Revista do GELNE, v. 22, número 2, 2020 ISSN: 2236-0883 ON LINE
DOI 10.21680/1517-7874.2020v22n2ID23284

UM ESTUDO EXPLORATÓRIO SOBRE A CLASSIFICAÇÃO DE MORFEMAS POR AGRUPAMENTO HIERÁRQUICO PARA COMPARAÇÃO TIPOLÓGICA[1]

AN EXPLORATORY STUDY INTO MORPHEME CLASSIFICATION THROUGH HIERARCHICAL CLUSTERING FOR TYPOLOGICAL COMPARISON

João Paulo Lazzarini Cyrino[2]
Eudes Barletta Mattos[3]

RESUMO
Em Tipologia Linguística é crucial que os fenômenos ou categorias observadas nas línguas sejam, de fato, comparáveis entre si. Conforme Croft (2002), essa comparabilidade é especialmente difícil quando estamos tratando de categorias sem uma definição semântico-pragmática clara, como categorias do tipo nome, verbo, sujeito, etc. Neste trabalho apresentamos um estudo exploratório sobre o uso de classes distribucionais de morfemas como uma forma de fazer comparação tipológica. Essas classes distribucionais são obtidas a partir da análise de adjacência entre os morfemas de uma língua e subsequente aplicação de algoritmo de agrupamento hierárquico. Trata-se de uma introdução ao uso de técnicas de processamento de linguagem natural, especialmente aprendizado não-supervisionado, para auxiliar na pesquisa em Tipologia Linguística. Realizamos o estudo com quatro línguas não relacionadas geográfica ou geneticamente: araweté, yakkha, pite saami e khwarshi. As classes obtidas parecem revelar informações sobre como as línguas organizam a morfologia gramatical em função da lexical, formando ou não classes de palavras que podemos reconhecer como nomes, verbos, etc.
PALAVRAS-CHAVE: tipologia linguística; comparabilidade; indução gramatical; processamento de linguagem natural.

ABSTRACT
It is crucial for the field of Linguistic Typology that features are comparable among languages. According to Croft (2002), such comparability is especially challenging whilst dealing with categories without a clear semantic-pragmatic definition, such as *nouns, verbs, subjects, etc.* Here we present an exploratory study into the usage of distributional classes of morphemes as a means for typological comparison. Such distributional classes are obtained from the analysis of adjacency between the morphemes of a given language and subsequent application of a hierarchical clustering algorithm. This is to be understood as an introduction to the usage of Natural Language Processing techniques, especially unsupervised learning, as a tool for Typological research. The study was conducted with data from four geographically and genetically unrelated languages: Araweté, Yakkha, Pite Saami and Khwarshi. The classes yielded seem to give information on how languages organize their grammatical morphology with respect to lexemes, tending or not to form word classes that can be reckoned as nouns, verbs, etc.
KEYWORDS: linguistic typology; comparability; grammar induction; natural language processing.

[1] Agradecimentos especiais a Ricardo Potozky de Oliveira, pelo registro e processamento dos dados da língua khwarshi. Também agradecemos aos pareceristas pelos valiosos comentários. Algumas das sugestões foram implementadas aqui, outras estão sendo consideradas para futuros trabalhos.
[2] Professor da Universidade Federal da Bahia (UFBA).
[3] Graduando em Letras pela Universidade Federal da Bahia (UFBA), bolsista FAPESB de Iniciação Científica.

INTRODUÇÃO

De acordo com Croft (2002), o pré-requisito básico para comparar línguas é a capacidade de identificar o mesmo fenômeno gramatical entre elas. Esta premissa se chama *Comparabilidade*: para ser estudado através de diferentes línguas, um dado fenômeno necessita ser comparável, e a questão de estudo, formulada tendo esta necessidade em vista. Trata-se de um tema bastante amplo em Tipologia Linguística e que tem sido objeto de bastante atenção nos últimos anos, sendo também formulado como o *Problema da Correspondência* (CORBETT, 2008).

A maneira mais típica de se lidar com o *Problema da Correspondência* é a de relacionar propriedades formais das categorias de marcação de uma determinada língua, como tempo e aspecto por exemplo, a conceitos semântico/pragmáticos que seriam universais. Essa possibilidade se perde, no entanto, quando o objeto de estudo são as categorias formais das línguas, como *nome, verbo, sujeito, adjunto, núcleo, complemento*. Não parece haver definições semântico-pragmáticas ou funcionais claras para essas categorias, que são justamente as categorias centrais para a análise linguística (CROFT, 2002).

Diversas formas de tornar comparáveis categorias como essas têm surgido. Podemos mencionar a *Tipologia Canônica* (BROWN; CHUMAKINA, 2012), que organiza definições particulares a cada língua com o propósito de extrair um cânone geral, e a *Tipologia Multivariada* (BICKEL, 2006), que define as categorias em termos de correspondências comuns de comportamentos variáveis específicos de cada língua.

Enquanto estas são, certamente, contribuições bastante relevantes para a Tipologia Linguística, com importantes resultados, aqui sugerimos outra forma de lidar com o *Problema da Correspondência*, uma forma que acreditamos ser mais harmoniosa com a metodologia indutiva característica da *Tipologia Linguística*. Nesse tipo de metodologia, não é interessante trabalhar com conceitos definidos mormente com base na tradição gramatical (*nome, verbo, sujeito*, por exemplo), mas sim basear-se em métodos rigorosos de obtenção de dados e em uma atitude parcimoniosa para com a postulação de entidades abstratas, pois estas precisam refletir rigorosamente o que é observado. Desta forma, o ônus da prova de se criarem objetos teóricos deve recair sempre sobre o cientista que os propõe (GOLDSMITH, 2007) e não sobre uma vaga tradição.

Neste artigo, portanto, introduzimos a comparação de línguas baseada em categorias distribucionais de morfemas, obtidas por técnicas de aprendizagem de máquina não-supervisionada, também conhecidas como *indução gramatical* (CLARK; LAPPIN, 2010). Essas técnicas formam parte da área de Processamento de Linguagem Natural (PLN), e inspiram-se no estruturalismo, em especial nos trabalhos de Zellig Harris que, ao longo de sua carreira, buscou relacionar teorias probabilísticas e teoria da informação à distribuição dos morfemas em uma sentença (GOLDSMITH, 2010). Embora possamos considerar que a Tipologia Linguística surge como uma resposta ao não-universalismo visto em abordagens estruturalistas, é importante ponderar, como em Haspelmath (2010), que esse particularismo se dá a respeito das categorias descritivas; os métodos utilizados para se obterem essas categorias, no entanto, devem ser aplicáveis a quaisquer línguas. Nesse sentido, podemos propor uma linha de comparação tipológica baseada no tipo de resposta que os dados dão para determinado método descritivo.

Apresentamos, portanto, um estudo inicial e exploratório que ilustra o potencial desse tipo de comparação a partir de um método de classificação não-supervisionada de morfemas em uma dada língua, baseando-se em suas propriedades distribucionais. Aplicamos a técnica a dados de quatro línguas geográfica e geneticamente diversas: khwarshi (nakh-daguestaniana, tsézica: Rússia; KHALILOVA, 2009), araweté (tupiana, tupi-guarani: Brasil; SOLANO, 2009), yakkha (sino-tibetana, kiranti: Nepal; SCHACKOW, 2014) e pite saami (urálica, fínica: Suécia; WILBUR, 2014). Os resultados apontam haver associação entre algumas estatísticas, como a razão de morfemas que se agrupam em classes distribucionais menores e regulares sobre o número total de morfemas, e a centralidade – em cada língua – do que comumente reconhecemos como classes de palavras.

O artigo se inicia com uma explicação sobre método utilizado e os conceitos-chave de bigramas, grafos, matrizes de adjacência e o algoritmo de agrupamento hierárquico dentro do contexto do PLN não-supervisionado. Disso se segue uma breve apresentação sobre a inovação de se implementar esse tipo de classificação aos morfemas da língua e não a palavras, o que é possibilitado a partir do uso de dados de gramáticas descritivas. Por fim, aplica-se o método aos dados das quatro línguas mencionadas e se discute a interpretação dos resultados.

1 Agrupamento Hierárquico de Morfemas por Propriedades Distribucionais

O método que trazemos neste artigo está baseado em uma técnica de classificação não-supervisionada de palavras como um primeiro passo para que a máquina possa fazer análise sintática. O método é inspirado em Schütze (1993) e Schütze (1995)[4] e constitui na aplicação de um algoritmo de agrupamento hierárquico a uma matriz de adjacência (vizinhança imediata) de morfemas.

Para entender no que consiste o método, tomemos alguns exemplos da língua yakkha:

(1) a. *ta-ya-na*
vir-PST-NMLZ.SG
"Ela/ele veio."

b. *ta-yatasa-na*[5]
vir-PST.PROG-NMLZ.SG
"Ela/ele estava vindo."

c. *ta-meʔ-na*
vir-NPST-NMLZ-SG
"Ela/ele vem/virá."

Com o intuito de representar as relações de morfemas em diferentes sequências possíveis na língua, lançamos mão de um campo da matemática chamado *teoria dos grafos*, de grande aplicação em PLN (NASTAE, 2015). Um grafo é um objeto que contém um conjunto de pontos -- nós ou vértices -- conectados por linhas -- arestas ou arcos (BONDY, 1976). As árvores sintáticas, bastante conhecidas dos linguistas, são um exemplo de grafo do tipo binário, cujos nós conectam-se a, no máximo, mais dois outros. Baseando-nos nos dados acima, podemos representar as sequências possíveis de morfemas em yakkha utilizando grafos direcionados, cujas arestas indicam relações orientadas em um só sentido, como ilustrado a seguir:

[4] Para o autor era importante classificar as palavras em classes de palavra consolidadas, como verbos, adjetivos, etc. Para tanto,utilizam-se vários artifícios tanto para conseguir tal classificação como para testar essa classificação. Aqui apenas mostramos a parte mais trivial de se obter uma classificação a partir de uma matriz de adjacência.

[5] Apesar da semelhança com o morfema *ya* (passado), o que poderia levar à análise de *yatasa* como sendo composto de **ya-tasa*, SCHACKOW (2014) traz a forma verbal *tayatasa* como não fazendo parte de um paradigma, e o morfema *yatasa* (passado progressivo), como não-segmentável.

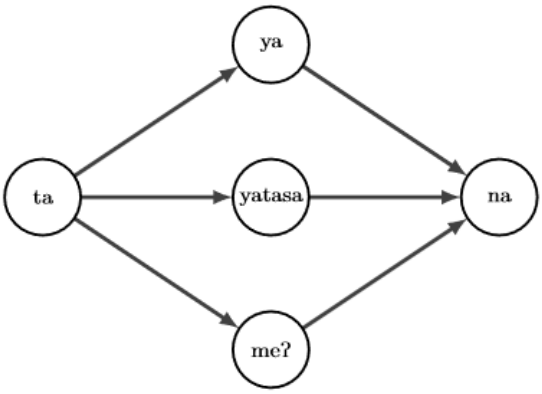

Figura 1: **Grafo ilustrando sequências possíveis de morfemas em yakkha**

Além da representação gráfica, um grafo também pode ser representado por uma matriz de adjacência. Isso nos permite realizar importantes operações matemáticas como as que envolvem o algoritmo de agrupamento hierárquico, que forma classes distribucionais de morfemas. Em uma matriz de adjacência para os morfemas do yakkha tais quais representados na Figura 1, cada linha representa o primeiro morfema de um par e cada coluna representa o segundo morfema. Se o par existir nos dados, o valor é 1, caso contrário, é 0. Há outras possibilidades de valor, como por exemplo a contagem de pares ou probabilidade do par, mas para o modelo exploratório apresentado neste artigo nos ativemos à versão mais simples.

	ta	*ya*	*yatasa*	*me?*	*na*
ta	0	1	1	1	0
ya	0	0	0	0	1
yatasa	0	0	0	0	1
me?	0	0	0	0	1
na	0	0	0	0	0

Tabela 1: **Matriz de Adjacência para morfemas do yakkha**

À matriz de adjacência podemos aplicar um algoritmo de agrupamento que classificará cada morfema em função de seus vizinhos. Trata-se, portanto, de um agrupamento baseado em critérios distribucionais (SCHÜTZE, 1995). Para este estudo, utilizamos uma versão de algoritmo de agrupamento denominado *AgglomerativeClustering*, que é parte do conhecido pacote de aprendizagem de máquina *scikit-learn* para a linguagem de programação *Python*.

Também chamados de algoritmos de *Clustering*, algoritmos de agrupamento analisam uma matriz de *m x n* dimensões como um conjunto de *n* vetores em um espaço *m*-dimensional. Dessa forma, ele mede a distância entre esses vetores e agrupa os que estiverem mais próximos. *AgglomerativeClustering* é um algoritmo de agrupamento especial que agrupa os dados hierarquicamente em termos de suas semelhanças e diferenças (Agrupamento Hierárquico)[6]. É possível observar a atuação do algoritmo a partir de uma visualização denominada *dendrograma*:

[6] Para uma descrição mais técnica desse tipo de algoritmo sugerimos ver Friedman; Hastie & Tibchirani (2001).

Figura 2: **Dendrograma baseado no agrupamento hierárquico de morfemas do yakkha**

O dendrograma obtido a partir dos dados do yakkha apresentados em (1) mostra que os morfemas *ya*, *yatasa* e *me?* encontram-se no mesmo grupo. Esse grupo forma um outro grupo junto com o morfema *na*, que forma um grupo distinto do que abarca os morfemas *ya*, *yatasa* e *me?*. O morfema *ta* foi avaliado como o mais diferente dentre os cinco.

2 Morfemas, não palavras, como unidade de formação da sentença

Cabe uma explicação sobre porque orientamos nossa análise para morfemas e não palavras. A noção de palavra, embora aludida à exaustão em publicações dos campos da morfologia e da sintaxe, segue uma das mais indefinidas em toda a linguística. Assume-se que qualquer falante saiba o que é uma palavra, e, desta forma, o conceito é utilizado para circunscrever a competência da morfologia e também como ponto de partida para o estudo da sintaxe – "morfologia lida com a composição das palavras enquanto a sintaxe lida com a combinação das palavras" (DIXON; AIKHENVALD, 2003). É aí, portanto, que reside o problema; até o momento, não se chegou numa definição de *palavra* que seja consistentemente válida para um conjunto significativo de línguas do mundo, ou até mesmo para uma só delas (BOOIJ, 2012; DIXON; AIKHENVALD, 2003; HASPELMATH, 2011; HASPELMATH, 2013).

É inevitável que esta noção soe um tanto mais adequada a línguas europeias ditas sintéticas; não por acaso, a linguística ocidental se iniciou com a tradição greco-romana, que se debruçava sobre o grego antigo e o latim – línguas fusionais, de difícil segmentação em morfemas. A abordagem favorecida pelos gramáticos clássicos foi, desde o início, a paradigmática, em que uma palavra tentativamente bem-delimitada assumiria diferentes formas de acordo com sua função numa sentença, ou da informação veiculada. É o caso das listas de conjugação verbal, conhecidas pelos estudantes de qualquer uma das línguas românicas, assim como os estudantes do grego e latim clássicos se deparam com tabelas de declinação. A compreensão do que seria palavra, portanto, surge com uma carga bastante alta de particularismo; o conhecimento e descrição de diferentes línguas que não indo-europeias, bem como a inserção destas línguas na sociedade escrita mostra que palavra não é algo tão natural quanto tem sido considerado pela linguística moderna.

Tomemos como exemplo o caso das línguas bantu do sul da África, que possuem uma estrutura verbal tida por aglutinante; mais importante, no entanto, é nos determos no fato de que estas línguas são sujeitas, em sua modalidade escrita, a estratégias ortográficas diferentes. Van Wyk (1968) sintetiza este panorama em duas tendências; o que ele chama de *disjuntivismo*, segundo o qual unidades linguísticas curtas são escritas como sendo palavras discretas, e o *conjuntivismo*, no qual unidades simples são unidas para formar palavras longas e morfologicamente complexas. Em seguida, traz um exemplo da língua sotho do norte, o qual consideramos oportuno reproduzir.

(2) *re* *tlo* *e* *bua* *ka* *thipa* *ya* *gagwe*
1.PL.SUJ FUT 3.SG.OBJ esfolar INS faca CLASS9 dele
"Nós o esfolaremos com a faca dele."

Esta mesma frase poderia ser escrita de duas formas, de acordo com a estratégia ortográfica a ser adotada: *re tlo e bua ka thipa ya gagwe*, de acordo com o *disjuntivismo*; e *retloebua kathipa yagagwe*, de acordo com o *conjuntivismo*. As diversas línguas bantu no sul da África seguem convenções ortográficas diferentes, afiliando-se a uma ou outra tendência entre as duas mencionadas, situando como discutível não só a validade da noção de *palavra*, como também as fronteiras entre a morfologia e a sintaxe.

Desconsiderar o conceito de palavra, no entanto, implica propor formas de lidar com outros conceitos que estão na base de nosso entendimento sobre a natureza estrutural das línguas. Inicialmente, faz-se necessário definir uma unidade mínima para trabalho que seja razoavelmente bem estabelecida e prontamente distinguível. Bronisław Malinowski, eminente antropólogo do início do século XX, propunha que a unidade linguística de trabalho na pesquisa de campo fosse a sentença produzida, e não unidades menores desprovidas de contexto. Como expoente do funcionalismo, chegou a afirmar que palavras isoladas seriam artefato de análises linguísticas avançadas, ou seja, uma mera invenção da linguística (MALINOWSKI, 1933). Entre linguistas, pensadores como Charles Bally e André Martinet sempre punham *palavra* entre aspas, como se evidenciando seu caráter um tanto artificial. Dixon & Aikhenvald (2003) tratam minuciosamente em seu livro a noção de palavra, sem chegar num conceito universal; ao longo de sua argumentação, que discute o que seria a palavra segundo definições fonológicas, gramaticais ou ortográficas, o único consenso entre as línguas é justamente não existir consenso algum quanto à existência ou não da *palavra* como unidade natural e bem circunscrita.

Em outra direção, já desde 1946, Zellig Harris propõe não a palavra, mas sim o morfema como unidade básica de análise morfossintática (HARRIS, 1946), e é a esta compreensão que nos alinhamos. Para ele, o morfema seria "o mais simples [item] observável", prontamente identificável e de fácil visualização. De fato, ao contrário da palavra, a noção de morfema é consistente translinguisticamente e amplamente utilizada e veiculada em gramáticas descritivas; definir o morfema como item de trabalho, portanto, torna possível a comparação entre descrições das mais diversas línguas – de grande interesse para estudos em tipologia –, ao passo em que se evitam conceitos controversos.

3 Os dados e a plataforma Òcun

Os dados utilizados neste estudo foram retirados de gramáticas descritivas, que constituem o melhor tipo de fonte para estudos tipológicos, dada a existência de informações sobre as línguas em seus vários níveis de análise (da fonética à pragmática) e, particularmente relevante para o presente trabalho, pela diversidade estrutural dos dados.

Para a análise estatística de dados de gramáticas descritivas, o LATIP, Laboratório de Tipologia Linguística, lotado no Instituto de Letras da Universidade Federal da Bahia, vem desenvolvendo, desde 2019, uma ferramenta online denominada plataforma Òcun (https://ocun.latip.com.br). A plataforma consiste em um banco de frases[7] provenientes de gramáticas descritivas, segmentadas por morfemas, e produz – a partir desses dados – tabelas de frequência de formas, significados e de morfemas, histogramas de probabilidade e complexidade (logaritmo positivo da probabilidade), dentre outras análises estatísticas. Ela está sendo

[7] Por conveniência, tratamos por *frase* mesmo que no banco de dados haja sequências de morfemas menores que uma frase, como sintagmas nominais, adjetivais etc.

desenvolvida de forma que possa ser alimentada por qualquer usuário após um breve cadastro. A inserção de dados é feita a partir de gramáticas das mais diversas línguas, contanto que descritivas, com exemplos segmentados em morfemas e possuindo licença ou autorização que as permitam ser utilizadas dessa forma.

A plataforma, quanto à maneira em que é alimentada ou fornece os dados armazenados, opera sob a mesma lógica de seu material-fonte, a gramática descritiva: na primeira linha, os dados da língua são apresentados morfema a morfema; na segunda, a glosa, contendo a informação veiculada por cada morfema da frase; e na terceira, a tradução. Por consequência, podem-se recuperar morfemas do banco de dados tanto por forma quanto por significado, ou frases inteiras; todo o conteúdo inserido da língua também é disponibilizado para cópia por meio de uma ferramenta oferecida pela própria plataforma.

Os dados utilizados neste estudo encontram-se registrados na plataforma Òcun e de lá foram extraídos para pré-processamento por um programa feito especialmente para o estudo. O programa, escrito na linguagem *Python*, constrói um grafo a partir dos dados da língua e dele obtém-se a matriz de adjacência utilizada pelo algoritmo de agrupamento hierárquico.

4 Aplicação do algoritmo e discussão

Na lida com modelos estatísticos, uma preocupação frequentemente presente é a ocorrência de sobreajuste do modelo – é dizer, o modelo perde o poder de generalização por descrever com acurácia excessiva a situação encontrada no estudo, e apenas ela. No caso de um modelo de classificação não supervisionado como o de agrupamento hierárquico, o poder preditivo não está no fim da árvore hierárquica e sim no começo; o intuito deste modelo, portanto, não é o de oferecer uma classificação definitiva, trazida nos ramos terminais, para os itens de estudo, mas sim o de evidenciar padrões de afinidade por meio de diferentes simulações de agrupamento. Isso envolve querermos observar poucas classes; por mais confusas que algumas se apresentem, haverá outras delas que revelem fatos de interesse acerca da língua.

Ao aplicarmos o algoritmo *AgglomerativeClustering* aos dados, podemos selecionar o número de classes em que queremos dividir o conjunto de morfemas da língua. Essa ferramenta é bastante importante, já que permite que o algoritmo não construa a árvore total; assim, teremos agrupamentos mais significativos. Ainda não definimos um número exato de classes que seria o mais desejável, e acreditamos que esse número será estabelecido conforme mais pesquisas com dados de gramáticas descritivas forem sendo feitas segundo esta metodologia. A título de comparação, SCHÜTZE (1995), em *corpora* de mais de 45 mil palavras, se utiliza de 50 classes para obter uma correspondência desejável com classes gramaticais mais tradicionais. Aqui mostraremos os resultados com 8 classes, considerando que nossos *corpora* são menores e que estamos lidando com morfemas, e não palavras. Exceção fica aos dados da língua araweté, para os quais mostramos, progressivamente, os resultados com 2, 4 e 8 classes, intentando ilustrar o comportamento do algoritmo.

Pela natureza da plataforma Òcun, bem como a organização dos dados, quaisquer análises que utilizem o método aqui exposto podem levar em conta somente as formas dos morfemas de uma língua, somente os significados dos morfemas, ou ambos. Optamos aqui por analisar somente os significados dos morfemas; desta maneira, analisamos relações de vizinhança entre significados, desconsiderando possíveis distorções decorrentes de casos de alomorfia, por exemplo.

4.1 Araweté

Araweté é uma língua tupi-guarani falada no interior do Pará, na aldeia Araweté Igarapé Ipixuna, e conta com cerca de 280 falantes. A fonte de dados utilizada é a gramática descritiva de Solano (2009) da qual extraímos 563 frases que nos forneceram 419 morfemas diferentes (por

significado)[8]. O agrupamento dos morfemas em duas classes resulta numa classe com a maior parte dos morfemas e outra classe menor (classe 0), com os morfemas apresentados abaixo:

Classe	Significado	Descrição da Abreviatura
0	1	Pronome de primeira pessoa
0	2	Pronome de segunda pessoa
0	3	Pronome de terceira pessoa
0	13	Pronome de primeira pessoa do plural exclusiva
0	23	Pronome de primeira pessoa do plural inclusiva
0	R1	Prefixo relacional 1

Tabela 2: **Araweté – classe menor na divisão em 2 classes**

Desde esta primeira divisão em duas classes, já fica patente que nesta língua há algo de próprio aos pronomes e marcas relacionais. Na divisão em quatro classes, em seguida, o algoritmo subdivide a classe 0 em três classes, conforme apresentamos a seguir; desta vez, o pronome de primeira pessoa fica isolado em uma classe para si, e o prefixo relacional, em outra.

Classe	Significado	Descrição da Abreviatura
0	1	Pronome de primeira pessoa
1	2	Pronome de segunda pessoa
1	3	Pronome de terceira pessoa
1	13	Pronome de primeira pessoa do plural exclusiva
1	23	Pronome de primeira pessoa do plural inclusiva
2	R1	Prefixo relacional 1

Tabela 3: **Araweté – classes menores na divisão em 4 classes**

Por sua vez, pedindo que o algoritmo separe os dados em oito classes, tem-se a divisão da classe maior de divisões prévias em duas classes, ambas extensas e de difícil interpretação. As demais seis classes consistem em três classes de apenas um morfema cada, e mais três classes apresentadas na Tabela 3. As seis classes resultantes – que não as duas maiores e de interpretação complicada por reunirem morfemas dos quais aparentemente não se extrai uma semântica comum – visualizam-se na Tabela 4 a seguir.

Classe	Significado	Descrição da Abreviatura
0	1	Pronome de primeira pessoa
1	2	Pronome de segunda pessoa
1	13	Pronome de primeira pessoa do plural exclusiva
1	23	Pronome de primeira pessoa do plural inclusiva
2	3	Pronome de terceira pessoa
3	R1	Prefixo relacional 1
4	R2	Prefixo relacional 2
5	FOC	Foco

Tabela 4: **Araweté – classes menores na divisão em 8 classes**

Seguiremos então para a aplicação do algoritmo aos dados das demais línguas contempladas no estudo. Para estas, apresentaremos diretamente a divisão dos morfemas em oito classes.

[8] Um dos pareceristas apontou que a ocorrência de morfemas zero na gramática, assim como a escolha de glosar morfemas cumulativos como morfemas únicos, poderia interferir em como os dados são analisados. Reconhecemos a importância desses problemas e estamos considerando formas de lidar com essas questões.

4.2 Yakkha

Yakkha é uma língua sino-tibetana, do ramo Kiranti, falada no Nepal e na região de Sikkim na Índia por cerca de 20 mil falantes. A fonte de dados utilizada é Schackow (2014), da qual retiramos 669 frases com 1028 morfemas.

Na divisão em oito classes pelo algoritmo de agrupamento, veem-se três classes de difícil interpretação, as três contendo a maior parte dos morfemas. Uma quarta classe, denominada Classe 0, contém 128 morfemas que são sistematicamente lexemas denotando eventos, conforme mostra-se a seguir.

Classe	Significado	Descrição da Abreviatura
0	eat	comer
0	give	dar
0	become	tornar-se
0	call	chamar
0	entangle	embaraçar
0	do	fazer
0	exist	existir
0	happen	acontecer
0	ask	perguntar
...	...	...

Tabela 5: **Yakkha – Classe 0, lexemas que denotam eventos**

As demais quatro classes contêm um morfema cada uma:

Classe	Significado	Descrição da Abreviatura
1	NMLZ.SG	Nominalizador singular
2	GEN	Genitivo
3	LOC	Locativo
4	NEG	Negação

Tabela 6: **Yakkha - 4 classes menores na divisão em 8 classes**

4.3 Pite Saami

A língua pite saami, ou lapônico de Pite, é da família fino-úgrica, sendo falada por cerca de 50 pessoas ao longo do rio Pite, na Suécia e Noruega. A fonte de dados utilizada é a gramática descritiva de Wilbur (2014). Extraímos as 241 frases da gramática, que resultaram em 567 diferentes morfemas.

Dentre as quatro línguas observadas, esta é a de mais difícil interpretação dos resultados, seja pelo menor número de frases na gramática, ou por possíveis idiossincrasias tipológicas da língua. Com a divisão pelo algoritmo, tem-se uma grande classe de 482 morfemas, que abarca, portanto, a maior parte dos morfemas. Dentre as demais classes, veem-se uma com 43 morfemas, composta majoritariamente por lexemas que denotam eventos, e outra de 35 morfemas sem nada de semelhante detectável. As demais cinco classes são de extensão menor e estão representadas na Tabela 7.

Classe	Significado	Descrição da Abreviatura
1	1SG.NOM	Pronome de primeira pessoa do singular nominativo
1	ACC.SG	Acusativo singular
1	ACC.PL	Acusativo plural

2	be|3SG.PRS	Ser, terceira pessoa do singular no presente
3	1SG.PRS	Primeira pessoa do singular no presente.
4	then	então
5	and	e

Tabela 7: **Pite Saami – 5 classes menores na divisão em 8 classes**

4.4 Khwarshi

Khwarshi é uma língua do Cáucaso, da família nakh-daguestaniana, falada na república do Daguestão, Rússia, por cerca de mil pessoas. Os dados foram retirados da gramática descritiva de Khalilova (2009): 1149 frases e 1446 morfemas diferentes.

Nas 8 classes resultantes, há uma grande e de difícil interpretação, contendo 1106 morfemas. As demais classes, por sua vez, são bastante claras. Das três classes seguintes em tamanho, uma apresenta apenas lexemas indicando entidades (nomes); outra, lexemas indicando eventos; e mais outra contendo, em sua maioria, as diversas marcas de caso da língua. Dentre as quatro classes restantes, de tamanho reduzido, há uma com as marcas de gênero (cinco classes nominais) e número (plural humano e não-humano), e outras três como um único membro cada. Estas últimas quatro classes são apresentadas a seguir:

Classe	Significado	Descrição da Abreviatura
0	I	Classe/Gênero I
0	II	Classe/Gênero II
0	III	Classe/Gênero III
0	IV	Classe/Gênero IV
0	V	Classe/Gênero V
0	HPL	Plural humano
0	NHPL	Plural não-humano
1	AND	e
3	ERG	Caso ergativo
4	PST.UW	Passado não-testemunhado

Tabela 8: **Khwarshi – 4 classes menores na divisão em 8 classes**

5 Discussão

A Tabela 9 mostra uma síntese das informações encontradas com a aplicação do algoritmo de agrupamento hierárquico nos dados. Abaixo temos a descrição de cada sigla a ser utilizada:

- **(NSI) Número de classes sem interpretação:** classes normalmente grandes, que parecem agrupar muitos elementos distintos.
- **(N1) Número de classes de um único membro:** classes com um único membro, independentemente de sua interpretação.
- **(N10) Número de classes de mais de 10 membros:** classes maiores, independentemente de sua interpretação.
- **(MC/TM) Proporção de morfemas em classes menores e regulares em relação ao número total de morfemas:** determina a proporção de dados agrupados segundo alguma regularidade.
- **(CD) Significados que formam classes distribucionais:** quais significados possuem uma distribuição regular na língua de forma que seja possível detectá-los pelo algoritmo.

Língua	NSI	N1	N10	MC/TM	CD
Araweté	2	5	2	1,9%	Pessoa, Relação, Foco
Yakkha	3	4	4	12%	Eventos, Nominalizador
Pite Saami	7	4	3	8,8%	Eventos
Khwarshi	1	3	4	23%	Eventos, Entidades, Gênero

Tabela 9: **Síntese da informação extraída a partir do Agrupamento Hierárquico**

Primeiramente, façamos a ressalva de estarmos assumindo que a quantidade de dados de cada língua, contanto que suficiente, não terá maior interferência no resultado. A proporção de morfemas em classes menores ou regulares (MC/TM) para o khwarshi é a maior de todas, e isso coincide com o fato de khwarshi ser a língua da qual extraimos mais frases. No entanto, a língua com a menor proporção, araweté, não é a língua com o menor número de frases. Desta forma, se o número de frases possui algum impacto nesses resultados, o impacto pode não ser forte o suficiente para que as observações feitas a seguir sejam descartadas de imediato.

O indicador MC/TM nos informa a proporção dos morfemas para os quais o algoritmo conseguiu ou esgotar as possibilidades de agrupamento (classes de um único grupo), ou encontrar classes significativas; por exemplo, um grande agrupamento de lexemas que denotem eventos. Sabendo que as classes são determinadas por propriedades distribucionais, uma alta proporção de MC/TM indica que a língua possui mais restrições com relação à distribuição de um número maior de morfemas e vice-versa, ou seja, divide seus morfemas em mais classes, não permitindo que estes se agrupem livremente, mas sim com demais morfemas que exibam determinadas propriedades semelhantes. Caso analisemos esta relação em conjunto ao número de classes referentes a lexemas, podemos entender que há uma tendência da língua a associar certos morfemas a certos tipos de denotação de lexemas. Por exemplo, em khwarshi, a razão MC/TM é alta e há duas classes grandes de lexemas, uma de eventos e outra de entidades. Provavelmente a língua apresenta uma morfologia específica para o que consideraríamos serem verbos e outra para o que consideraríamos serem nomes. No extremo oposto, araweté, não há nenhuma classe de lexemas por tipo de denotação e a proporção MC/TM é bastante baixa, indicando que a morfologia da língua não especifica classes como nomes, verbos e adjetivos.

De fato, em sua gramática, Solano (2009) descreve o araweté como sendo uma língua sem adjetivos e que não distingue formalmente verbos e nomes: ambos podem ocorrer com os mesmos morfemas e podem ocupar as mesmas funções sintáticas. Por outro lado, o khwarshi (KHALILOVA, 2009) é considerada uma língua com forte distinção entre nomes e verbos e as marcas de gênero encontradas são marcas de concordância. Nesta língua, verbos não concordam com pessoa, mas sim com os gêneros de seus argumentos.

Ainda sobre araweté, os morfemas relacionais, glosados como R1, R2 e R4 na gramática, são cruciais para o estabelecimento das relações entre elementos da oração em função de sua ordem. O mais frequente, R1, indica adjacência entre determinante e determinado (possuidor-possuido, objeto-verbo, etc.):

(3) Araweté (SOLANO, 2009: 2, 36)

a. *padidi-i* *r-awe*
banana-pau R1-folha
"folha de bananeira."

b. *he* *ku* *ne* *r-aʔaʔa*
1 FOC 2 R1-arranhar
"Eu te arranhei."

Yakkha e pite saami são línguas que também apresentam uma certa tendência à sistematicidade morfológica para com lexemas verbais. Saliente-se que em yakkha a maior parte

dos adjetivos são, na verdade, lexemas verbais nominalizados com um sufixo nominalizador. Além disso, o nominalizador também ocorre como estratégia de relativização; veja-se:

(4) Yakkha (SCHACKOW, 2014: 1,30)

a. *ci-há* *mangcwa*
ficar_frio-NMLZ.SG água
"água fria."

b. *ka-ya-na* *anusar*
dizer-PRET-NMLZ.SG de_acordo
"de acordo com o que você prometeu."

Os resultados do pite saami são de análise mais difícil. Apesar de termos encontrado por meio do algoritmo uma classe de lexemas que denotem eventos, a interpretação das demais sete classes escapa à argumentação. De toda sorte, é interessante observar como em três das quatro línguas há um certo grau de similaridade no comportamento de lexemas que denotam eventos.

CONSIDERAÇÕES FINAIS

O presente trabalho relata os primórdios de um campo de pesquisa potencialmente frutífero, com uma grande promessa de aplicações e trabalhos subsequentes. Duas perspectivas concretas, por exemplo, são desde já aventadas: a utilização de uma matriz ponderada, com valores de probabilidade atribuídos a cada par de vizinhos em vez dos valores binários 0 e 1; e a decomposição de valores singulares (SVD, *single value decomposition*) da matriz. Com tudo quanto exposto, intentamos demonstrar um pouco do que a utilização de dados de gramáticas descritivas aliadas a ferramentas estatísticas e algoritmos de aprendizado não-supervisionado permite que seja feito; acreditamos que, ao se seguirem os métodos indutivos tão caros à tipologia linguística, revelações importantes a respeito das línguas possam ser alcançadas, não somente no campo das classes de palavras, mas em seu funcionamento como um todo.

REFERÊNCIAS BIBLIOGRÁFICAS

BICKEL, Balthasar. Towards a multivariate typology of clause linkage, with particular reference to non-embedded structures. In: **International symposium on the grammar and pragmatics of complex sentences (LENCA-3), Tomsk**. 2006. p. 27-30.

BONDY, John Adrian et al. **Graph theory with applications**. London: Macmillan, 1976.

BOOIJ, Geert. **The grammar of words: An introduction to linguistic morphology**. Oxford University Press, 2012.

BROWN, Dunstan; CHUMAKINA, Marina. What there might be and what there is: An introduction to Canonical Typology. **Brown D./Chumakina M./Corbett GG (Hrsg.), Canonical Morphology & Syntax, Oxford**, p. 1-19, 2012.

CLARK, Alexander; LAPPIN, Shalom. Unsupervised learning and grammar induction. **The Handbook of Computational Linguistics and Natural Language Processing**, v. 57, 2010.

CORBETT, Greville G. Universals and features. In: **Universals of language today**. Springer, Dordrecht, 2009. p. 129-144.

CROFT, William. **Typology and universals**. Cambridge University Press, 2002.

DIXON, Robert MW; AIKHENVALD, Alexandra Y. (Ed.). **Word: A cross-linguistic typology**. Cambridge University Press, 2003.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. **The elements of statistical learning**. New York: Springer series in statistics, 2001.

GOLDSMITH, John. Probabilistic models of grammar: Phonology as information minimization. **Phonological Studies**, v. 5, p. 21-46, 2002.

GOLDSMITH, John A. Segmentation and Morphology. **The handbook of computational linguistics and natural language processing**, v. 57, 2010.

GOLDSMITH, John. Towards a new empiricism. **Recherches Linguistiquesa Vincennes**, v. 36, p. 9-36, 2007.

HARRIS, Zellig S. From morpheme to utterance. In: **Papers on Syntax**. Springer, Dordrecht, 1946. p. 45-70.

HASPELMATH, Martin. Comparative concepts and descriptive categories in crosslinguistic studies. **Language**, v. 86, n. 3, p. 663-687, 2010.

HASPELMATH, Martin. The indeterminacy of word segmentation and the nature of morphology and syntax. **Folia linguistica**, v. 45, n. 1, p. 31-80, 2011.

HASPELMATH, Martin; SIMS, Andrea D. **Understanding morphology**. Routledge, 2013.

KHALILOVA, Zaira et al. **A grammar of Khwarshi**. 2009. Tese de Doutorado. LOT, Netherlands Graduate School of Linguistics, Utrecht.

MALINOWSKY, Bronislaw. Coral Gardens and their Magic, vol. II: The Language of Magic and Gardening. 1933.

NASTASE, Vivi; MIHALCEA, Rada; RADEV, Dragomir R. A survey of graphs in natural language processing. **Natural Language Engineering**, v. 21, n. 5, p. 665-698, 2015.

SCHACKOW, Diana. **A grammar of Yakkha**. 2014. Tese de Doutorado. University of Zurich.
SCHÜTZE, Hinrich. Part-of-speech induction from scratch. In: **Proceedings of the 31st annual meeting on Association for Computational Linguistics**. Association for Computational Linguistics, 1993. p. 251-258.

SCHÜTZE, Hinrich. Distributional part-of-speech tagging. **arXiv preprint cmp-lg/9503009**, 1995.

SOLANO, Eliete de Jesus Bararuá. Descrição gramatical da língua Araweté. 2009. Tese de Doutorado, UnB.

VAN WYK, E. B. Notes on word autonomy. **Lingua**, v. 21, p. 543-557, 1968.
WILBUR, Joshua. **A grammar of Pite Saami**. Language Science Press. 2014.