecto. Por um lado, apresenta exemplos de *supé* regendo prefixo inativo de 3ª pessoa, ver (304) na p. 201 e (363) na p. 218. Por outro, em mais de uma passagem, explicitamente nega essa possibilidade, por exemplo quando afirma na p. 197 que as posposições locativas do grupo *pé*, entre as quais classifica *supé*, são incompatíveis com prefixos inativos, impossibilidade que confirma na p. 200 e no Quadro 28 na p. 223.

Do ponto de vista da implementação computacional, um dos aspectos mais complexos do nheengatu são as classes de palavras multiformes, que constituem subgrupos de substantivos, verbos inativos e posposições. Distinguem-se pela incompatibilidade com o prefixo silábico *i* do Quadro 1. Em vez disso, selecionam o alomorfe assilábico *s*- (*t*- no caso de uns poucos substantivos).

Os substantivos multiformes apresentam, no caso canônico, representado por *tendaua* em (5), (12) e (15), três formas, que se distinguem, conforme Navarro (2011), pelos prefixos de relação *t*-, *r*- e *s*-, representados nas glosas por IMPR, RLL e INACT, respectivamente.[28] A terceira forma resulta da concatenação com *s*-. A segunda forma ocorre com todos os outros tipos de PSOR, isto é, NP pleno ou prefixo inativo de 3ª pessoa do plural e de 1ª e 2ª pessoas, ver (91). A primeira forma, chamada absoluta, que constitui a forma de citação do substantivo, assinala que o substantivo não realiza PSUM de construção possessiva.

(91) se r-uka nhaã kiá s-apé upé
1s.INACT RLL-casa DEM.DIST sujo 1s.INACT-rua em
'a minha casa é naquela suja rua dele'

O Quadro 4 sistematiza os diferentes tipos de substantivos multiformes descritos por Navarro (2011). Os microgrupos (ii)-(v) resultam de desvios do padrão canônico representado pelo microrupo (i), decorrentes de alomorfias do primeiro ou terceiro prefixos.

QUADRO 4 – Microgrupos de substantivos multiformes

Grupo	Forma absoluta	PSOR ≠ 3s.INACT	PSOR = 3s.INACT
	tendaua 'comunidade'	*rendaua*	*sendaua*
	sangaua 'retrato'	*rangaua*	*sangaua*
	uka 'casa'	*ruka*	*suka*
	pé 'caminho'	*rapé*	*sapé*
	tuí 'sangue'	*ruí*	*tuí*

Fonte: Elaboração própria.

[28] Cf. nota 1.

As posposições e os verbos inativos multiformes exibem comportamento análogo ao dos substantivos multiformes prototípicos, com a diferença de que não possuem a forma em *t*-, uma vez que são sempre empregados relacionalmente, ver (9), (10), (14) e (15).

Os verbos inativos multiformes possuem tema iniciado por vogal (CRUZ, 2011, p. 135). O alomorfe *s* do prefixo de concordância de 3ª pessoa do singular concatena-se diretamente ao tema, ver (1), (6), (10), (85). Todas as demais flexões pessoais se adjungem à forma prefixada por *r*-, ver (2), (9) e (13).

3 O formalismo GF

O termo GF possui três acepções distintas, mas intimamente relacionadas (RANTA, 2010). Em primeiro lugar, designa um formalismo para especificação de gramáticas que constitui, em si mesmo, uma linguagem de programação. Isso significa que uma gramática de uma língua como nheengatu formalizada no GF pode ser não só interpretada por um especialista humano com conhecimento do formalismo, mas também compilada num código binário passível de ser processado por um computador na execução de tarefas como análise (*parsing*) e geração de sentenças, tradução automática etc. Por extensão metonímica, o termo GF designa também o software que permite realizar essas tarefas. Finalmente, na terceira acepção, GF designa uma teoria gramatical, i.e., um modelo da organização e funcionamento das gramáticas de línguas naturais.[29]

De um ponto de vista formal, a sintaxe de uma língua natural consiste no conjunto de regras para formação de unidades complexas (sintagmas) a partir de unidades elementares (palavras) (JURAFSKY; MARTIN, 2009). Essa noção estende-se à morfologia, com a diferença de que as unidades elementares são, tipicamente, morfemas e as maiores, palavras.

[29] Diferentemente de outras teorias gramaticais computacionalmente implementadas como a LFG (BRESNAN, 2001), Ranta (2011) não reivindica qualquer validade tipológica ou plausibilidade psicológica para a GF. A justificação da teoria advém da engenharia de software, que procura otimizar a elaboração e a eficiência de programas de computador. Sobre a relação entre teorização linguística e engenharia da gramática, ver Müller (2015).

O formalismo GF baseia-se na teoria matemática dos tipos, que trata da combinação de termos elementares de determinados tipos para formar termos complexos de outros tipos. Enquanto linguagem de programação, o GF enquadra-se no paradigma funcional. Desse modo, a construção dos tipos dá-se por meio da aplicação de funções sobre zero, um ou mais argumentos.

A característica distintiva do GF é a fatoração da combinatória de tipos em dois componentes, a sintaxe abstrata e a sintaxe concreta. No primeiro componente, abstrai-se da realização morfossintática dos elementos envolvidos, que é especificada apenas no segundo.

Na sintaxe abstrata da ITALENG, Ranta (2011) propõe os tipos Comentário, Item, Classe (*Kind*) e Qualidade, que se combinam por meio das funções em (92)-(97). Comentário é o tipo de sentenças declarativas do tipo de (1), (2) etc., em que se atribui uma qualidade a uma entidade (ou conjunto de entidades) de uma determinada classe.

(92) Pred: Item –> Quality –> Comment;

(93) This, That, These, Those: Kind –> Item;

(94) Mod: Quality –> Kind –> Kind;

(95) Very: Quality –> Quality;

(96) Food, Milk, Fish: Kind;

(97) Warm, Cheap, Expensive, Red: Quality;

Nessa notação, os termos à esquerda dos dois pontos são os nomes das funções e os tipos à direita separados por setas as caracterizam em termos dos tipos de argumentos e do valor que produzem, de tal modo que o último (ou o único) desses elementos é o tipo do valor da função. Por exemplo, a função *Pred* produz, a partir de um Item e de uma Qualidade, um Comentário, que é o tipo mais complexo, não entrando na composição de nenhum outro tipo.

A função *Mod* produz uma Classe a partir de uma Classe e uma Qualidade. Por exemplo, em (4), a classe designada por *peixes vermelhos* resulta da aplicação de *Mod* sobre a Qualidade e a Classe designadas pelo adjetivo e substantivo, respectivamente. O tipo Item resulta da aplicação, a uma Classe, de uma das funções nomeadas pelos demonstrativos do

inglês *This* ‘este’, *That* ‘aquele’, *These* ‘estes’ e *Those* ‘aqueles’. Em outras palavras, essas funções extraem indivíduos ou grupos de indivíduos das Classes designadas pelos seus argumentos. Por exemplo, em (4), a aplicação da função *These* à Classe *peixes vermelhos* resulta no Item *estes peixes vermelhos*. O Comentário expresso por (3) exemplifica a função *Very* ‘muito’, que, aplicada à Qualidade *caro*, gera a Qualidade *muito caro*.

Em (96) e (97), temos funções sem argumentos que produzem as Classes e as Qualidades correspondentes aos conceitos básicos designados, respectivamente, pelos substantivos e adjetivos do inglês que nomeiam essas funções. Por exemplo, a função *Fish* produz a Classe correspondente ao conceito de peixe, enquanto a função *Red* gera a Qualidade correspondente ao conceito de vermelho.

Com base nessa sintaxe abstrata, Ranta (2011) define, na ITALENG, duas sintaxes concretas: uma para o inglês e outra para o italiano. Cada um desses módulos especifica como os diferentes tipos e funções da sintaxe abstrata são *linearizados*, ou seja, expressos nessas duas línguas.

Para especificar as propriedades formais das categorias que realizam os diferentes tipos da sintaxe abstrata, utilizam-se *registros*, uma estrutura de dados que permite organizar em campos as diferentes propriedades dessas categorias. Exemplifiquemos. Em inglês, a função *Red* lineariza-se como (98), ou seja, um registro cujo único campo *s* tem como valor “red”, que constitui uma cadeia de caracteres (tipo *Str*, do inglês *string*, doravante Forma). Em português, a função *Fish* lineariza-se por meio do registro (99), que possui dois campos, *s* e *g*, para representar o paradigma flexional e o gênero do substantivo *peixe*, respectivamente. O valor do primeiro campo é uma tabela bidimensional do tipo Número=>Forma (QUADRO 5), que associa, a cada número gramatical, a cadeia de caracteres correspondente. O valor do segundo campo indica o gênero masculino. Os traços gramaticais gênero, número, caso etc. precisam ser definidos por meio de parâmetros, exemplificados em (100) e (101).

(98) Red={s=“red”};

(99) Fish={s=table{Sg=>“peixe”; Pl=>“peixes”}; g=Masc};

(100) Gender=Masc|Fem ;

(101) Number=Sg|Pl ;

QUADRO 5 – Paradigma do lexema *peixe*

Número	**Forma**
Sg	“peixe”
Pl	“peixes”

Fonte: Elaboração própria.

Diferentemente de línguas desprovidas de flexão adjetival como o inglês, *Red* lineariza-se em português como (102), cujo único campo tem como valor uma tabela tridimensional do tipo Gênero=>Número=>Forma. Para extrair um membro de um paradigma *q* desse tipo, GF utiliza a sintaxe *q.s!g!n*, onde *s* é o valor do campo *s* de *q*, o qual constitui uma tabela da qual se extrai, por meio da aplicação sucessiva do operador “!”, a forma de gênero *g* e número *n*.

(102) Red={s=table{Masc=>table{Sg=>“vermelho”; Pl=>“vermelhos”};
table{Fem=>table{Sg=>“vermelha”; Pl=>“vermelhas”}};

A exemplo das linguagens de programação comuns, como Java, C ou Python, GF permite definir operações parametrizadas (funções) para execução de tarefas repetidas. Esse recurso torna desnecessário especificar, no léxico, todas as formas de cada adjetivo do português, uma vez que se flexionam conforme padrões predizíveis a partir do lema. Desse modo, definimos a operação mkAdj, que gera os paradigmas dos diferentes tipos de adjetivo em português, permitindo simplificar (102) como (103). Essa operação utiliza o operador de concatenação de cadeias “+”, chamado “operador de colagem” (*glue operator*), que permite formar palavras a partir de unidades menores, no caso em tela, radical adjetival e flexão de gênero e número. GF permite modelar não só processos morfológicos concatenativos, mas também não concatenativos, como a metafonia. Operações análogas a (103) podem ser definidas para gerar todos os paradigmas de todas as classes de palavras flexionáveis.

(103) Red=mkAdj “vermelho”;

A linearização dos tipos complexos da sintaxe abstrata é especificada de forma análoga, utilizando registros para representar as informações morfossintáticas dos constituintes envolvidos e operações

para manipular esses registros, de modo a dar conta de fenômenos como concordância, regência, variações na ordem dos constituintes etc. Vejamos alguns exemplos. Em português, o tipo Item possui o tipo de linearização de (104), onde o valor do campo *s* é uma tabela do tipo Caso=>Forma. A especificação de caso é necessária porque elementos do tipo Item são realizados não só por sintagmas nominais com núcleo substantival, mas também por pronomes pessoais, que variam em caso conforme a função sintática exercida. Os outros três campos especificam as propriedades inerentes do sintagma nominal. Em (105), temos o tipo de lexicalização de Itens em inglês, que difere de (104) apenas pela ausência de especificação de gênero, que não participa de processos de concordância nessa língua.

(104) {s: Case=>Str; n: Number; p: Person; g: Gender};

(105) {s: Case=>Str; n: Number; p: Person};

As definições de linearização de funções que combinam dois ou mais tipos especificam não somente a ordem em que se concatenam as suas linearizações, mas também, para tipos expressos por tabelas, as células compatíveis. Por exemplo, no caso de *peixes vermelhos*, é exigida compatibilidade de gênero e número. Para concatenar cadeias que representam formas de paradigmas lexicais, GF utiliza o operador ++, exemplificado na função de linearização de *Very* em inglês definida em (106). Como vimos em (92), a função *Very* aplica-se a uma Qualidade para produzir uma Qualidade. Por exemplo, em nheengatu, de *sepiasu* 'caro' obtém-se *sepiasu retana* 'muito caro', ver (3). Em inglês, o intensificador precede o adjetivo. Desse modo, a função (106) especifica que, em inglês, a cadeia "very" é concatenada ao valor do campo *s* do argumento da função, i.e., à forma do adjetivo, representado pela variável *a*.

(106) Very a={s="very"++a.s};

No caso da função *Mod* de (92), é preciso especificar, para o inglês, que Qualidade precede Classe. Em português, de um modo geral, assim como em nheengatu, Qualidade pode tanto preceder quando subseguir Classe. Ao contrário do inglês e do nheengatu, o português exige concordância de gênero e número entre os dois constituintes.

A função de linearização (107) da gramática concreta do português aplica-se aos dois registros que constituem as linearizações de Qualidade e Classe, referidos pelas variáveis *q* e *k*, respectivamente. Na primeira linha, é atribuído, por meio do operador *let*, o valor *k.g* à variável *g* do tipo *Number*. Na segunda linha, o valor de *g* é atribuído ao campo *g* do registro que resulta da combinação dos dois argumentos da função. O valor do campo *s* desse registro é uma tabela do tipo Número=>Forma, análoga à de (102), com a diferença de que se constitui de uma única linha, visto que uma combinação particular de Qualidade e Classe (por ex., *peixes vermelhos*) possui um número específico, representado pela variável *n*. O valor dessa tabela é o resultado da operação *shuffle* aplicada sobre as cadeias que constituem as linearizações de *q* e *k*. Essa operação, que definimos no módulo Oper.gf, gera todas as permutações possíveis entre os seus argumentos. Desse modo, a gramática gera adjetivos tanto pós-nominais quanto pré-nominais (ordem marcada no caso de adjetivos de cores).

(107) Mod q k=let g: Gender=k.g in
{s=\\n=>shuffle (q.s!n!g) (k.s!n); g=g};

A linguagem de programação GF determina uma estrita separação entre a concatenação de *tokens* (palavras) para formar listas de *tokens* (sintagmas) por meio do operador "++", exemplificado em (106), e a colagem de cadeias para formar *tokens* (palavras individuais ou locuções) por meio do operador "+", utilizado pela operação (103).[30] Este último processo ocorre em tempo de compilação do código-fonte, ao passo que o primeiro ocorre em tempo de execução do código binário. A exigência de todos os *tokens* estarem formados em tempo de execução confere maior eficiência à análise e geração de sentenças e tem uma consequência importante para a implementação linguística: regras de geração de formas representadas por *tokens* individuais, como as formas com prefixo inativo *s-* do nheengatu, precisam ser modeladas como regras lexicais. Não é possível, portanto, na definição da função de linearização de *Pred* numa sintaxe concreta do nheengatu, concatenar *s-* com uma base verbal para formar, por exemplo, *suri*, como em (10). Por outro lado, é possível gerar

[30] Limitamo-nos aqui a uma descrição simplificada, informal dos dois operadores, definidos formalmente em Ranta (2011).

formas do tipo de *se ruri* de (9) ou na sintaxe por meio da concatenação dos *tokens* "se" e "ruri", solução que adotamos, ou no léxico como um único *token* "se ruri" (análogo a uma locução).

4 Descrição da implementação

Uma descrição detalhada das sintaxes concretas do nheengatu, português e inglês e da sintaxe abstrata subjacente às três línguas extrapolaria os limites de um artigo. O leitor familiarizado com o GF pode consultar os códigos-fonte de todos os componentes, disponíveis *on-line* sob licença de uso de software livre e suficientemente autoexplicativos.[31] Desse modo, limitamo-nos aqui a destacar as particularidades mais importantes da implementação.

Um princípio fundamental da programação é a modularização, a divisão de um programa complexo em componentes organicamente articulados. Desse modo, dividimos a sintaxe abstrata nos seguintes componentes: (i) Gra.gf, (ii) Func.gf e (iii) Cont.gf. As sintaxes concretas do nheengatu, português e inglês constituem os módulos GraYrl.gf, GraPor.gf e GraEng.gf, respectivamente, cada um dos quais recorre a operações definidas num módulo Oper.gf comum e nos módulos específicos OperYrl.gf, OperPor.gf e OperEng.gf.

O módulo Gra.gf constitui-se de regras de formação de tipos complexos, como as exemplificadas em (108)-(112). Algumas dessas funções utilizam tipos definidos nos demais componentes, como Polaridade, Locação e Qualidade.

(108) Pred: Polarity –> Item –> State –> Comment;

(109) StageLevelState: Property –> State;

(110) IndLevelState: Property –> State;

(111) mkPropLoc: Location –> Property;

(112) mkPropQual: Quality –> Property;

[31] Disponível em: http://github.com/leoalenc/nheengatu. Acesso em: 27 ago. 2020.

O módulo Func.gf contém funções que definem tipos linearizados por palavras funcionais ou sinsemânticas (BUSSMANN, 2002), ou seja, em GraYrl.gf, o advérbio de intensidade *retana* 'muito', clíticos de negação, posposições, pronomes e determinantes. Por exemplo, as funções de (113), análogas a (93), aplicam-se ao tipo Classe para formar o tipo Não-Dêitico, de que trataremos mais adiante, linearizado sob a forma de um DP pleno. A presente versão da gramática não contempla DPs com quantificadores ou numerais, exemplificados em (49), (77) e (78). *TheSG* e *ThePL* permitem construir DPs definidos no singular e no plural, respectivamente. Em GraYrl.gf, essas duas funções linearizam-se como DPs com artigo zero. Desse modo, exemplos do nheengatu com esse tipo de constituinte serão sempre traduzidos em português e inglês como DPs plenos nucleados por artigo definido. Essa solução, porém, tem um caráter apenas prático, uma vez que, em narrativas em nheengatu, uma interpretação indefinida é, por vezes, preferível.[32] No entanto, a solução teoricamente mais adequada, que seria implementar as duas interpretações possíveis, provocaria um aumento exponencial da ambiguidade de sentenças com esse tipo de DP.[33]

(113) TheSG, ThePL, This, That, These, Those: Kind –> NonDeictic;

O módulo Cont.gf define tipos expressos por palavras lexicais (*content words*) ou autossemânticas (BUSSMANN, 2002), no caso, apenas substantivos e adjetivos, pois verbos plenos não foram ainda incluídos. As cópulas são tratadas como palavras sincategoremáticas, definidas por Ranta (2011, p. 100) como palavras sem representação própria na sintaxe abstrata. Desse modo, são introduzidas na regra de linearização da função *Pred* de (108).

Para a elaboração da sintaxe abstrata, tomamos como ponto de partida os tipos e as funções de (92), extraídos da minigramática do inglês e do italiano de Ranta (2011). Essa gramática, porém, restringe-se a uma fração dos fenômenos gramaticais do Quadro 2. Desse modo,

[32] Cf. nota 20.

[33] Abstraindo de outras fontes de ambiguidade, o número de leituras de uma sentença com *n* DPs com núcleo zero, pressupondo a ambiguidade desse núcleo entre uma interpretação definida e outra indefinida, seria 2^n. Desse modo, uma sentença com quatro DPs desse tipo teria, pelo menos, 16 leituras.

foi necessário criar diversos novos tipos, reformular várias das regras propostas por Ranta (2011) e elaborar novas regras.

Para dar conta da negação, da predicação locativa e da distinção entre nível de fase e nível de indivíduo, substituímos a regra (92) pelas regras de (108)-(112). Conforme (108), um Comentário é constituído de uma Polaridade, um Item e um Estado. O tipo Polaridade, por sua vez, é formado em Func.gf a partir dos tipos *Yes* ou *No*, que codificam polaridade positiva e negativa, respectivamente. Apenas esta última é expressa na linearização.

O tipo Estado constitui-se de uma Propriedade, que, por sua vez, pode ser uma Locação ou uma Qualidade, conforme (111) e (112). As funções (109) e (110) geram estados contingentes e não-contingentes, respectivamente. Qualidades são geradas nos módulos Cont.gf e Func.gf por funções análogas a (95) e (97), respectivamente, Locações no módulo Func.gf pelas funções de (114) e (115). Por exemplo, em (15), a aplicação da função *Near* ao Item designado por *pe taua* 'a cidade de vocês' gera a Locação referida por *pe taua ruaki* 'perto da cidade de vocês'. Em (11), o advérbio *iké* 'aqui' constitui em si mesmo uma Locação, dado que lineariza a função *Here* de (115).

(114) On, With, In, Inside, Near: Item -> Location;

(115) Here, There: Location;

Do ponto de vista da semântica referencial, é indiferente se um Item é linearizado como pronome pessoal ou DP pleno, compare (8) e (7). Ambos os exemplos consistem numa predicação qualificativa que atribui uma qualidade a uma entidade. Pronomes pessoais e DPs plenos, porém, não compartilham exatamente a mesma distribuição, uma vez que os primeiros não realizam PSUM, ver (50) e (51). Desse modo, como meio de evitar a hipergeração, a implementação, em sintaxe abstrata, das construções possessivas e dos sujeitos pronominais, inclusive nulos, distingue, no âmbito dos Itens, por meio das regras (116) e (117), entre os subtipos Dêitico e Não-Dêitico, linearizados como pronomes pessoais e DPs plenos, respectivamente.[34]

[34] Seguimos aqui Lyons (1995, p. 307), que considera dêiticos tanto os pronomes pessoais de 1ª e 2ª pessoas quanto os anafóricos de 3ª.

(116) mkItemDeictic: Deictic -> Item;

(117) mkItemNonDeictic: NonDeictic -> Item;

O tipo Dêitico, por sua vez, é construído por meio das funções de (118), cujos nomes se baseiam nas formas dos pronomes pessoais do inglês. Dada a ambiguidade da forma pronominal *you*, a 3ª pessoa do singular e a do plural são designadas por meio de *YouSG* e *YouPL*, respectivamente, analogamente a *TheSG* e *ThePL* de (113).

(118) He, She, It, They, I, YouSG, YouPL, We: Deictic;

Completam o módulo Gra.gf as regras de (119)-(124), responsáveis pela geração de construções possessivas do tipo de (5), (7), (14), (15), (52) e (53). Nesta versão da gramática, limitamo-nos a um subconjunto das construções possessivas exemplificadas por Cruz (2011), deixando, para uma versão futura, construções recursivas com mais de um núcleo nominal encaixado, como em (54).

(119) Poss: Psor –> NonDeictic –> Item;

(120) Poss_: PossPro –> SimpleKind –> PossKind;

(121) mkPsor: Num –> SimpleKind -> Psor;

(122) mkPsor_: Num –> PossKind –> Psor;

(123) mkKind: SimpleKind –> Kind;

(124) mkKind_: PossKind –> Kind;

Conforme (123) e (124), há dois tipos de Classes: Classe Simples (*SimpleKind*), como *tendaua* ‘comunidade’ em (12), e Classe Possuída (*PossKind*), como *sendaua* ‘comunidade dele’ e *pe taua* ‘cidade de vocês’ em (5) e (15). O primeiro subtipo é gerado no módulo Cont.gf por meio de regras do tipo de (125). O segundo é construído por meio de (120) pela combinação dos tipos Classe Simples e PossPro, este último construído por meio das funções de (126), análogas às de (118).

(125) Food: SimpleKind;

(126) His, Her, Its, Their, My, YourSG, YourPL, Our: PossPro;

Por que pronomes possessivos (prefixos inativos em nheengatu) não foram implementados como linearizações de Itens? A razão é que, diferentemente do inglês e do português, o nheengatu não licencia um DP pleno como PSOR, mas apenas um NP ou um prefixo inativo, conforme (55) e (61). De modo a possibilitar a tradução, entre as três línguas, de construções possessivas do subconjunto referido, evitando, ao mesmo tempo, a hipergeração em nheengatu, criamos o tipo *Psor*, construído pelas regras (121) e (122) a partir do tipo *Num*, por um lado, e do tipo Classe Simples ou Classe Possuída, por outro. O tipo *Num* é construído no módulo Func.gf por meio de (127), linearizando-se como morfema de número, o qual determina a forma de plural ou singular da expressão que realiza o segundo argumento das funções (121) e (122).

(127) PL, SG: Num;

Em nheengatu, conforme (57), o PSOR é um NP, podendo traduzir-se em português tanto por um NP quanto por um DP nucleado por artigo definido, ver (71). Como essa última opção tradutória é a mais comum em Navarro (2011), implementamo-la nas funções de linearização de (119) em português e inglês, seguindo o tratamento conferido ao DP pleno de núcleo zero.

A função (128) define o tipo de linearização de Classe como registro cujo único campo constitui uma tabela do tipo Número=>*NForm*=>Forma. O parâmetro *NForm*, definido em (129), modela a distribuição das formas de substantivos multiformes, tendo dois construtores, Rel e Abs. O primeiro indica que a forma é relativa, ou seja, constitui complemento genitivo, o segundo que se trata da forma absoluta. Como há duas formas no primeiro caso, conforme o tipo de argumento interno, uma para prefixo inativo de 3ª pessoa do singular, a outra para todos os demais casos, o construtor Rel possui como argumento o parâmetro *ArgForm*, definido em (130), cujos valores correspondem, respectivamente, aos dois tipos de argumento interno.

(128) KIND: Type={s: Number=>NForm=>Str};

(129) NForm=NRel ArgForm|NAbs;

(130) ArgForm=SG3|NSG3;

A função (131) define o tipo de linearização de Classe Simples. Constitui uma extensão de (128), diferindo pela inclusão do campo *nc*, que codifica a classe nominal. O parâmetro *NClass*, definido em (132), classifica os substantivos em uniformes e multiformes. Assume um dos dois valores *NCI* e *NCS*, segundo o alomorfe do prefixo inativo de 3ª pessoa do singular exigido, *i* e *s*-, respectivamente, permitindo a linearização correta de (120).

(131) SIMPLEKIND: Type=KIND**{nc: NClass};

(132) NClass=NCI|NCS;

Como evidenciam os exemplos (1)-(17), o tipo Item lineariza-se em nheengatu tanto como DP pleno quanto como pronome livre, pronome nulo ou prefixo inativo, dependendo de uma série de fatores. O primeiro é a função sintática, que determina caso nominativo ou genitivo. O segundo é a classe do predicado regente: apenas verbos inativos flexionáveis licenciam incondicionalmente um sujeito nulo, os outros tipos de predicado o fazem condicionalmente, dependendo do terceiro fator, o nível da predicação. Por exemplo, em (18), o sujeito nulo é licenciado porque, não obstante o predicado regente consistir num PP, a predicação é de nível de fase, marcada, portanto, pelo auxiliar *iku*, compare com (24) e (23). O quarto fator é a classe flexional (*NClass*), que regula a alomorfia entre *i* e *s*-.

O tipo de linearização de Item em nheengatu, definido em (133), que incorpora (134), leva em conta todos esses fatores. Embora desprovida, como em inglês, de especificação de gênero gramatical, essa definição, que se aplica também ao tipo Dêitico, é bem mais complexa do que as correspondentes nas outras duas línguas, compare-se com (104) e (105).

(133) ITEM: Type={s: Class=>Level=>FORM; n: Number; p: Person}**{pos: POS};

(134) FORM: Type=Case=>NClass=>Str;

Tipicamente, para linearizar uma função que combina dois ou mais tipos, como *Pred* e *Mod*, basta concatenar sequencialmente as cadeias que representam os diferentes argumentos da função, como em (106) e (107). Em nheengatu, esse também é geralmente o caso. No

entanto, para linearizar corretamente (119), a representação ortográfica do PSOR deve ser enxertada dentro da do PSUM. Por exemplo, em (17), o DP sujeito *Pedro* (PSOR) deve ser inserido em *nhaã pindá* 'aquele anzol' (PSUM). A simples concatenação desses dois constituintes resultaria numa construção agramatical, ver (135).

(135) *Pedro nhaã pindá
Pedro DEM.DIST anzol

(136) {s: {d: Str; h: NForm=>Str}; n: Number; p: Person} ** {pos: POS};

A definição de linearização do tipo Não-Dêitico em (136) permite gerar exemplos como (17), evitando, ao mesmo tempo, a geração de exemplos como (135). Nessa definição, o campo *s* consiste num registro com dois campos. O primeiro abriga a forma do determinante, o segundo, uma tabela do tipo *NForm*=>Forma, que determina a forma do substantivo, conforme (129). Na linearização de (119), a forma do PSOR é inserida entre as dos campos *d* e *h*.

Vejamos agora como o léxico é codificado na sintaxe concreta. Em GF, entradas lexicais possuem o formato CONCEITO=LINEARIZAÇÃO, como vimos em (98)-(102). Consideremos primeiro estes exemplos:

(137) River=regNoun ("paranã"|"paraná");

(138) Son_Of_Woman=regNoun ("mimbira"|"mbira");

(139) Brother_Of_Woman=regNoun "kiuíra";

(140) Language=regNoun "nheenga";

(141) Word=regNoun "nheenga";

O léxico modela uma relação de muitos para muitos, porque um dado conceito, numa dada língua, pode ter mais de uma linearização, ver (137) e (138),[35] ao mesmo tempo que diferentes conceitos podem compartilhar a mesma linearização, ver (140) e (141).

[35] GF permite implementar esse tipo de variação como resultado da aplicação de regras que alteram representações ortográficas de itens lexicais, análogas a regras fonológicas. Deixamos isso para uma versão futura da gramática.

A notação especial dos conceitos de (138) e (139) decorre de que não são lexicalizados em inglês (nem em português), apenas em nheengatu. O termo *mimbira* (ou *mbira*) designa tanto o filho quanto a filha de uma mulher. No nheengatu, tanto os designativos de parentes colaterais de 2º grau (i.e., irmãos) quanto os dos demais descendentes imediatos (i.e., filho e filha de homem) lexicalizam não somente o gênero da própria pessoa, mas também do parente.

Para facilitar a codificação das entradas lexicais, seguindo o exemplo de (103), implementamos no módulo OperYrl diversas operações que geram os paradigmas flexionais de substantivos e adjetivos. Nas entradas acima, a operação regNoun gera as formas de singular e plural de substantivos regulares, ou seja, uniformes.

Conforme o Quadro 4, os substantivos multiformes distribuem-se em cinco microgrupos. A operação RelPrefNoun, exemplificada em (142)-(147), permite reduzir esses microgrupos a apenas dois macrogrupos, facilitando a codificação de novas entradas lexicais. O primeiro engloba os microgrupos (i)-(iv), reunindo, portanto, a maioria dos substantivos multiformes, com entradas lexicais no formato de (142)-(146). Nesse caso, a operação RelPrefNoun tem como argumento apenas a forma de citação do substantivo, a partir da qual as demais formas são computadas. O segundo macrogrupo é constituído pelos membros do microgrupo (v), para os quais é necessário especificar, além da primeira, a terceira forma do Quadro 4, ver (147).

(142) Community=RelPrefNoun “tendaua”;

(143) Picture=RelPrefNoun “sangaua”;

(144) Street=RelPrefNoun “pé”;

(145) Path=RelPrefNoun “pé”;

(146) House=RelPrefNoun “uka”;

(147) Blood=RelPrefNoun “tuí” “tuí”;

As entradas das demais classes de palavras possuem uma estrutura análoga. Vejamos alguns exemplos. Para codificação das posposições locativas, implementamos a operação mkLoc, exemplificada em (148)-(151). Essa operação possui duas variantes, uma monoargumental para as

posposições uniformes e outra biargumental para as multiformes, compare (148)-(150) com (151). As duas primeiras entradas correspondem às preposições *on* e *in* do inglês e *em* do português. Enquanto *upé* lineariza os dois conceitos, o segundo é passível de linearização também pela posposição *pupé* e pela locução pospositiva *kuara upé*, que expressam localização no interior de algo. Na última entrada, o segundo argumento da operação indica que se trata de posposição multiforme.

(148) On=mkLoc “upé”;

(149) In=mkLoc (“upé”|”pupé”|”kuara upé”);

(150) With=mkLoc “irũmu”;

(151) Near=mkLoc “ruaki” NCS;

Verbos inativos são codificados por meio de variantes da operação mkQual: a monoargumental se aplica aos não flexionáveis, a biargumental, aos flexionáveis, compare (152) e (153) com (154) e (155). Nesse último caso, o parâmetro *NClass* permite distinguir verbos uniformes de multiformes, por meio de NCI e NCS, respectivamente.

(152) New=mkQual “pisasu”;

(153) Red=mkQual “piranga”;

(154) Heavy=mkQual “pusé” NCI;

(155) Hot=mkQual “raku” NCS;

Pronomes pessoais são codificados por meio da operação *PersPron*, que recebe três argumentos, a saber, número, pessoa e forma ortográfica, como nos exemplos (156)-(160). Observe que, na 3ª pessoa, uma única forma do nheengatu lineariza três pronomes diferentes do inglês, dado que a primeira língua não marca o gênero gramatical.

(156) YouSG=PersPron Sg P2 “indé”;

(157) YouPL=PersPron Pl P2 “penhẽ”;

(158) He=PersPron Sg P3 “aé”;

(159) She=PersPron Sg P3 “aé”;

(160) It=PersPron Sg P3 “aé”;

De forma análoga, os possessivos de (126) são linearizados em nheengatu por meio da operação *PossPron*, que possui apenas dois argumentos, número e pessoa, uma vez que a forma é gerada por uma outra operação, responsável pela geração dos prefixos inativos.

(161) YourSG=PossPron Sg P2;

(162) YourPL=PossPron Pl P2;

(163) His=PossPron Sg P3;

(164) Her=PossPron Sg P3;

(165) Its=PossPron Sg P3;

Concluímos esta seção com dados quantitativos da cobertura lexical da gramática do nheengatu. Conforme a Tabela 1, foram implementados apenas cerca de um sexto dos substantivos e posposições e um quinto dos adjetivos do glossário de Navarro (2011). A Tabela 2 apresenta um levantamento comparativo das quantidades de conceitos implementados na GrammYEP e na ITALENG. Pode-se constatar que a primeira é bem mais abrangente do que a segunda.

Há ainda um longo caminho a percorrer na codificação do léxico do nheengatu, de modo a cobrir todo o vocabulário das 13 lições de Navarro. Essa tarefa, contudo, será bastante facilitada pelas operações lexicais implementadas, as quais geram as representações de todas as subclasses de substantivos e adjetivos a partir unicamente da forma de citação na maioria dos casos, necessitando de apenas um argumento adicional para os demais.

TABELA 1 – Percentual implementado das principais categorias do glossário de Navarro (2011)

Categoria	Número de lexemas	Percentual implementado
Substantivo	305	15,1%
Adjetivo	72	20,8%
Verbo	186	0,5%
Advérbio	62	3,2%
Posposição	26	15,4%
Numeral	23	0%

Fonte: Elaboração própria.

TABELA 2 – Quantidades dos principais conceitos na GrammYEP e na ITALENG

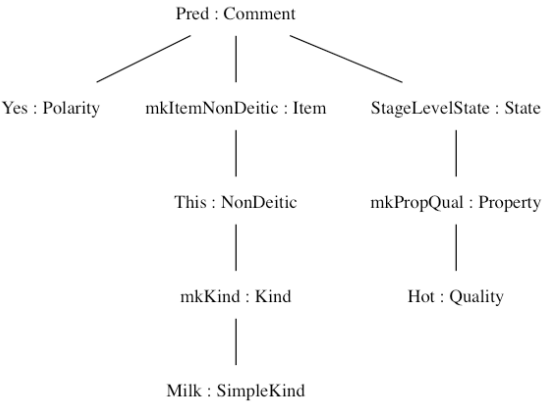
Tipo	GrammYEP	ITALENG
Classe	50	4
Qualidade	15	6
Locação	5	0

Fonte: Elaboração própria.

5 Avaliação da gramática

Nesta seção, apresentamos dados que permitem avaliar a qualidade da gramática computacional do nheengatu na análise e geração de sentenças. No primeiro caso, cada sentença definida como gramatical recebe uma ou mais representações semânticas, correspondentes às diferentes leituras da sentença conforme a sintaxe abstrata. No segundo, são geradas todas as linearizações possíveis de uma dada representação semântica, conforme as sintaxes concretas dadas.

FIGURA 5 – Análise da sentença (1) em termos de funções e tipos semânticos

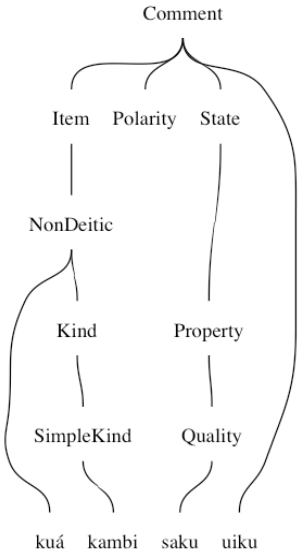


Fonte: Elaboração própria.

São três os tipos de formatos de representação semântica que o software GF pode produzir para cada leitura de sentença passível de geração por uma gramática dada: (i) árvore em formato textual, (ii) gráfico arbóreo de funções e tipos e (iii) gráfico arbóreo ligando palavras e tipos semânticos. O primeiro formato é a base para geração dos dois últimos, os quais constituem variações notacionais mais amigáveis. Exemplifiquemos com base em (1). Para essa sentença não ambígua, o analisador gera unicamente a árvore (166). Essa árvore indica as sucessivas aplicações funcionais para construir os tipos de que se constitui a representação semântica da sentença. A partir dessa representação, podem ser gerados os gráficos da Figura 5 e na Figura 6. No primeiro, cada nó constitui um par ordenado *f:t*, onde *f* é a função correspondente de (166), e *t* é o tipo produzido por essa função, cujos argumentos, no caso de funções com aridade maior que zero, são os respectivos nós filhos. Por exemplo, *Pred*, a função hierarquicamente mais alta, gera um Comentário a partir dos tipos Polaridade, Item e Estado, gerados, por sua vez, respectivamente, pelas funções indicadas à esquerda dos dois pontos em cada caso.

(166) Pred Yes (mkItemNonVar (This (mkKind Milk))) (StageLevelState (mkPropQual Hot))

FIGURA 6 – Correspondência entre palavras e tipos semânticos do exemplo (1)



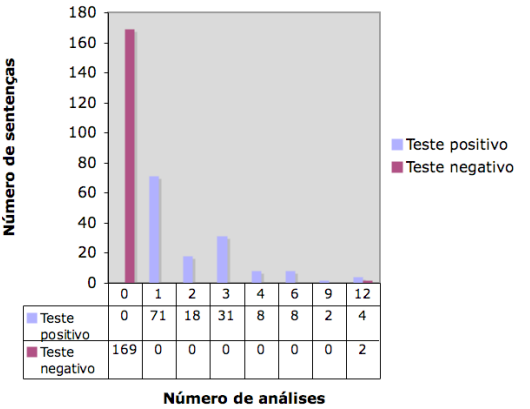
Fonte: Elaboração própria.

No gráfico da Figura 6, os nós não terminais são os tipos semânticos, enquanto os terminais, ou seja, as "folhas" da árvore, são as palavras da sentença, que constituem linearizações desses tipos. Conforme esse gráfico, a sentença é do tipo Comentário, o nó raiz da árvore, que se constitui dos tipos dos nós filhos Item, Polaridade e Estado, os quais, por sua vez, se constituem dos tipos dos respectivos filhos. O gráfico distingue formalmente as palavras sincategoremáticas *kuá* 'este' e *uiku* 'está' das categoremáticas *kambi* 'leite' e *saku* 'quente'. Enquanto estas são os únicos filhos dos respectivos pais, aquelas têm tipos como irmãos.

Um fragmento de gramática consiste num modelo de um dado recorte de fenômenos gramaticais. Desse modo, deve satisfazer duas exigências simultâneas: (i) analisar todas as sentenças gramaticais passíveis de ser construídas utilizando os recursos desse recorte, (ii) não gerar nenhuma sentença agramatical. Para tanto, foram compilados dois conjuntos-teste iniciais, um negativo e outro positivo, denominados GraYrl-Neg e GraYrl-Pos, constituídos, respectivamente, de 171 sentenças agramaticais, como (20)-(23), (81)-(84) e (90), e de 142 sentenças

gramaticais, incluindo todos os exemplos deste trabalho, com exceção de dois grupos. O primeiro contém DPs quantificados ou múltiplos possuidores sob a forma de núcleo nominal, ver (49), (54) e (60). O segundo consiste de construções que extrapolam o domínio das predicações qualificativas, ver (32), (64), (65), (67), (72)-(74), (77), (78), (86) e (88).

GRÁFICO 1 – Número de análises por número de sentenças nos conjuntos-teste positivo e negativo



	0	1	2	3	4	6	9	12
Teste positivo	0	71	18	31	8	8	2	4
Teste negativo	169	0	0	0	0	0	0	2

Fonte: Elaboração própria.

O Gráfico 1 apresenta os resultados da aplicação da gramática do nheengatu na análise desses dois conjuntos-teste. Todas as sentenças gramaticais foram analisadas, com uma média de 2.44 análises por sentença. Apenas duas sentenças do conjunto negativo foram analisados, a saber, os exemplos (81) e (82), cada um dos quais recebeu doze análises. Essas análises, porém, referem-se a leituras em que o prefixo inativo na função de complemento genitivo de *uka* 'casa' e *igara* 'canoa' não é correferencial do DP asteriscado entre parênteses, como nas linearizações em português de (81) em (167) e (168). Esses exemplos não parecem semanticamente muito felizes, mas não violam nenhuma regra sintática ou morfológica. De fato, as sentenças estruturalmente análogas (169) e (170) são aceitáveis.

(167) a casa dele da mulher é na cidade

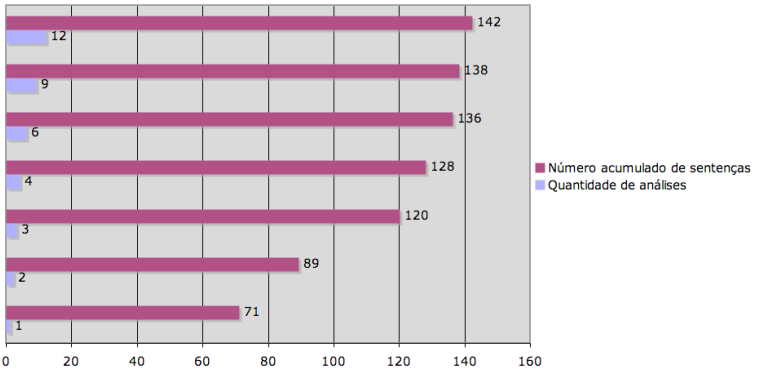
(168) a mulher é na cidade da casa dele

(169) a fotografia dele da canoa está ali

(170) a casa é na cidade da irmã dele

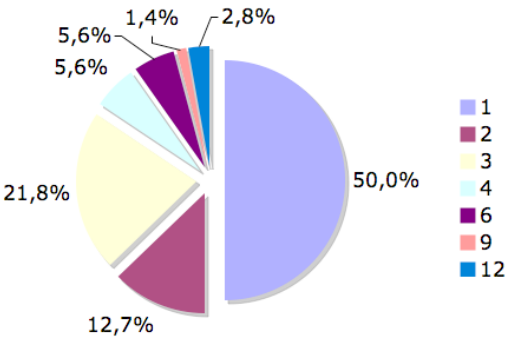
Uma das maiores dificuldades da análise sintática automática é o elevado número médio de análises geradas para as sentenças de entrada, uma vez que, via de regra, apenas uma interpretação está em jogo num determinado contexto (LJUNGLÖF; WIRÉN, 2010). O Gráfico 2 e o Gráfico 3 mostram que o grau de ambiguidade produzido pela gramática nesse conjunto-teste é baixo, uma vez que a grande maioria das sentenças (120 do total ou 84,5%) recebeu entre uma e três análises, sendo que a metade recebeu apenas uma análise.

GRÁFICO 2 – Quantidade de análises por número acumulado de sentenças do conjunto-teste positivo



Fonte: Elaboração própria.

GRÁFICO 3 – Percentual de sentenças do conjunto-teste positivo por número de análises



Fonte: Elaboração própria.

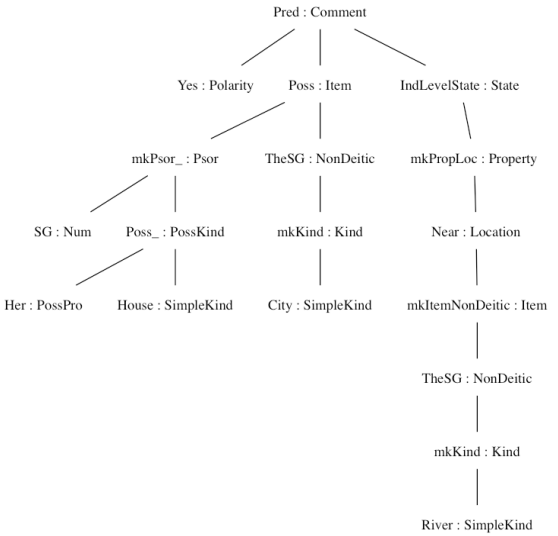
A ambiguidade pode ser de dois tipos: lexical e gramatical (LYONS, 1995). No primeiro caso, uma palavra lineariza mais de um conceito, como em (144) e (145) ou (158)-(160). No segundo, uma mesma sequência de palavras corresponde a duas ou mais configurações arbóreas, como nas leituras gramaticais de (81) e (82) acima referidas. Com frequência, os dois tipos de ambiguidade interagem, o número de leituras de um multiplicando o do outro. Esse é o caso de (171). Por um lado, *taua* 'casa' liga-se ou à palavra precedente, funcionando como seu núcleo regente, ou à subsequente como seu complemento nominal, conforme as representações da Figura 7 e da Figura 8, respectivamente, para as quais o sistema gera as linearizações em português (172) e (173). Por outro lado, o prefixo inativo *s-*, na função de possessivo nesse exemplo, possui três leituras, conforme (163)-(165). A interação entre todas essas ambigüidades produz 2x3=6 leituras.

(171) s-uka taua paranã r-uaki
3s.INACT-casa cidade rio RLL-perto.de

(172) a cidade da casa dela é perto do rio

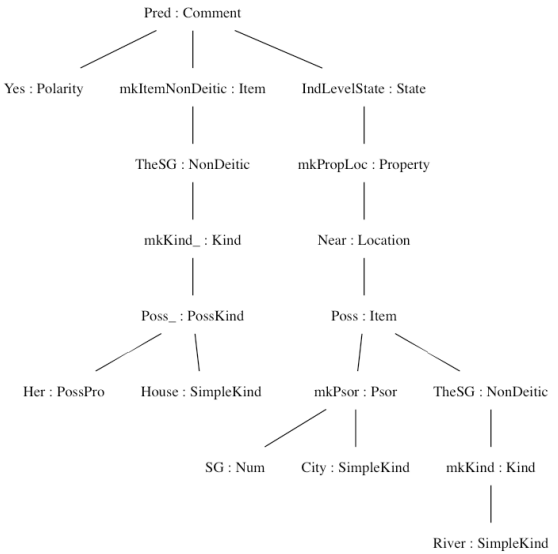
(173) a casa dela é perto do rio da cidade

FIGURA 7 – Leitura de (171) correspondente a (172)



Fonte: Elaboração própria.

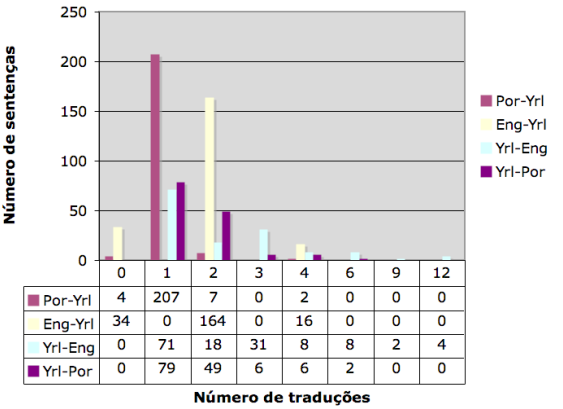
FIGURA 8 – Leitura de (171) correspondente a (173)



Fonte: Elaboração própria.

A gramática foi avaliada também como recurso de tradução automática entre o nheengatu, o inglês e o português, com o primeiro idioma tanto como língua-fonte quanto língua-alvo. No primeiro caso, o sistema produz análises do tipo de (166) para as sentenças do nheengatu, a partir das quais gera as linearizações correspondentes do inglês e do português. No segundo caso, o processo é inverso: as representações semânticas geradas para as sentenças do inglês e do português são linearizadas em nheengatu. O Gráfico 4 apresenta os resultados. Os 4 pares de língua-fonte e língua-alvo (nesta ordem) estão identificados pelos respectivos códigos no padrão ISO 639-3 (EBERHARD; SIMONS; FENNIG, 2020), aqui com inicial maiúscula.

GRÁFICO 4 – Número de traduções por número de sentenças nos conjuntos-teste Por-Yrl, Eng-Yrl, Yrl-Eng e Yrl-Por



	0	1	2	3	4	6	9	12
Por-Yrl	4	207	7	0	2	0	0	0
Eng-Yrl	34	0	164	0	16	0	0	0
Yrl-Eng	0	71	18	31	8	8	2	4
Yrl-Por	0	79	49	6	6	2	0	0

Fonte: Elaboração própria.

Todas as 142 sentenças do conjunto-teste positivo foram traduzidas para o português e o inglês, com uma média de 1,63% e 2,44 traduções por sentença, respectivamente, eliminando as repetições.[36] Comentemos primeiro os resultados obtidos com o par Yrl-Por. Mais da metade das sentenças-fonte (79 ou 55.63%) recebeu uma única tradução,

[36] As repetições decorrem de distinções semânticas que não são expressas em todas as três línguas, por exemplo, a distinção entre gênero masculino, feminino e neutro do inglês ou entre predicado de nível de indivíduo e predicado de nível de fase do nheengatu e do português.

mais de um terço (49 ou 34,51%) exatamente duas e a pequena parcela restante (14 ou 9,86%) entre 3 e 6. Apenas duas sentenças não foram traduzidas corretamente, a saber (174), baseada em (53), e variante com o alomorfe *aintá* em vez de *ta*. Na tradução dessas sentenças, a gramática do português produziu a forma agramatical *delea* em vez de *deles*.

(174) ne kiuíra ta nheenga puranga
2s.INACT irmão.de.mulher 3p.INACT palavra bonito
'a palavra deles do teu irmão é bonita'

Como mostra o Gráfico 4, na tradução do nheengatu para o inglês (identificada por Yrl-Eng), 120 sentenças, portanto, a grande maioria (84,51% do total), receberam entre uma e três traduções, ao passo que as 22 restantes (15,49%), entre quatro e doze. Todas as traduções permitem compreender as diferentes leituras das sentenças-fonte. No entanto, 59 traduções (de 7 sentenças-fonte) apresentam desvios em relação ao inglês padrão contemporâneo, no que tange à ordem de demonstrativos e possessivos. Por exemplo, em vez de (177), (175) é traduzida como (176), ao lado de duas outras possibilidades, dada a não especificação de gênero do possessivo na sentença-fonte. Enquanto essa construção ocorria normalmente em estágios anteriores do inglês, hoje em dia é considerada obsoleta (COMBINING, 2019). Por outro lado, o sistema traduz (91) como (178), quando (179) seria o correto. Esses dados apontam para a necessidade de corrigir a linearização dos possessivos na sintaxe concreta do inglês.

(175) nhaã x-imbiú puranga
DEM.DIST 3s.INACT-comida bonito
'aquela comida dele é bonita'

(176) ?that his food is beautiful

(177) that food of his is beautiful

(178) *my house is in that dirty his street

(179) my house is in that dirty street of his

A partir das traduções em português e inglês corrigidas, foram compilados os conjuntos-teste Por-Pos e Eng-Pos, com 220 e 214 sentenças cada, a fim de avaliar a tradução automática para o nheengatu. Conforme o Gráfico 4, 216 e 180 sentenças, representando 98,18% e 84,11% de cada conjunto, obtiveram traduções em nheengatu, com uma média de 1,04 e 1,83 tradução por sentença, respectivamente.

Todas as traduções do inglês estão corretas. No caso do português como língua-fonte, o único erro cometido pelo sistema foi usar o substantivo necessariamente possuível *simiriku* 'esposa' em (182) e (183) como tradução de (180) e (181), respectivamente, ao lado das traduções que realizou corretamente com o substantivo *kunhã* 'mulher', capaz de ser usado autonomamente. A fonte desse problema é a ambiguidade da palavra *mulher* em português, que lineariza tanto o conceito *Wife* quanto *Woman*.

(180) a mulher está bonita

(181) aqueles filhos da mulher estão com eles

(182) simiriku puranga uiku
3.INACT-esposa bonito 3p.ACT-estar
'a esposa dele está bonita'

(183) nhaã-itá s-imiriku mimbira uiku aintá irũmu
DEM.DIST-PL 3.INACT-esposa filho.de.mulher 3p.ACT-estar 3p.INACT com
'aqueles filhos da esposa dele estão com eles'

O pior desempenho da tradução a partir do inglês, em que 15,89% das sentenças-fonte não foram traduzidas, contra apenas 1,82% com o português como língua-fonte, explica-se pelo problema apontado na sintaxe concreta dos possessivos, ver (176) e (178). Essa questão representa uma dificuldade também na tradução a partir do português, porém, em menor grau, uma vez que só afeta os possessivos de 3ª pessoa, no caso de sentenças-fonte como (174).

Desta avaliação da gramática resultou um *treebank* com 243 sentenças do nheengatu, reunindo o conjunto-teste positivo e as traduções dos conjuntos Por-Yrl e Eng-Yrl, emparelhadas com as sentenças equivalentes em português e inglês.

6 Considerações finais

Neste artigo, apresentamos a implementação computacional de um fragmento do nheengatu abrangendo cerca de um quinto do conteúdo gramatical de Navarro (2011). Esse fragmento integra a GrammYEP, uma gramática computacional multilíngue no formalismo GF, da qual fazem parte fragmentos análogos do português e do inglês. Com isso, o sistema traduz do nheengatu para essas duas línguas e vice-versa. No momento, limita-se a orações que atribuem qualidades e localizações contingentes e não-contingentes a pessoas e coisas. O vocabulário é reduzido, mas pode ser facilmente expandido por meio das operações lexicais implementadas.

Pré-requisito para a implementação computacional de uma língua é uma rigorosa formalização das estruturas gramaticais e lexicais. Dado o caráter não formalizado de Navarro (2011) e Cruz (2011), as duas descrições do nheengatu utilizadas, formalizamos inicialmente as estruturas de constituintes na CFG. Em seguida, modelamos formalmente as restrições de concordância e valência na combinatória de constituintes. Finalmente, integramos ambas as dimensões na implementação em GF.

A formalização revelou lacunas e inconsistências daquelas duas abordagens no tratamento das expressões nominais, que em parte sanamos com base nos dados desses mesmos autores. No entanto, algumas questões levantadas ficaram por esclarecer à luz de novos dados, entre as quais destacamos as seguintes:

i. A complementação genitiva é recursiva?
ii. Qual o estatuto do possuidor nessa construção? DP ou NP?
iii. A dupla marcação de plural com *-itá* é licenciada?
iv. Os substantivos necessariamente possuíveis possuem uma forma absoluta?

A GrammYEP obteve resultados bastante satisfatórios tanto na análise quanto na tradução de sentenças das três línguas. Traduziu para o português e o inglês todas as 142 sentenças do conjunto-teste positivo do nheengatu. Inversamente, verteu para o nheengatu 98,18% e 84,11% dos conjuntos-teste correspondentes do português e do inglês, com 220 e 214 sentenças, respectivamente. A partir do conjunto-teste do nheengatu e das traduções do português e do inglês constituiu-se um *treebank* do

nheengatu com 243 sentenças, emparelhadas com as equivalentes nas duas outras línguas.

O sistema ainda padece de hipergeração, problema a ser enfrentado nas próximas versões. Nesse quesito, o nheengatu representa o maior desafio, por conta de particularidades lexicais e sintáticas sem equivalência nas outras duas línguas. A primeira dessas dificuldades são os substantivos necessariamente possuíveis, cujo complemento genitivo precisa sempre realizar-se, ver (32). A segunda são os termos designativos de descendentes imediatos e parentes colaterais de 2º grau, que lexicalizam o gênero do parente.

Dado seu caráter de software livre, a GrammYEP oferece diversas oportunidades de colaboração por parte da comunidade de linguistas e programadores em GF visando à solução dos problemas apontados e à expansão do sistema.

Agradecimentos

Agradecemos aos dois pareceristas anônimos pelos comentários e sugestões.

Referências

ÁVILA, M. T. *Estudo e prática da tradução da obra infantil* A terra dos meninos pelados*, de Graciliano Ramos, do português para o nheengatu*. 2016. 199f. Dissertação (Mestrado em Estudos da Tradução) – FFLCH, Universidade de São Paulo, São Paulo, 2016.

BENDER, E. M. Grammar Engineering for Linguistic Hypothesis Testing. *In*: GAYLORD, N. *et al.* (org.). *The Proceedings of the Texas Linguistics Society 10*: Computational Linguistics for Less-Studied Languages. Stanford: CSLI, 2008. p. 16-36.

BENDER, E. M. Reweaving a Grammar for Wambaya. *Linguistic Issues in Language Technology*, Stanford, v. 3, n. 3, p. 1-36, 2010.

BENTLEY, D. Copular and Existential Constructions. *In*: DUFTER, A.; STARK, E. (org.). *Manual of Romance Morphosyntax and Syntax*. Berlin: De Gruyter, 2017. p. 332-366. DOI: https://doi.org/10.1515/9783110377088-009

BERNSTEIN, J. B. The DP Hypothesis: Identifying Clausal Properties in the Nominal Domain. *In*: BALTIN, M.; COLLINS, C. (org.). *The Handbook of Contemporary Syntactic Theory*. Malden: Blackwell, 2003. p. 536-561. DOI: https://doi.org/10.1111/b.9781405102537.2003.00019.x

BRESNAN, J. *Lexical-Functional Syntax*. Malden: Blackwell, 2001.

BUSSMANN, H. (org.). *Lexikon der Sprachwissenschaft*. 3. ed. Stuttgart: Kröner, 2002.

CARNIE, A. *Syntax*: A Generative Introduction. 3. ed. Malden: Blackwell, 2012.

CASASNOVAS, A. *Noções de língua geral ou nheengatú*: gramática, lendas e vocabulário. 2. ed. Manaus: Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, 2006.

COMBINING Demonstrative and Possessive Pronoun. [*S.l.*]: [*S.n.*], 2019. Disponível em: https://english.stackexchange.com/questions/476384/combining-demonstrative-and-possessive-pronoun. Acesso em: 19 jun. 2020.

COMRIE, B. *Language Universals and Linguistic Typology*: Syntax and Morphology. Oxford: Blackwell, 1983.

CRUZ, A. *Fonologia e gramática do nheengatú*: a língua falada pelos povos Baré, Warekena e Baniwa. Utrecht: LOT, 2011.

CRUZ, A. The Rise of Number Agreement in Nheengatu. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, Belém, v. 10, n. 2, p. 419-439, 2015. DOI: https://doi.org/10.1590/1981-81222015000200011

DUCHIER, D.; PARMENTIER, Y. High-level Methodologies for Grammar Engineering, Introduction to the Special Issue. *Journal of Language Modelling*, Warszawa, Poland, v. 3, n. 1, p. 5-19, 2015. DOI: https://doi.org/10.15398/jlm.v3i1.117

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (org.). *Ethnologue*: Languages of the World. 23. ed. Dallas: SIL International, 2020. Disponível em: http://www.ethnologue.com. Acesso em: 12 jun. 2020.

FÁBREGAS, A. Adjectival and Genitival Modification. *In*: DUFTER, A.; STARK, E. (org.). *Manual of Romance Morphosyntax and Syntax*. Berlin: De Gruyter, 2017. p. 771-803. DOI: https://doi.org/10.1515/9783110377088-021

FRANCEZ, N.; WINTNER, S. *Unification Grammars*. Cambridge: CUP, 2012.

FREIRE, J. R. B. *Rio Babel*: a história das línguas na Amazônia. 2. ed. Rio de Janeiro: EdUERJ, 2011.

GYNAN, S. Morphological Glossing Conventions for the Representation of Paraguayan Guaraní. *In*: ESTIGARRIBIA, B.; PINTA, J. (org.). *Guarani Linguistics in the 21st Century.* Leiden: Brill, 2017. p. 86-130.

HAJIČOVÁ, E. *et al.* Treebank Annotation. *In*: INDURKHYA, N.; DAMERAU, F. J. (org.). *Handbook of Natural Language Processing*. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010. p. 167-188.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2. ed. Upper Saddle River: Prentice Hall, 2009.

KARVOVSKAYA, L. *The Typology and Formal Semantics of Adnominal Possession*. Utrecht: LOT, 2018.

LEANDRO, W. M.; AMARAL, L. A. The Interpretation of Multiple Embedded Genitive Constructions by Wapichana and English Speakers. *Revista LinguíStica*, Rio de Janeiro, v. 10, n. 2, p. 149-162, 2014.

LJUNGLÖF, P.; WIRÉN, M. Syntactic Parsing. *In*: INDURKHYA, N.; DAMERAU, F. J. (org.). *Handbook of Natural Language Processing*. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010. p. 59-91.

LYONS, J. *Linguistic Semantics*: An Introduction. Cambridge: CUP, 1995. DOI: https://doi.org/10.1017/CBO9780511810213

MATHESIUS, V. *A Functional Analysis of Present Day English on a General Linguistic Basis*. Haia: Mouton, 1975. DOI: https://doi.org/10.1515/9783110813296

MÜLLER, S. The CoreGram Project: Theoretical Linguistics, Theory Development and Verification. *Journal of Language Modelling*, Warszawa, Poland, v. 3, n. 1, p. 21-86, 2015. DOI: https://doi.org/10.15398/jlm.v3i1.91

NAVARRO, E. A. *Curso de Língua Geral (nheengatu ou tupi moderno)*: a língua das origens da civilização amazônica. São Bernardo do Campo: Paym, 2011.

NAVARRO, E. A.; ÁVILA, M. T.; TREVISAN, R. G. O Nheengatu, entre a vida e a morte: a tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, Campina Grande, v. 6, n. 2, p. 9-29, 2017. DOI: https://doi.org/10.35572/rlr.v6i2.768

PIRINEN, T. *et al.* Introduction. *In*: INTERNATIONAL WORKSHOP FOR COMPUTATIONAL LINGUISTICS OF URALIC LANGUAGES, 3., 2017, St. Petersburg. *Proceedings* […]. Stroudsburg, USA: Association for Computational Linguistics, 2017. p. iii.

PRAÇA, W. N.; MAGALHÃES, M. M. S.; CRUZ, A. Indicativo II da família Tupi-Guaraní: uma questão de modo? *Liames*, Campinas, v. 17, n. 1, p. 39-58, 2017. DOI: https://doi.org/10.20396/liames.v17i1.8646480

RANTA, A. *Grammatical Framework Tutorial*. [*S.l.*]: [*S.n.*], 2010. Disponível em: https://www.grammaticalframework.org/doc/tutorial/gf-tutorial.html. Acesso em: 15 jun. 2020.

RANTA, A. *Grammatical Framework*: Programming with Multilingual Grammars. Stanford: CSLI, 2011.

RODRIGUES, A. D. As línguas gerais sul-americanas. *Papia*, São Paulo, v. 4, n. 2, p. 6-18, 1996.

RODRIGUES, A. D. Prefácio. *In*: FREIRE, J. R. B. *Rio Babel*: a história das línguas na Amazônia. 2. ed. Rio de Janeiro: EdUERJ, 2011. p. 13-14.

SAG, I. A.; WASOW, T.; BENDER, E. *Syntactic Theory*: A Formal Introduction. 2. ed. Stanford: CSLI, 2003.

SYMPSON, P. L. *Grammatica da lingua brazilica geral, fallada pelos aborigines das provincias do Pará e Amazonas*. Manaus: Typographia do Commercio do Amazonas, 1877.

ZOMPÌ, S. *Case Decomposition Meets Dependent-Case Theories*. 2017. 108 f. Dissertacão (Mestrado em Linguística) – Corso di Laurea Magistrale in Linguistica, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Pisa, 2017. Disponível em: https://ling.auf.net/lingbuzz/003421. Acesso em: 14 jun. 2020.