Ser necessário fazer e ordenar fazer:

classificadores gêneros literários por meio de classes verbais do grego

Caio B. A. Geraldes*

FFLCH–USP

caio.geraldes@usp.br

23 de setembro de 2023

Resumo

Neste trabalho, utilizo métodos de processamento de linguagem natural e quantitativos para produzir evidência formal para esta hipótese de distribuição de classes verbais entre gêneros literários. Utiliza-se o modelo Naïve-Bayes de classificador para verificar se a seleção lexical de verbos é suficientemente diferente entre os gêneros historiográfico e filosófico e são extraídos os coeficientes de peso dos verbos de interesse na classificação dos gêneros para observar se as classes verbais semânticas possuem efeito significativo na classificação. O resultado positivo permite considerar a associação entre autoria e gênero e atração de caso espúria e causada pela associação entre classe verbal e gênero literário.

Palavras-chave: semântica; gêneros literários; métodos quantitativos; grego antigo

1 Introdução

Em trabalhos anteriores (GERALDES, 2020, 2021), mostrei que atração de caso infinitiva em grego clássico, isto é, a concordância de caso entre objeto indireto da matriz e predicado secundário de uma oração infinitiva (exemplificadas em (1), sendo (1-a) um exemplo sem a concordância de caso e (1-b) com), está correlacionada com os seguintes fatores: (a) distância entre controlador e alvo; quanto menor, mais frequente; (b) classe de verbo infinitivo; a atração ocorrendo preferencialmente com cópulas; (c) classe do verbo matriz, preferencialmente com verbos com sentido modal/deôntico; (d) autoria, sendo a atração mais frequente em Platão do que em Xenofonte e praticamente inexistente em Heródoto.

*Este projeto foi financiado pela FAPESP por meio dos processos de número 2017/23334-2, 2019/18473-9 e 2021/06027-4. Este *paper* resulta do trabalho final da disciplina FLL5133 *Linguística Computacional*. Agradeço a Marcos Lopes (USP), Martina Rodda (Univerisity of Oxford) e Richard McElreath (Max-Planck-Institut für evolutionäre Anthropologie) pelas discussões sobre os métodos e a interpretação dos dados. Quaisquer falhas são, naturalmente, da minha parte. Todos os dados e o código utilizado neste trabalho estão disponíveis em https://github.com/caiogeraldes/2023sbec.

(1) a. symbo:léw-e: tô:j Ksenoph̊ô:nti elthónt-a e:s delpho:s anakojnoô:s-aj
aconselha-3SG X.DAT.SG.M indo-ACC.SG.M para-Delfos interrogar.INF
tô:j theô:j peri tê:s poré:as
o-deus.DAT.SG sobre-a-viagem
Ele aconselha Xenofonte ir a Delfos interrogar o deus sobre a viagem. (Xen. Anab. 3.1.5)

b. aphê:k-e moj elthónt-i pros hymâs légein
permitiu-3SG PRON(1SG.DAT.SG) indo-DAT.SG.M frente-a-vós dizer.INF
talɛ:thê:
a-verdade-ACC.PL.N
Ele me permitiu ir frente a vós [e/para] dizer a verdade.
(Xen. Hell. 6.1.13)[1]

Os dois fatores com maior correlação são a classe do verbo principal [2] e a autoria. [3] Uma vez que é impossível comparar resultados de testes estatísticos e valores p, não se pode deduzir a partir deles qual fator é mais determinante para a ocorrência da atração, sendo assim necessário estabelecer, a partir da teoria, um modelo causal gerativo – não confundir com modelo *sintático* gerativo.

Antes de apresentar o modelo causal que defendo, são necessárias algumas explicações sobre as "classes verbais" estatisticamente correlacionadas com a atração. Os verbos que permitem a atração de caso regem um objeto em caso oblíquo (dativo ou genitivo) e uma oração infinitiva, organizando-se em três classes:

1. verbos pessoais com a semântica de '*ordenar*/*pedir* /*aconselhar*/etc. [a alguém] [fazer algo]' (e.g. παραγγέλλω, δέομαι e συμβουλεύω);
2. a forma impessoal do verbo δοκέω 'parecer bom [a alguém] [fazer algo]';
3. verbos impessoais que denotam a possibilidade ou necessidade da ação infinitiva 'ser possível/suficiente/necessário [a alguém] [fazer algo]' (e.g. ἔξεστι, ἐξαρκεῖ e προσήκει).

A classe "independente" do verbo δοκέω, embora sintaticamente mais próxima da terceira classe, é semanticamente mais próxima da primeira, dado que via de regra a interpretação mais comum de δοκεῖ μοι πράττειν é antes 'decido/prefiro fazer' do que 'parece-me (bom/melhor) fazer'.[4] Tratando as classes 1 e 2 como uma única, temos duas classes semânticas de verbos principais: os que denotam uma ação (ordenar, pedir, aconselhar ou decidir/acreditar) e as que denotam algum tipo de modalidade (ser possível, necessário ou suficiente).

Retornando às hipóteses de modelo causal gerativo dos dados, parece-me que as mais defensáveis são:

1. os autores se valem de maneiras distintas da atração por preferências individuais de estilo e a classe do verbo principal atua *independentemente* da autoria;

[1] Os exemplos utilizados foram retirados das edições disponibilizadas no TLG (PANTELIA, 1972-2023). Para garantir o cotejo dos exemplos por público não especializado, realizei a transliteração automaticamente utilizando o pacote `cltk` (JOHNSON et al., 2014–2021), o qual segue a reconstrução fonológica apresentada em Probert (2010).
[2] Pearsons's $\chi^2 = 36.370(1) p = 1.6 - 09$.
[3] Pearson's $\chi^2 = 16.506(2), p = 2.6e - 04$.
[4] Outros argumentos para o pertencimento da construção impessoal de δοκέω à primeira classe incluem: 1. possibilidade do emprego da construção pessoal δοκῶ μοι πράττειν e 2. possibilidade da substituição do sujeito da infinitiva δοκῶ μοι ἡμᾶς πράττειν 'parece-me melhor que *nós* façamos'.

2. os autores se valem de maneira distinta da atração por razão do dialeto de cada um (ático para Platão e Xenofonte, jônico para Heródoto) e a classe do verbo principal atua *independentemente* da autoria;
3. os autores não se distinguem *diretamente* no uso da atração e a correlação observada entre autoria e atração é produzida indiretamente pela preferência de cada um por classes verbais específicas, as quais atuam diretamente na seleção de construções com ou sem atração.

Das hipóteses levantadas acima, a mais interessante do ponto de vista linguístico é a terceira, posto que ela permite uma explicação mais estrutural das condições que produzem a atração de caso. No entanto, ela depende de razões e evidências de que os autores se valem de classes verbais distintas na produção dos seus textos.

A razão pela qual os autores utilizariam verbos distintos é sugerida pela posição intermediária ocupada por Xenofonte entre Heródoto e Platão na frequência de uso da atração de caso. Uma vez que dividimos os textos de Xenofonte entre diálogos filosóficos e prosa historiográfica, torna-se claro que o autor inverte a preferência entre construções atraídas e não atraídas, cf. Tabela 1.

	Sem atração	Com atração
Platão	13 (48.1%)	14 (51.8%)
Xenofonte (diálogo filosófico)	4 (40.0%)	6 (60.0%)
Diálogo filosófico (Total)	17 (45.9%)	20 (54.1%)
Xenofonte (prosa histórica)	29 (64.5%)	16 (35.5%)
Heródoto	35 (92.1%)	3 (7.8%)
Prosa histórica (Total)	64 (77.1%)	19 (22.8%)

Tabela 1: Distribuição de sentenças sem e com atração de caso por autoria e gênero. Xenofonte está próximo de Platão no uso de atração em seus diálogos socráticos, enquanto em sua prosa historiográfica, ele se aproxima de Heródoto, muito embora não as evite tanto quanto este.

Uma hipótese plausível para a posição de Xenofonte e sua preferência assimétrica pelas construções com e sem atração de caso a depender do gênero de seus escritos é que, ao trocar de gênero literário, a seleção lexical do autor muda. Textos de prosa historiográfica talvez usem verbos que denotam ações mais do que verbos que denotam modalidade, enquanto textos de diálogo filosófico prefeririam o oposto. O experimento desta apresentação busca produzir as evidências necessárias para a defesa dessa hipótese.

2 Modelagem

Embora testes estatísticos sejam úteis para apontar correlação, os coeficientes produzidos por eles não são interpretáveis. Assim, se lançássemos mão de um teste para evidenciar se os gêneros literários prosa historiográfica ou diálogo filosófico estão estatisticamente associados ao maior ou menor uso de verbos que denotam modalidade, não seria possível identificar quais verbos estão mais

associados a qual gênero ou qual a medida do efeito da troca de gênero na troca de classes verbais. Modelos *bayesianos*, como o Naïve-Bayes, permitem uma modelagem mais específica e a extração de informações com maior granularidade, de modo que se prestam às questões deste trabalho melhor do que os testes de frequência.

Assim, utilizo o modelo Naïve-Bayes, que é utilizado na construção de *classificadores* de texto, podendo ser empregado para *análise de sentimentos*, classificação de autoria, gênero sexual, gênero literário e outras tarefas. O modelo criado aqui calcula para palavra do texto p_i a probabilidade P de que ela ocorra em um texto do gênero literário g_j: $P(p_i|g_i)$. A soma das probabilidades de cada palavra *atestada* define a probabilidade de um dado texto t_h ser do gênero g_i, sendo que neste modelo *texto* é uma unidade de parágrafo. Neste modelo, substitui-se a palavra *como ocorre* pela entrada lexical, utilizando apenas o vocabulário *verbal* dos textos, e antes de se incluir as palavras no modelo, foi utilizada uma técnica de pré-processamento chamada Td-idf (*Term Frequency – Inverse Document Frequency*) que evita que o modelo produza confiança excessiva no caso de palavras extremamente raras.

O corpus selecionado compreende os documentos supérstites de Heródoto, Xenofonte, Platão e Tucídides, divididos em prosa filosófica e prosa historiográfica da seguinte maneira:

- Historiografia:
    1. Histórias de Heródoto;
    2. Xenofonte: (a) Ciropédia; (b) Anábase; (c) Helênica.
    3. História da Guerra do Peloponeso de Tucídides.
- Prosa filosófica (diálogos socráticos):
    1. Platão;[5]
    2. Xenofonte: (a) Agesilau; (b) Hierão; (c) Simpósio; (d) Apologia (e) Memoráveis.[6]

Utilizou-se do banco de dados anotados Diorisis (VATRI; MCGILLIVRAY, 2020) para a obtenção dos textos convertidos para entradas lexicais e para a filtragem dos verbos.

2.1 Modelo utilizado

O modelo utilizado para extrair os pesos do léxico verbal para classificar um determinado parágrafo como do gênero filosófico ou historiográfico foi treinado com 80% do banco de dados e avaliado com os 20% restantes. As métricas obtidas estão em Tabela 2 e a matriz de confusão em Figura 1. Em resumo, este modelo prevê corretamente o gênero textual de um parágrafo aproximadamente 80% das vezes utilizando-se da entrada lexical apenas do verbos utilizados nos parágrafos.

[5] Estão excluídos do corpus os textos de atribuição duvidosa ou apócrifos, a saber: (a) Amantes, (b) Cartas, (c) Alcibíades Primeiro, (d) Alcibíades Segundo, (e) Clítofon, (f) Epinomis, (g) Hiparco, (h) Menexeno, (i) Minos, e (j) Teages.

[6] Alguns textos de Xenofonte foram deixados de fora por pertencerem a gêneros fora das classes aqui observadas.

	precisão	cobertura	F_1	suporte
historiografia	0.81	0.76	0.79	4696
filosofia	0.79	0.84	0.82	5136
acurácia			0.80	9832
macro avg	0.80	0.80	0.80	9832
weighted avg	0.80	0.80	0.80	9832

Tabela 2: Resumo das métricas do modelo utilizado

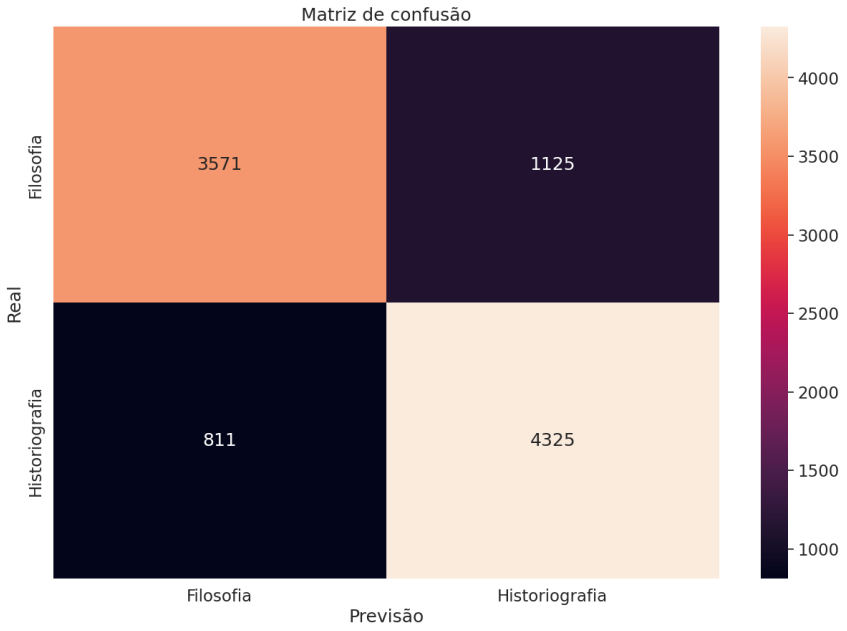


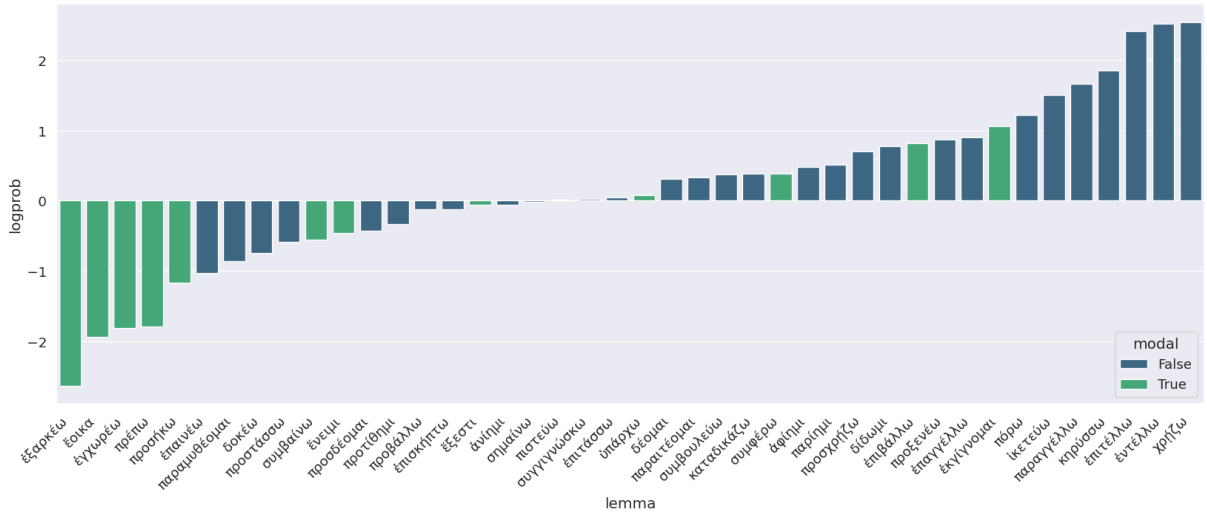
Figura 1: Matriz de confusão para o modelo utilizado, métricas em Tabela 2

3 Discussão

O próximo gráfico contém os coeficientes derivados pelo modelo para cada um dos verbos que pode reger um objeto oblíquo e oração infinitiva e que estão atestados no corpus de estudo.[7] Valores positivos de coeficiente indicam que, estando esse verbo em um dado parágrafo, a probabilidade de que este parágrafo corresponda a um texto historiográfico sobe, e os valores negativos favorecem a hipótese de que o texto vem de um diálogo filosófico. Por fim, as barras em verde marcam os verbos que denotam modalidade, as azuis verbos que denotam ações. 5 Os verbos com semântica modalizadora ocupam em sua maioria a região esquerda do gráfico, onde a presença de um item favorece a atribuição do parágrafo ao gênero filosófico, como previa a hipótese inicial. Apenas três verbos dessa classe possuem peso contrário, a saber sumferw, hyparxw e porw, sendo particularmente excepcional

[7] Os coeficientes aqui são produzidos a partir da diferença entre as probabilidades de ocorrer em um texto filosófico e de ocorrer em um texto historiográfico. A classe de modelo Naïve-Bayes utilizada aqui estima a probabilidade de pertencer tanto ao gênero g quanto ao gênero $\hat{g}$ (não-g), sendo assim necessário produzir um valor de diferença. A prática de contraste de coeficientes é comum na análise de modelos estatísticos bayesianos, cf. McElreath (2020).

o caso do último.[8] Por outro lado, os verbos com semântica de 'ordenar/aconselhar/etc.' estão concentrados na direita do gráfico, indicando a preferência por eles em textos historiográficos, embora a dispersão seja muito maior.



Uma palavra sobre as unidades pode ser importante: a escala logarítmica do eixo y é utilizada em modelos porque computadores sofrem calculando probabilidades de 0 a 1 ou 0 a 100% e ela é um tanto estranha de se interpretar. No exemplo acima, o verbo entellw tem um valor de aproximadamente $logit(2.5)$, o que equivale a algo na casa de 0.97. Isso é um valor **muito** alto. Qualquer verbo cujo coeficiente ultrapasse a casa de $\pm logit(1.0)$ indica que a probabilidade condicional de que aquele verbo ocorra mais em um gênero textual do que outro é altíssima, ou ao menos que o modelo tem muita certeza disso e a faixa entre $logit(0.5)$ e $logit(0.75)$ não pode ser desprezada. Assim, esse experimento oferece evidência de que:

1. a maioria dos verbos de interesse desta pesquisa com semântica modal estão associados ao gênero do diálogo filosófico grego;
2. alguns destes tem associação extremamente forte, servindo de ponto de contraste entre o gênero filosófico e historiográfico;
3. a maioria dos verbos de interesse que denotam ordens, conselhos etc. estão associados ao gênero de prosa filosófica; e
4. esta última classe de verbos é mais dispersa na escala do que a classe dos modalizadores.

Referências

GERALDES, Caio Borges Aguida. *Case Attraction on Infinitive Clauses of Ancient Greek: A case study on Herodotus, Plato and Xenophon*. 2020. Diss. (Mestrado) – Universidade de São Paulo, São Paulo, 2020. DOI: https://doi.org/10.11606/D.8.2020.tde-12042021-174449.

[8] A justificativa provavelmente está na baixíssima frequência de ocorrência deste verbo no corpus grego como um todo, com exceção apenas de Homero.

GERALDES, Caio Borges Aguida. *Dataset for Case Attraction on Infinitive Clauses of Ancient Greek in Herodotus, Plato and Xenophon*. Zenodo, 2021. DOI: 10.5281/zenodo.4906110.

JOHNSON, Kyle P. et al. *CLTK: The Classical Language Toolkit*. 2014–2021. Disponível em: <https://github.com/cltk/cltk>.

MCELREATH, Richard. *Statical Rethinking: A Bayesian Course with Examples in R and STAN*. 2. ed.: Chapman e Hall/CRC, 2020.

PANTELIA, Maria C. (Ed.). *Thesaurus Linguae Graecae® Digital Library.* University of California. 1972-2023. Disponível em: <http://www.tlg.uci.edu>.

PROBERT, Philomen. Phonology. In: BAKKER, Egbert J. *A companion to Ancient Greek*. London: Wiley-Blackwell, 2010. P. 85–103.

VATRI, Alessandro; MCGILLIVRAY, Barbara. Lemmatization for Ancient Greek: An experimental assessment of the state of the art. *Journal of Greek Linguistics*, Brill, Leiden, The Netherlands, v. 20, n. 2, p. 179–196, 2020. DOI: https://doi.org/10.1163/15699846-02002001. Disponível em: <https://brill.com/view/journals/jgl/20/2/article-p179%5C%5F4.xml>.