DO MORFEMA PARA A PALAVRA: PADRÕES DE ASSOCIAÇÃO ENTRE MORFEMAS A PARTIR DE MÉTODOS ESTATÍSTICOS

João Paulo Lazzarini Cyrino [1]
Eudes Barletta Mattos [2]

RESUMO

O conceito de palavra tem sido rediscutido na linguística como um construto teórico não necessariamente válido, principalmente em línguas não indoeuropeias, calcado numa tradição gramatical que se ajustava mais às línguas clássicas que à totalidade – e inclusive maioria – das línguas do mundo. Com o intuito de examinar fenômenos de relevância tipológica que não dependam do conceito tradicional de palavra, bem como de tentar revelar padrões translinguísticos de relação entre morfemas, o presente estudo utiliza ferramentas do aprendizado de máquina não-supervisionado para processar dados captados de gramáticas descritivas. Organizando os dados em matrizes de adjacência que representam a ocorrência ou não de bigramas – associações de dois morfemas –, avaliamos como diferentes línguas organizam seu inventário de morfemas e associam itens deste. Para tal, aliamos a técnica Fatoração de Matriz Não-Negativa (NMF) à clusterização por K-Médias (K-Means), de maneira a agrupar os morfemas de dada língua em um número variável de classes; em seguida, as entradas dos morfemas foram substituídas pelo índice de suas respectivas classes, e as matrizes de adjacência, reanalisadas com atenção às associações entre as classes formadas, aplicando-se o teste Z para medir a significância das relações bigrâmicas língua a língua. Os resultados demonstraram grande variação no comportamento dos morfemas das línguas, assim como algumas correlações, de fracas a moderadas, entre uma maior proporção de bigramas significantes e a existência de classes distribucionais lexicais.

Palavras-chave: Tipologia Linguística, Aprendizado de máquina, Morfossintaxe, K-Médias, Fatoração de Matriz Não-Negativa.

INTRODUÇÃO

Recentemente há uma problematização na literatura tanto de Teoria como de Tipologia linguística a respeito do conceito de palavra. Essa problematização é bem sintetizada em Haspelmath (2011), no âmbito da Tipologia Linguística, que mostra não haver uma definição de palavra apta a ser aplicada translinguisticamente. Sendo assim, linguistas não deveriam presumir uma diferença entre morfologia e sintaxe – ao menos não ao comparar línguas – e deveriam orientar-se pela unidade mínima da gramática: o

[1] Professor do curso de Letras da Universidade Federal da Bahia - UFBA, jpcyrino@gmail.com;
[2] Graduando pelo Curso de Letras da Universidade Federal da Bahia - UFBA, eudesbarlettam@gmail.com;

morfema ou, mais recentemente, o morfe (cf. HASPELMATH, 2020). No âmbito da Teoria Gerativa, podemos também mencionar desenvolvimentos como a Morfologia Distribuída (HALLE & MARANTZ, 1993), que prevê transparência entre os níveis da Morfologia e Sintaxe, e que a estrutura sintática, ao menos em um nível mais abstrato, é fundamentalmente formada por morfemas.

Por outro lado, a Tipologia Linguística tem em sua tradição a classificação de línguas em termos de relação entre morfemas e palavras. Essa tradição, que remonta a trabalhos como os dos irmãos Schlegel – dividindo as línguas em *isolantes*, *aglutinantes* e *fusionais* –, ainda encontra eco em trabalhos tipológicos, como Bickel & Nichols (2007), que mostram haver parâmetros como *fusão* (o quanto morfemas tendem a se juntar em torno de uma base), *exponência* (quantos significados distintos podem ser carregados por um único morfema) e *flexividade* (o quanto de alomorfia há no sistema morfológico). Desses três parâmetros, o de *fusão* depende ou de algum conceito como palavra, ou de alguma métrica que permita entender o quanto um morfema está preso ao outro.

Algo que é também não é enfatizado nessa discussão é o fato de que algumas línguas aparentemente tendem a juntar determinados tipos de morfemas, a exemplo de morfemas de tempo/modo/aspecto, a morfemas com significados lexicais, formando unidades tentativamente delimitadas e que se agrupam no que conhecemos por classes de palavras; é dizer, muito do que entendemos por palavra está relacionado ao resultado de diferentes padrões de junção. Matthews (2003) mostra, por exemplo, que o conceito de palavra que utilizamos nas gramáticas é oriundo dos gramáticos latinos – e, em latim, há grande regularidade na junção de determinados morfemas com determinados significados lexicais.

Aqui buscamos estabelecer uma métrica para a junção entre morfemas das línguas. Inspirados nas técnicas de aprendizado de máquina não supervisionado propostas em Schütze (1995) para obter as classes das palavras de um *corpus*, desenvolvemos um método para, a partir dos dados morfologicamente segmentados de uma gramática descritiva, classificar automaticamente os morfemas de acordo com sua distribuição local (vizinhos à esquerda e à direita). A partir dessa classificação, observamos o quanto os morfemas das diferentes classes tendem a se juntar. Isso é feito por um teste estatístico de proporção (teste Z), que mede o quão estatisticamente significativa é a coocorrência de dois morfemas (bigramas), um antecessor e outro sucessor.

Para este trabalho buscamos estabelecer correlações entre a proporção destes pares de morfemas na língua e o quanto as classes de morfemas são semanticamente identificáveis. As correlações encontradas não são fortes, mas permitem levantar algumas hipóteses interessantes.

Realizamos o estudo com 15 línguas de diferentes famílias e de diferentes partes do mundo:

- *khwarshi e aghul*, línguas do nordeste do cáucaso, faladas no Daguestão (Rússia). A primeira é do ramo tsézico, e a segunda do ramo lezgiano.
- *moloko e bathari,* línguas afro-asiáticas, sendo a primeira do ramo chádico e a segunda semítica. Faladas respectivamente nos Camarões e em Omã.
- *mandarim e yakkha*, línguas sino-tibetanas, sendo a primeira do ramo sinítico, falada na China, e a segunda do ramo kiranti, falada no Nepal.
- *nenets e pite saami*, línguas urálicas, a primeira samoiédica e a segunda fínica, faladas, respectivamente, na Sibéria (Rússia) e no interior da Suécia.
- *rapa nui e cheke holo*, línguas austronesianas, do ramo malaio-polinésio. A primeira é falada na Ilha da Páscoa e a segunda nas ilhas Solomon.
- *araweté*, língua indígena brasileira, do tronco tupi, falada no estado do Pará.
- *kanoê*, língua indígena brasileira, isolada, falada no estado do Amazonas.
- *kuot*, língua isolada, falada na Papua Nova Guiné.
- *sumério*, língua isolada, extinta, falada no sul da Mesopotâmia antiga.
- *zapoteca de San Bartolomé Zoogocho*, língua da família oto-mangueana, falada na região de San Bartolomé Zoogocho, México.

O presente texto está organizado da seguinte forma. Primeiramente desenvolvemos os aspectos metodológicos, divididos em três etapas: coleta de dados, estabelecimento dos hiperparâmetros dos algoritmos de análise e aplicação de teste de significância estatística. Em seguida, apresentamos os resultados, mostrando (i) a proporção de bigramas significativos em cada língua e (ii) a possível correlação entre as métricas e a semântica das classes de morfemas obtidas. Encerramos, então, com as Considerações Finais.

METODOLOGIA

Este trabalho foi realizado em 3 diferentes etapas, cada qual com uma metodologia particular: (i) coleta de dados, (ii) estabelecimento dos hiperparâmetros dos algoritmos de análise, (iii) aplicação de teste de significância estatística para os bigramas obtidos.

Coleta de Dados e Matrizes de Adjacência

Os dados foram coletados a partir de 15 gramáticas descritivas de diferentes línguas. Como este trabalho não pretende mostrar correlações entre parâmetros tipológicos nem estabelecer universais, não houve uma metodologia de amostragem específica para as línguas escolhidas, apesar de haver diversidade entre elas em termos genéticos, geográficos e mesmo históricos (por exemplo, figuram dados de uma língua sem falantes vivos há milênios, como o sumério). Coletamos, a partir de cada gramática, o número de frases apresentado na Tabela 1. Devido à grande disparidade entre estes números, utilizamos um algoritmo para escolher, aleatoriamente e sem reposição, 200 frases de cada língua que contasse com mais de 200 frases no banco de dados. Para as línguas com menos de 200 frases, utilizamos o conjunto de frases já disponíveis em sua inteireza.

Língua	**Gramática**	**Frases**
Aghul	Maslak (2014). *Агульские тексты 1900—1960-х годов*[3]	191
Araweté	Solano (2009). *Descrição Gramatical da Língua Araweté*	563
Bathari	Gasparini (2018). *The Baṭḥari Language of Oman: Towards a Descriptive Grammar*	79
Cheke Holo	Boswell (2018). *A grammar of Cheke Holo*	108
Kanoe	Bacelar (2004). *Gramática da Língua Kanoê*	934
Khwarshi	Khalilova (2009). *A grammar of Khwarshi*	1149
Kuot	Lindström (2002). *Topics in the grammar of Kuot*	81
Mandarim	Sheng Ma (2006). *Modern Mandarin Chinese Grammar*[4]	213
Moloko	Friesen et al (2017). *A Grammar of Moloko*	181
Nenets	Nikolaeva (2014). *A Grammar of Tundra Nenets*	130
Pite Saami	Wilbur (2014). *A Grammar of Pite Saami*	241
Rapa Nui	Kieviet (2017). *A Grammar of Rapa Nui*	221
Sumério	Jagersma (2010). *A descriptive grammar of Sumerian*	379
Yakkha	Schackow (2015). *A grammar of Yakkha*	669

[3] Transliteração do russo: *Agulskie Teksty 1900-1960-ch godov*. Tradução: *Textos em Aghul dos anos 1900 a 1960.*
[4] Com glosas nossas.

Zapoteca de SBZ[5]	Sonnenschein (2004). *A descriptive grammar of San Bartolomé Zoogocho Zapotec*	387

Tabela 1 – Fontes de Dados e Números de Dados

As gramáticas descritivas normalmente fornecem os dados com segmentação morfológica e a glosa morfema a morfema. O presente estudo se baseou na análise distribucional das glosas de cada morfema, desconsiderando, portanto, casos de alomorfia. Apenas os dados do mandarim foram retirados de uma gramática cujas frases não apresentavam segmentação morfológica nem glosas; para adequação de seus dados, portanto, esta análise foi realizada no decorrer da pesquisa[6].

Munidos das respectivas glosas de cada frase de uma dada língua, listamos todos os bigramas (pares ordenados de dois morfemas, no caso) e os organizamos em uma matriz de adjacência. Este procedimento foi repetido para todas as línguas elencadas. As linhas da matriz correspondem a cada morfema, e as colunas, a cada vizinho anterior e posterior ao morfema, conforme ilustrado na Tabela 2. Os valores dessa matriz são zero (0) para a não existência do par nos dados e um (1) para a existência[7]. A construção dessas matrizes conclui a fase de coleta dos dados.

	Morfema 1_dir	**Morfema 2_dir**	**...**	**Morfema n_esq**
Morfema 1	0	0	...	1
Morfema 2	0	0	...	0
Morfema 3	0	1	...	0
...	...	...	...	..
Morfema n	1	1	...	0

Tabela 2 – Exemplo de Matriz de Adjacência

Fatoração da Matriz de Adjacência e *Clustering*

A segunda etapa da pesquisa consistiu em aplicar métodos de aprendizado de máquina não-supervisionado que agrupem em classes os morfemas mais semelhantes entre si (por critério distribucional). Para este estudo, utilizamos a Fatoração de Matriz Não-Negativa (doravante, NMF, de *Non-Negative Matrix Factorization*) seguida da

[5] Zapoteca de San Bartolomé Zoogocho
[6] A segmentação morfológica e subsequente descrição dos dados do mandarim foram realizadas por Eudes Barletta Mattos.
[7] Foram tentados outros valores para a matriz, como a frequência das adjacências, assim como a probabilidade. Esses métodos parecem privilegiar o agrupamento dos morfemas em termos semânticos, mais do que gramaticais. Um comparativo a respeito disso está em preparação.

clusterização por K-Médias (doravante, K-Means, de *K-Means Clustering*). Foram empregados os algoritmos disponíveis no pacote de aprendizado de máquina *scikit-learn* (PEDREGOSA et al, 2011), para a linguagem Python[8].

Tais algoritmos requerem que forneçamos o número de agrupamentos em que queremos dividir o total de morfemas. Uma vez que não sabemos de antemão em quantas classes os morfemas de uma dada língua podem se dividir, utilizamos os seguintes métodos.

A ação do algoritmo NMF consiste em fatorar a matriz de adjacência obtida na primeira etapa, de forma a se obterem duas matrizes W e H; da multiplicação destas duas resulta uma aproximação da matriz de adjacência. Trata-se de uma técnica bastante utilizada em processamento de linguagem natural (PNL), uma vez que ela permite classificar textos em tópicos a partir da frequência das palavras. Considerando estudos em PNL, o usual é construir-se uma matriz em que cada linha corresponda a um documento e cada coluna a uma palavra, e os valores da matriz correspondam à frequência de cada palavra no documento. Com NMF, torna-se possível agrupar documentos com frequências semelhantes de determinadas palavras. No entanto, é necessário fornecer ao algoritmo, de antemão, o número de agrupamentos – componentes – em que se deseja dividir os documentos. Portanto, para estudos que envolvam análise de documentos, o critério para definir o número de componentes pode ser relativamente simples: se houver misturados documentos sobre cinco assuntos, basta que sejam escolhidos cinco componentes para NMF. Após a aplicação, a matriz obtida terá uma linha para cada documento, mas ao invés de uma coluna para cada palavra, passará a ter uma coluna para cada componente. Cada documento terá um escore diferente para cada componente e documentos de cada assunto terão, em tese, um escore maior no componente respectivo a cada assunto.

No entanto, o que seria um documento em estudos de PNL equivale a um morfema neste estudo, e o que seria uma palavra em um dado documento, ao vizinho daquele morfema. Neste caso, para determinar o número de componentes – ou seja, o número de agrupamentos a ser feitos –, foram efetuadas diferentes tentativas, indo de dois até seis agrupamentos. Utilizamos o número de componentes em que mais víamos ser possível haver uma semântica subjacente associada a cada agrupamento.

[8] Informações sobre NMF em: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html; sobre K-Means em: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Uma vez aplicada a fatoração à matriz de adjacência, torna-se possível comparar sua variância com a variância da matriz fatorada. A Tabela 3 fornece o número de componentes que utilizamos para cada língua e a variância estimada (V.E.) da matriz fatorada em relação à matriz original.

Língua	**Componentes**	**V.E.**	**Clusters**
Aghul	3	0.127	6
Araweté	5	0.348	6
Bathari	4	0.261	5
Cheke Holo	4	0.247	6
Kanoe	3	0.157	6
Khwarshi	5	0.201	6
Kuot	4	0.266	5
Mandarim	4	0.185	5
Moloko	4	0.219	6
Nenets	3	0.125	6
Pite Saami	2	0.075	5
Rapa Nui	5	0.218	6
Sumério	2	0.106	6
Yakkha	4	0.129	6
Zapoteca de SBZ[9]	5	0.267	6

Tabela 3 – Componentes, Variância Estimada e Clusters

Com a matriz fatorada, utilizamos o algoritmo de K-Means para agrupar em classes os morfemas. O algoritmo agrupa os morfemas que possuam valores semelhantes nos escores da matriz obtida por NMF. K-Means também requer que informemos o número de *clusters* (agrupamentos). Diferentemente de NMF, no entanto, há um método para obter esse número, denominado método *Elbow*[10]. Os agrupamentos obtidos estão relacionados na Tabela 3. A obtenção dos *clusters* encerra a segunda etapa.

Substituição dos Morfemas pelas Classes e Teste Z

Com a classificação de cada morfema realizada, podemos substituir nos dados a entrada de cada morfema por sua respectiva classe. Assim foi feito para todas as línguas, de modo que uma lista de bigramas a partir dessas novas combinações foi obtida. Essa

[9] Zapoteca de San Bartolomé Zoogocho
[10] Informações sobre o método em Kodinariya & Makwana (2011).

lista deu origem a uma tabela de contingência – que conta o número de combinações entre cada primeiro e cada segundo elemento possível em um bigrama. Ilustra a Tabela 4, feita a partir da classificação dos morfemas da língua nenets; nesta, cada linha corresponde à classe do primeiro morfema do bigrama e cada coluna, à classe do segundo. Nota-se que não nomeamos as classes, apenas nos referindo a elas por um índice iniciado em 0. No caso do nenets, são 6 classes obtidas, de 0 a 5.

	0	**1**	**2**	**3**	**4**	**5**
0	60	4	6	3	18	0
1	9	0	1	0	4	0
2	1	2	1	9	0	8
3	7	1	2	0	3	0
4	16	9	1	8	7	12
5	2	0	2	0	9	0

Tabela 4 – Tabela de contingência de bigramas em nenets

Como podemos observar na tabela, temos combinações de classes de morfemas bastante frequentes na língua, como entre dois morfemas representados pelo par (0,0), com 60 ocorrências, combinações menos frequentes, como o par (3,4), com 3 ocorrências, e combinações não atestadas, como (3,3) ou (1,3).

Diante desses dados, podemos finalmente responder à pergunta: existem classes de morfemas com mais afinidades para suceder ou anteceder outras classes? Para tanto, procedamos da seguinte forma. Considerando um bigrama *(i,j)*, assumimos que não há afinidade significativa entre *i* e *j* se a quantidade de *j* atestados para um dado *i* for menor que duas vezes a média de ocorrências atestadas na posição _ em um bigrama (*i*,_), ou se a quantidade de *i* atestados para todo *j* for menor que duas vezes a média de ocorrências atestadas na posição _ em um bigrama (_, *j*).

	0	**1**	**2**	**3**	**4**	**5**
0	60	4	6	3	18	0
1	9	0	1	0	4	0
2	1	2	1	9	0	8
3	7	1	2	0	3	0
4	16	9	1	8	7	12
5	2	0	2	0	9	0

Tabela 5 – Destaque para o par (0,4)

A Tabela 5 destaca um caso da língua nenets. Para saber se os morfemas da classe 0 antecedem os de classe 4 sistematicamente, calculamos a média dos valores na linha 0

(≈ 15,17) e a média dos valores na coluna 4 (≈ 6,83). Embora o valor para (0,4) seja 18, inferior ao dobro da média da linha 0, é superior ao dobro da média da coluna 4. Isso significa que, de acordo com o critério proposto aqui, há uma sistematicidade na associação entre as classes 0 e 4 na ordem (0,4).

Para estabelecer margens de erro, aplicamos dois testes Z[11], um em que a hipótese nula H_0 é a de que o valor em (*i*,*j*) é menor do que duas vezes a média da linha *i* e o outro em que é menor do que duas vezes a média da coluna *j*. Disto temos dois valores-p; para um dado nível de significância $\alpha = 0,05$, temos uma dupla associação significativa entre (i,j) se ambos os valores-p forem menores do que 0,05 e uma associação significativa se apenas um dos valores-p for menor do que 0,05. No caso de (0,4), temos os seguintes valores-p: $p_1 \approx 0,056$ e $p_2 \approx 0,006$. Desta forma, (0,4) é significativo em função do valor p_2, que se refere à classe 4 suceder a classe 0, mas não é significativo em função do valor p_1, que se refere à classe 0 anteceder a classe 1. Pode-se afirmar, portanto, que há uma associação significativa em (0,4).

De porte de todas estas métricas, comparemos na seção seguinte as línguas em função das associações entre as classes de morfemas, em diferentes níveis de significância. Também vamos examinar o que está contido em cada classe e tentar ver se existe uma semântica que possa caracterizar cada classe e as combinações entre as classes.

RESULTADOS E DISCUSSÃO

Nesta seção apresentamos os resultados sumarizados para as línguas e não olhamos especificamente para cada morfema. Nesse sentido, utilizamos as métricas explicadas na seção anterior de forma a obter indicadores holísticos sobre a morfologia da língua. Os resultados apresentados revelam algumas características sobre as línguas, mas também possuem limitações, conforme se discute.

Aglomerações Significativas de Morfemas

A Tabela 6 contém os bigramas para um nivel de significância $\alpha = 0,10$ e a Tabela 7, para um nível de significância $\alpha = 0,05$. É importante notar que a língua Kanoê é a

[11] O teste Z é um dos testes utilizados para rejeitar ou não uma hipótese determinando se a diferença entre a média da amostra e a média da população é significativa estatisticamente ou fruto do acaso, destacando-se por possuir um único valor crítico para as amostras.

única a não apresentar bigramas significativos em α = 0,05. Durante a pesquisa notamos que isso se preserva com α = 0,01, embora não abordemos esses níveis aqui.

Língua	cl	big	ant	suc	dupla	prop_big	prop_ant	prop_suc	prop_dupla
Aghul	6	5	3	2	0	0,139	0,083	0,056	0,000
Araweté	6	11	8	6	3	0,306	0,222	0,167	0,083
Bathari	5	5	5	3	3	0,200	0,200	0,120	0,120
Cheke Holo	6	4	1	3	0	0,111	0,028	0,083	0,000
Kanoe	6	5	3	4	2	0,139	0,083	0,111	0,056
Khwarshi	6	13	8	7	2	0,361	0,222	0,194	0,056
Kuot	5	9	6	5	2	0,360	0,240	0,200	0,080
Mandarim	5	5	3	2	0	0,200	0,120	0,080	0,000
Moloko	6	11	5	6	0	0,306	0,139	0,167	0,000
Nenets	6	11	7	6	2	0,306	0,194	0,167	0,056
Pite Saami	5	7	3	5	1	0,280	0,120	0,200	0,040
Rapa Nui	6	11	6	6	1	0,306	0,167	0,167	0,028
Sumério	6	11	4	8	1	0,306	0,111	0,222	0,028
Yakkha	6	9	5	5	1	0,250	0,139	0,139	0,028
Zapoteca SBZ	6	8	5	7	4	0,222	0,139	0,194	0,111

Tabela 6 – Bigramas para α = 0,10

Língua	cl	big	ant	suc	dupla	prop_big	prop_ant	prop_suc	prop_dupla
Aghul	6	4	2	2	0	0,111	0,056	0,056	0,000
Araweté	6	10	7	5	2	0,278	0,194	0,139	0,056
Bathari	5	4	4	3	3	0,160	0,160	0,120	0,120
Cheke Holo	6	3	0	3	0	0,083	0,000	0,083	0,000
Khwarshi	6	12	7	7	2	0,333	0,194	0,194	0,056
Kuot	5	7	4	5	2	0,280	0,160	0,200	0,080
Mandarim	5	4	3	1	0	0,160	0,120	0,040	0,000
Moloko	6	9	3	6	0	0,250	0,083	0,167	0,000
Nenets	6	9	6	5	2	0,250	0,167	0,139	0,056
Pite Saami	5	6	3	4	1	0,240	0,120	0,160	0,040
Rapa Nui	6	11	6	6	1	0,306	0,167	0,167	0,028
Sumério	6	9	3	7	1	0,250	0,083	0,194	0,028
Yakkha	6	6	3	4	1	0,167	0,083	0,111	0,028
Zapoteca SBZ	6	8	5	7	4	0,222	0,139	0,194	0,111

Tabela 7 – Bigramas para α = 0,05

A coluna *cl* refere-se ao número de classes (*clusters*) de morfemas que foram encontrados pelo método *Elbow*; a coluna *big*, ao número de bigramas significativos encontrados nos dados segundo qualquer um dos testes Z aplicados; a coluna *ant*, ao número de bigramas significativos segundo o teste Z quanto ao primeiro morfema do par

ser antecessor do segundo, e a coluna *suc* refere-se ao teste quanto ao segundo morfema do par ser sucessor do primeiro. A coluna *duplo* refere-se ao número de bigramas significativos segundo os dois testes. Por fim, as colunas iniciadas por *prop_* trazem a proporção de bigramas significativos segundo cada categoria (*big*, *ant*, *suc* e *duplo*). Esta proporção é o número de morfemas das respectivas colunas divido pelo quadrado do número de classes de morfemas da língua – c*l* –, e é decorrente de estarmos considerando a proporção de bigramas significativos em relação aos bigramas logicamente possíveis na língua.

Segundo a estatística *prop_big*, a língua com a maior proporção de bigramas significativos é *khwarshi*, para ambos os níveis de significância (0,361 para $\alpha = 0,10$; 0,333 para $\alpha = 0,05$). Como a estatística se refere a bigramas de classes de morfemas e não somente de morfemas, podemos interpretá-la como um indicativo de que as classes capturam o comportamento dos morfemas da língua e a língua tende a juntar morfemas de forma mais regular. Ao contrário, *cheke holo* é a língua com a menor proporção de bigramas significativos (0,111 para $\alpha = 0,10$; 0,083 para $\alpha = 0,05$). Nesse caso teríamos uma tendência à negativa lógica do que esperamos com o *khwarshi*: ou as classes não capturam o comportamento dos morfemas na língua, ou a língua não tende a aglomerar morfemas de forma mais regular. No primeiro caso, as classes distribucionais não seriam suficientes para descrever o comportamento dos morfemas da língua e no segundo, os morfemas podem se combinar de forma mais livre.

Por conta dessa disjunção para valores mais baixos, essa estatística sozinha não é muito útil para uma representação holística do comportamento da morfologia da língua. Uma forma de trazer mais interpretabilidade para esses números seria compreender que tipos de morfemas foram agrupados pelo algoritmo de *clustering*.

Classes e sua semântica

Ao observar os morfemas que formam as diferentes classes de cada língua, verificamos a ocorrência de três tipos diferentes de classes. O primeiro tipo são classes formadas por morfemas lexicais. Todas as classes desse tipo agrupam, em sua maior parte, morfemas que denotam eventos, sugerindo que há uma morfologia verbal delimitada na língua. Das 15 línguas observadas, 7 possuem uma classe desse tipo.

O segundo tipo de classe são as formadas por morfemas gramaticais, ou seja, conseguimos encontrar majoritariamente morfemas como de caso, tempo/modo/aspecto,

concordância, gênero, classificadores, nominalizadores, etc. Estas classes nos permitem entender o que é sistematicamente marcado na língua. A única língua em que não se isolam com clareza morfemas gramaticais é *pite saami*. As demais possuem de 2 a 4 classes de morfemas gramaticais.

O terceiro tipo de classe são as formadas por morfemas gramaticais e lexicais. São classes de elementos variados, portanto, é difícil entender se elas são um resultado da má aplicação dos algoritmos de clustering, do fato de termos utilizado um conjunto reduzido de dados (normalmente algoritmos para PLN dependem de *corpora* muito mais extensos), ou simplesmente são resultado da falta de regularidade da língua (flexibilidade na combinação de morfemas).

A Tabela 8 mostra uma visão geral destas classes nas línguas, mostrando o número de classes Lexicais, Gramaticais e Variadas encontradas em cada uma e também uma descrição dos significados Lexicais (sempre eventos) e Gramaticais agrupados nas respectivas classes.

Língua	Classes	Lexicais	Gramaticais	Variadas	Lexical	Gramatical
Pite Saami	5	0	0	5		
Nenets	6	0	3	3		Essivo/Genitivo, 1SG, 3SG
Zapoteca SBZ	6	1	2	3	Eventos	Aspecto, Pessoa
Mandarim	5	0	2	3		Nominalizador, Pronome
Aghul	0	0	3	3		Caso, Tempo/Aspecto, Modal
Rapa Nui	6	0	4	2		Artigo, Locação, Aspecto, Caso
Moloko	5	0	4	2		Pessoa, 3, Aspecto, Clítico
Bathari	5	0	2	2		Pronome, Artigo, Preposição
Cheke Holo	5	0	3	2		Pronome, Locativo, Demonstrativo
Khwarshi	6	1	4	1	Eventos	Caso, Gênero, Tempo/Aspecto
Kanoe	6	2	3	1	Eventos	Declarativo, Pronome, Concordância
Kuot	5	1	3	1	Eventos	3, 3S, Enfático/etc.
Araweté	6	1	4	1	Eventos	Relacional, Foco, Pronome, Modal
Sumério	6	1	4	1	Eventos	Casos, Pronomes
Yakkha	6	1	4	1	Eventos	Nominalizador, Tempo, Caso, Aspecto

Tabela 8 – Tipos de Classes Encontrados

Diante desses números, retomemos o que foi dito anteriormente de que uma alta proporção de bigramos significantes na língua *(prop_big)* estaria relacionda a (i) as classes distribucionais capturarem o comportamento dos morfemas na língua e (ii) a tendência da língua em aglomerar morfemas de forma regular. Disto podemos entender que haveria uma correlação negativa entre *prop_big* e o número de classes variadas encontradas e uma correlação positiva entre *prop_big* e o número de classes lexicais encontradas.

Medimos essa correlação utilizando o coeficiente de correlação de Pearson. Tal coeficiente varia de 1 a -1, sendo 0 um indicativo de ausência de correlação entre as variáveis, 1 um indicativo de forte correlação positiva e -1 um indicativo de forte correlação negativa. Calculamos a correlação entre *prop_big* e o número de classes variádas e classes lexicais. Também calculamos a correlação entre os dois tipos de classes e *prop_dupla*, estimando que uma maior proporção de bigramas duplamente significativos pudesse também ter relação com os tipos de classe encontrados. Os resultados estão na Tabela 9, para os dois níveis de significância de *prop_big* e *prop_dupla* obtidos.

	α = 0,10		**α = 0,05**	
	prop_big	**prop_dupla**	**prop_big**	**prop_dupla**
Classes variadas	-0,20414375	-0,137738836	-0,255304	-0,07610464
Classes lexicais	0,08456529	0,370348655	0,4175945	0,37770078

Tabela 9 – Coeficientes de Correlação de Pearson

O que encontramos de sistemático entre os dois níveis de significância é uma correlação negativa fraca entre *prop_big* e classes variadas e uma correlação positiva moderada para *prop_dupla* e classes lexicais. Cabe mencionar também que há uma correlação positiva moderada entre *prop_big* e classes lexicais para o nível de significância α = 0,05. Isso provavelmente resulta de uma distorção provocada pelos dados do kanoê, que possuem duas classes lexicais (constituídas de eventos) mas *prop_big* baixo. No nível de significância α = 0,05 não há bigramas significativos para o Kanoê, fazendo com que a correlação salte de aproximadamente 0,08 para 0,42.

Cabe mencionar que estas medidas de correlação, para esse conjunto de dados, são de caráter exploratório. Como as línguas estudadas são relativamente poucas, é natural que não seja possível realizar uma boa inferência e que haja distorções, como a resultante da saída dos dados do kanoê.

Voltando aos dados da Tabela 8, vemos que as duas das línguas com o maior número de classes variadas são pite saami e nenets. Essas duas línguas não possuem o *prop_big* baixo. Aliás, essas línguas são conhecidas por terem a morfologia bastante complexa, com muitas classes flexionais. Sucede que essas línguas aparentam ter muitos casos de supleção e, portanto, muitas das informações gramaticais das palavras mais frequentes estão fundidas à raiz. Dessa forma, as classes distribucionais, da forma como estão calculadas, por segmentação morfológica, não são suficientes para capturar o que ocorre com a língua. Outro caso que pode provocar ruídos é o que ocorre com moloko e bathari; estas duas línguas possuem boa parte da flexão realizada por infixos (tal qual é característico das línguas afro-asiáticas).

CONSIDERAÇÕES FINAIS

Os números sugerem bastante variação no comportamento dos morfemas das línguas e algumas correlações de fracas a moderadas entre a maior proporção de bigramas significantes em uma língua e a existência de classes distribucionais lexicais e um menor número de classes variadas. Apesar de demandarem que o estudo se replique em mais línguas para que se confirmem, são indicativos interessantes de se observar uma vez que revelam que o modelo de classes distribucionais adotado pode ser falho com línguas de alta supleção (nenets e pite saami) ou com um sistema de infixos (bathari e moloko).

Uma hipótese não mencionada mas que será levada para estudo futuro é a de que o tipo de relação estabelecida entre os dois morfemas em um bigrama pode desfavorecer um indicativo de distribuição regular. Concretamente, se temos alta incidência de bigramas (a,b) e (b,a), temos uma relação reflexiva, indicando que as classes *a* e *b* são permutáveis. Há casos desses nas línguas observadas, mas seu impacto não foi ainda medido. A relação transitiva entre classes também deve ser estudada: alta incidência de bigramas como (a,b) e (b,c) pode ser indicativo de distribuição regular e formação de aglomerações como palavras. Em khwarshi isso é bastante comum: morfemas de gênero estão sempre acompanhados de um morfema lexical e o morfema lexical está sempre acompanhado de morfema de tempo/aspecto. Isto é o que forma o “verbo” em khwarshi. Outras línguas não possuem relações transitivas tão significativas, o que pode sugerir menor regularidade na distribuição dos morfemas.

Enfim, o estudo trouxe um tipo de métrica e um caso de aplicação. Cabe estendê-la para outras aplicações, buscando novas correlações que venham a corroborar ou não esse tipo de abordagem.

AGRADECIMENTOS

Aos demais membros do Laboratório de Tipologia Linguística da UFBA, pela colaboração na etapa de coleta de dados: Ricardo Potozky, André Cardoso, Joseane Oliveira e Jeferson Barbosa. À Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) pelo financiamento do projeto a nível de iniciação científica.

REFERÊNCIAS

BACELAR, Laércio Nora. **Gramática da língua Kanoê**. Nijmegen:[Sn], 2004.

BICKEL, Balthasar; NICHOLS, Johanna. Inflectional morphology. **Language typology and syntactic description**, v. 3, n. 2, p. 169-240, 2007.

BOSWELL, Fredrick Alvin. **A grammar of Cheke Holo**. LOT, 2018.

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

FRIESEN, Dianne. **A grammar of Moloko**. Language Science Press, 2017.

GASPARINI, Fabio. **Baṭḥari Language of Oman: Towards a Descriptive Grammar**. 2018. Tese de Doutorado. Università degli studi di Napoli L'Orientale.

HALLE, Morris. A. Marantz.(1993).'Distributed Morphology and the pieces of inflection'. **The view from building**, v. 20, p. 111-76, 1993.

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. The morph as a minimal linguistic form. **Morphology**, v. 30, n. 2, p. 117-134, 2020.

JAGERSMA, Bram. **A descriptive grammar of Sumerian**. Universiteit Leiden, 2010.

KHALILOVA, Zaira. **A grammar of Khwarshi**. Netherlands Graduate School of Linguistics, 2009.

KIEVIT, Paulus. **A grammar of Rapa Nui**. Language Science Press, 2017.

KODINARIYA, Trupti M.; MAKWANA, Prashant R. Review on determining number of Cluster in K-Means Clustering. **International Journal**, v. 1, n. 6, p. 90-95, 2013.

LINDSTRÖM, Eva. **Topics in the grammar of Kuot, a non-Austronesian language of New Ireland, Papua New Guinea**. 2002. Tese de Doutorado. Department of Linguistics, Stockholm University.

ШИХАЛИЕВА, Сабрина Ханалиевна. Рецензия на монографию ТА Майсака" Агульские тексты 1900-1960-х годов". **Гуманитарный вектор**, v. 11, n. 5, p. 198-200, 2016.

NIKOLAEVA, Irina. **A grammar of Tundra Nenets**. Walter de Gruyter GmbH & Co KG, 2014.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, v. 12, p. 2825-2830, 2011.

SCHACKOW, Diana. **A grammar of Yakkha**. Language Science Press, 2015.

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

MA, Jing-heng Sheng; ROSS, Claudia. **Modern Mandarin Chinese grammar: a practical guide. Hauptbd**. Routledge, 2006.

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

SONNENSCHEIN, Aaron Huey. **A descriptive grammar of San Bartolomé Zoogocho Zapotec**. University of Southern California, 2004.

WILBUR, Joshua. **A grammar of Pite Saami**. Language Science Press, 2014.