UiO : Det humanistiske fakultet

Diana Santos

Temos de contar com as contas?

Argumentos a favor e contra o uso de métodos quantitativos na linguística

ELC 2012
13 de Setembro de 2012

Porquê este assunto?

Questão atual: A ciência ganha terreno… demais? (algumas hipocrisias e incompetências)

Questão prática: Quando há muito material (com por exemplo grandes corpos), como proceder?

Questão filosófica: a contagem é algo separado da qualificação?

Questão pessoal: há muito que queria aprender estatística e não tinha tempo…

Plano da apresentação

Uma discussão esclarecedora dos conceitos quantitativos e qualitativos baseada em Karlgren

Um pouco de história

Alguns exemplos de abordagens populares

Gramática do português baseada em corpos: demonstração empírica do interesse, ou quimera?

Citação de P. Guiraud

La linguistique est la science statistique type; les statisticiens le savent bien; la plupart des linguistes l'ignorent encore.

Pierre Guiraud. *Problèmes et Méthodes de la statistique linguistique*, Paris, P.U.F., 1960.

Quantitativo ou qualitativo?

- Estamos a falar de estatística, ou de linguística matemática?
- De acordo com Ferenc Kiefer (1964), na linguística matemática há os seguintes tipos:
  - Baseados na teoria dos conjuntos, na lógica matemática, na álgebra, na estatística e na teoria da informação
  - Exemplos são Lambek, Bar-Hillel, Chomsky, Adjukiewicz, Kulagina

Curiosidade

Kiefer menciona, na conclusão, que estatística não é carne nem peixe

- *The statistical "models" of language that are known today [...] cannot be represented as a formal system. Thus, the statistical "models" cannot be called models at all [...]* (p. 25-26)
- (Um modelo linguístico matemático é definido como um sistema mais ou menos formalizado que ilumina um ou mais aspetos da língua. E formalização considera apenas as funções operativas e não eidéticas (significado))

Um modelo na estatística

De facto, a palavra *modelo* significa coisas diferentes em áreas diferentes…

Modelo no tricô, modelo no automobilismo, passagem de modelos, modelo na engenharia, modelismo, modelo matemático, … modelo na estatística:

É uma descrição matemática dos dados, que é proposta sem assumir mecanismos que causam os dados (Johnson, 2008:106)

Linguística quantitativa, para Hans Karlgren (1975)

- “Linguística quantitativa” define-se pelo método, ou também pela visão sobre a língua?
- Modelos quantitativos usados em linguística
  - Argumentos quantitativos em prol de questões qualitativas
  - Descrições quantitativas da língua
  - Explicações quantitativas de fenómenos linguísticos

Questões qualitativas

Teste quantitativo de uma hipótese

- Origem comum de duas línguas: se mais do que uma dada percentagem das palavras forem semelhantes

Geração de hipóteses

- Mudança de assunto, ou de autor, é marcada/detetável por contagem de palavras de um dado tema ou de um dado registo

Descrição linguística através de métodos quantitativos

- Descrição de géneros
- Descrição de estilos de textos
- Descrição de variedades de uma língua
- Descrição de estilos literários
- Descrição de estratos linguísticos

Explicações quantitativas de propriedades linguísticas

- Palavras frequentes são mais curtas
- Mudança na forma das palavras escandinavas na história: aumento de diferenciação fonológica e redução no número de sílabas estão correlacionadas, a primeira compensa a segunda (= reduzir duas dimensões a uma)

Métodos quantitativos…

(Karlgren, 1975:30)

- São, acima de tudo, métodos de **generalização** (mas não são os únicos)
- Os problemas ou os objetos não são **em si** quantitativos.
- O que é quantitativo é o **método** para combinar e "**reduzir**" observações a uma forma mais geral

Não se pode prever o futuro!

- Dado um tipo de questões linguísticas, não se pode prever que tipo de modelos matemáticos podem contribuir
- Dado um tipo de modelos matemáticos (ou um matemático com certas preferências), não se pode saber a priori que tipo de questões linguísticas podem esclarecer
- Conclusão: o mais que se pode fazer em planeamento de pesquisa é misturar e deitar fogo :-)

Encontro de linguística com corpos

O corpo

- É uma ferramenta e/ou uma fonte de dados
- Esses dados são geralmente **quantitativos**, de duas formas
  - frequência
  - distribuição

Porque é que estamos aqui?

- Qual é o problema/questão que temos?
- Ou temos um trabalho e queremos usar os corpos nesse trabalho?
- Traduzir, procurar informação, escrever, ensinar uma língua, rever, verificar se foi copiado/plagiado, aprender, fazer palavras cruzadas, relembrar...
- **O que nos faz mover?**

Agrupamento – ou anotação / classificação ?

Método científico

- Análise do problema
- Operacionalização da solução
- Teste
- Resultado

A previsão do clima e da temperatura

- É muito mais fácil prever o estado final do que a maneira de lá chegar

Abstração e generalização

- É o deixar cair dos pormenores...
- Imitação, ensino explícito, ... até naturalizar/concetualizar, e automatizar
- Exemplos: aprender a andar, a conduzir, a nadar, usar botões ou écrans/telas, torneiras, bolas, ringues, cordas, patins
- E outras coisas consideradas mais culturais como cumprimentar

Era uma vez dois homens...

- Um fragmento da história dos métodos quantitativos e da estatística na língua

George Udny Yule (1871-1951)

- Engineer at University College London
- Worked with his former teacher Karl Pearson, working in statistics
- *Introduction to the Theory of Statistics* (1911), 14 editions, based on his lecture
- One of the few members of *the Royal Statistical Society*
- Moved to Cambridge 1912, retired 1931
- ***The Statistical Study of Literary Vocabulary*** (Cambridge University Press, 1944)

GEORGE UDNY YULE

The Statistical Study of Literary Vocabulary

- *This book arose from a desire to study a particular vocabulary in a case of disputed authorship. When I had advanced some way in that particular study, it became only too clear into how thorny a field of statistics I had strayed. (p. ix)*
- *The vocabulary and diction of Thomas à Kempis are discussed as evidence. These discussions left in my mind a sense of inadequacy. They deal with such details (...); but they give no faintest notion as to what his vocabulary is really like as a whole.*

George Kingsley Zipf (1902-1950)

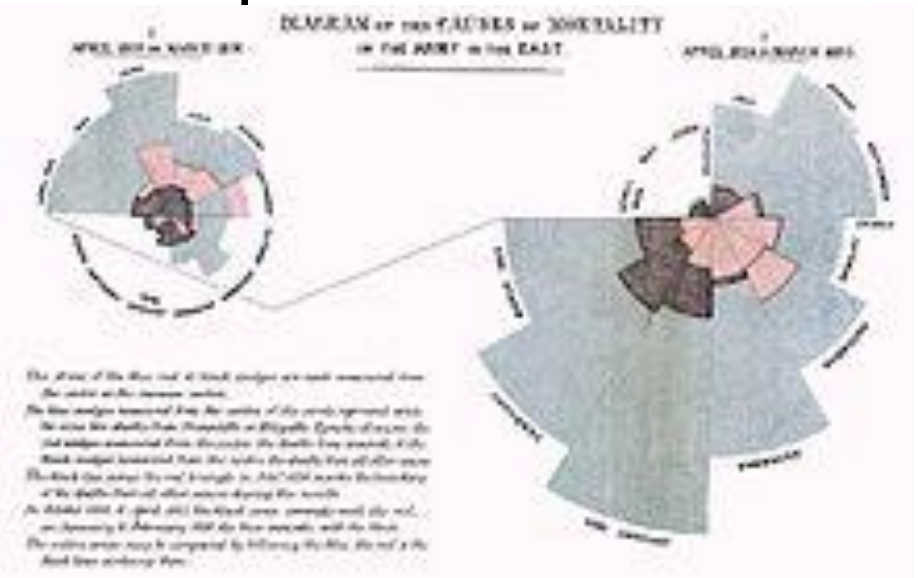
- Harvard philologist
- Chairman of the German Department and University Lecturer (meaning he could teach any subject he chose) at Harvard University
- *The Psychobiology of Language* (1935)
- *Human Behavior and the Principle of Least Effort* (1949)

Human behavior and the principle of least effort: an introduction to human ecology

- *Disclosure of some fundamental principles that seem to govern important aspects of our behavior, both as individuals and as members of social groups*
- *Discover the nature of the underlying principles that govern our conduct*
- *Neither the natural scientist nor the practical social engineer can afford to ignore the power of such preconceptions (...) Nevertheless, to the natural scientist man's preconceptions do not belong to some other world, but instead are further natural phenomena*
- *The expressed purpose of this book [is] to establish The Principle of Least Effort as the primary principle that governs our entire individual and collective behavior of all sorts*

E uma mulher…

- Considerada como a fundadora da logística
- Generais e homens de Estado vinham procurá-la para se informar de como tratar e dimensionar as suas tropas
- Iniciadora da
- visualização
- de dados



Florence Nightingale (1820-1910)

Primeira mulher numa sociedade de estatística (Royal Statistical Society, 1859)

*a true pioneer in the graphical representation of statistics, and is credited with developing a form of the pie chart now known as the **polar area diagram** or occasionally the **Nightingale rose diagram**,*

Primeira mulher a receber a ordem de mérito britânica (em 1907)

Enorme influência na política médica e social no mundo

Reações a Zipf e a Yule

- Herdan (1963:121)
- *Yule, when comparing the two values of r [.960 e .986] considers them being appreciately different. However, taking into account the sampling errors attached to each correlation coefficient such a view is not justified, [...]*

Reações a Zipf e a Yule

- Tweedie & Baayen,1998: How variable may a constant be?
- *When we estimate the slope of the regression line at 40 equally spaced intervals for varying text sizes, the estimated slope changes systematically.*

  Baayen (2008:226ff)
- Distribuições de frequência das palavras têm muitos elementos com probabilidades muito pequenas, são distribuições LNRE (“large number of rare events”) [modelo de Gauss-Poisson generalizado, e modelo de Zipf-Mandelbrot finito]

Reações a Zipf e a Yule

- George Miller na sua introdução/prefácio de 1965 ao livro linguístico de Zipf:
- *If we assume that word-boundary markers (space) are scattered randomly through a text (…) His facts were right enough, but not his explanations*
- Baroni (2008) discorda…
  - http://zipf.r-forge.r-project.org/

Aplicando as descobertas de Zipf e Yule

Quantas palavras conhecia Camões?

Os últimos Harry Potters são mais diversos lexicalmente que os primeiros?

Consegue distinguir-se níveis de falantes de uma língua estrangeira pela riqueza de vocabulário?

Quantas palavras sabe uma criança de 5 anos?

Mudanças no vocabulário podem prever Alzheimer?

Mas voltando à vaca fria…

Na linguística com corpos, quem são as pessoas mais conhecidas e mais influentes no uso da estatística?

Biber

Church

Labov

Halliday

Jellinek

Biber

- *Investigating macroscopic textual variation through multifeature/multidimensional analysis* (1985)
- Abordagens micro- e macroscópicas
- Abordagem macroscópica: tenta encontrar os parâmetros de variação subjacentes a um conjunto de tipos de texto, usando “análise de fatores”
- Fatores: **interpretação** dada a um conjunto de características que co-ocorrem

Biber (cont.)

Exemplos de fatores (dimensões textuais):

- Estilo distanciado ou imediato
- Avaliação estilística geral, emoção pessoal, ornamentação, abstração, seriedade e caracterização
- Interativo ou editado, abstrato ou situado, e distanciado ou imediato

Nota: a análise de fatores está dependente dos fundamentos teóricos que levam à escolha do material e das características

Biber: descrição detalhada

- Para cada dimensão textual esperada, incluiu pelo menos cinco características...
- Usou 42 num corpo de 2 milhões de palavras (545 textos)
- Fatores são expressos como uma combinação linear das quantidades das características, com pesos (“factor loadings”) associados
- Método: CFA extrair o máximo de variância partilhada, com rotação dos fatores (Promax)

Biber: descrição detalhada (cont.)

Como avaliar a correção de uma análise de fatores?

- Verificar se as características identificadas realmente são típicas do tipo de textos que quer descrever
- Análise confirmatória de fatores: adicionando novas características e ver se emparelham com os fatores esperados

Biber: classificação dos fatores ("factor scores")

- É calculada, para cada texto, através da soma das características únicas para cada fator
- Depois faz-se uma ANOVA entre o tipo de texto e estas classificações, para identificar se é possível prever o tipo de texto através destes valores

A estatística na literatura

- *Male and female language in Jane Austen's novels* (1980) – Kari-Anne Rand Schmidt
- Boa língua é sinal de bons princípios: As personagens boas falam bem, as más fazem erros de gramática... (além das diferenças sociais) no discurso direto
- Há diferenças entre homens e mulheres na maneira como falam? **Há**
- Igualdade entre os sexos significa igualdade de língua? **Não**

KARS (cont.)

- Uso dos adjetivos nos discurso direto dos seis livros de Jane Austen
- Atributivo vs. predicativo

	Atributivo	Predicativo
Homens	989	886
Mulheres	1234	1676

Usando X (chi quadrado)= 48,608 mostra que a diferença não é devida a simples sorte: p = 3.126e-12

KARS (cont)

- Restringindo a conversações “tête-à-tête”

	Atributivo	Predicativo
Homens	641	592
Mulheres	239	336

- X-squared = 16,6338 p-value = 4.534e-05
- Comparando com os outros casos (conversa só entre mulheres, e conversa mista), este é o caso mais flagrante de diferença na proporção dos 2 tipos de adjetivos

KARS (cont)

- Mais duas características comparadas entre homens e mulheres
- Uso de comparações (neutro, comparativo, superlativo)
- Conjuntos de adjetivos: 1, 2, 3 ou 4, ou 1 ou mais
- Conclusão: a língua das mulheres segue uma norma mais estrita (“stronger norm”) do que a dos homens

Church (2000)

A probabilidade de dois Noriegas… é mais próxima de p/2 do que p^2

Poisson não é apropriado, porque, depois de aparecer uma vez, a probabilidade de aparecer outra é muito maior

- para palavras que indicam o tópico

- conceito de adaptação (maior para boas palavras-chave), modelado por P (k>2|k>1) ou P(teste|história)

Katz: uma distribuição com 3 parâmetros para palavras e expressões com conteúdo

Para explicar a distribuição das palavras entre e dentro de documentos, são precisos 3 parâmetros, por exemplo α, β, e B (de burstiness)

A probabilidade de ver pelo menos uma ocorrência, α, depende do tamanho do documento $\alpha = f*L / B$ (L)

A probabilidade de repetir depende apenas de se existe ou não mais alguma coisa para dizer

Várias coisas de que gosto em Katz (1996)

1. Argumentos linguísticos (com base em com a língua funciona e com sugestões de causas)
2. Estudo empírico muito bem descrito e avaliado
3. Aplicação prática interdisciplinar: RI e fala
4. Visão crítica de muitas tentativas de adaptar a língua a modelos estatísticos, por muito que dêem uma boa aproximação: misturas de Poisson correspondem a um mecanismo estocástico em duas fases, incompatível com a realidade da produção da lingua

Widdows

- *Geometry and meaning*
- Usar distâncias entre textos ou palavras, calculadas através da sua representação vetorial de coocorrências

M. A. K. Halliday

Halliday (2005): “probability” as a theoretical construct is just the technicalising of “modality” from everyday grammar

- The grammar of a natural language is characterized by overall quantitative tendencies (two kinds of systems)
  - equiprobable: 0.5-0.5
  - skewed: 0.1-0.9 (0.5 redundancy) – unmarked categories
- In any given context, ... global probabilities may be significantly perturbed. ... the local probabilities, for a given situation type, may differ significantly from the global ones. *“resetting” of probabilities ... characterizes functional (register) variation in language*. This is how people recognize the “context of situation” in text. (pp. 236-8)

Fonte de estatística em linguística: os psicólogos!

- Style in language
- Citação de Miller
- Valor e surpresa

Aulas de português

- A língua é qualitativa, não conta.
- Ai não?
- *Está muito calor*
- *Junta um pouco de sal*
- *Não estava ninguém na sala*
- *Ele ia depressa demais*
- *Ele é meio parvo*
- *Ela já é suficientemente crescida*

Aulas de português

- A língua não fala de tendências
- Ai não?
- *As pessoas dantes não ligavam à forma de vestir*
- A língua não fala de probabilidades
- Ai não?

Ele deve ter chegado aí pelas 3 horas

Ele já deve ter feito o almoço

De acordo com declarações da prefeitura de São Paulo, ele terá sido assassinado...

Muito provavelmente todos os ramos da matemática começaram na língua

- Com questões sobre a língua
- Com exemplos da língua (e do raciocínio)
- A estatística não é necessariamente exceção

- Quando a inteligência artificial se virou para a língua como último reduto da inteligência, fechou-se o círculo

Gramática baseada em corpos

Uma gramática baseada em corpos, ou num corpo especialmente criado para o efeito

Forma abstrata de prosseguir:

- Definir as áreas
- Usar o corpo como oráculo
- Observar questões interessantes, por exemplo com o apoio de frequência
- Prosseguir por esses ramos

Uma questão puramente extensional

Só os verbos que estão no corpo podem ser usados… assim, verbos raros não vão contar, e verbos mal analisados também não

Mas poderá haver verbos que não constem de lista nenhuma

Quantos verbos?

Formas: Formas distintas:

Lemas: Lemas distintos :

O exemplo da passiva

Usando todos os corpos do AC/DC (corpo todosjuntos ☺)

Qual a frequência da passiva?

1862233 casos em 29467910 verbos principais (6,3%), ou em 21116214 verbos principais finitos (8,8%).

Com base nos verbos transitivos… Quais os verbos transitivos nos corpos do AC/DC?

- Aqueles que têm um objeto direto: 8532385
- Ou que estão na passiva: 1862233

São 27414 verbos diferentes, dos quais 7769 (28,3%) se encontram (também) na passiva

Contas na passiva

Quantos verbos? O que significa esta pergunta?

Quantas formas? Quantas formas distintas? OU

Quantos lemas? Quantos lemas distintos?

Com que auxiliar? (em 1806432)

ser 1462926 (81,0%) *estar* 283111 (15,7%) *ficar* 60395 (3,34%)

Preferência dos auxiliares…

Preferências dos auxiliares

Absoluta: qual o verbo com mais casos do auxiliar

ser: fazer,

estar: prever,

ficar: marcar

Relativa: qual o verbo que atrai mais a passiva com

ser: submeter

estar: concluir

ficar: surpreender

Relações com a passiva

Pergunta: A passiva é mais frequente em orações relativas?

Frequência em geral (nos verbos transitivos): 17,9%

Frequência em orações relativas (nos verbos transitivos)

234834 em 4578152 (5,1%) (ou em 1082620, 21,7%)

Verbos mais e menos frequentes com a passiva

Mais: inaugurar, nomear, condenar, adiar, aprovar, eleger, substituir, aplicar, exibir, obrigar, entregar, atribuir, prender, proibir, transmitir, interpretar, distribuir, utilizar, transportar, analisar

Menos: insistir, restar, depender, bastar, custar, pertencer, funcionar, poder, ser, existir, estar, acontecer, ficar, ir, faltar, permanecer, optar, concordar, contribuir, andar

Verbos mais frequentes com a passiva com *estar*

concluir, sujeitar, relacionar, vedar, sintonizar, descansar, normalizar, lesionar, confinar, arredar, ancorar, vincular, saturar, arrepender, ligar, fartar, cotar, preparar, sossegar, envolver, contaminar, condicionar, errar, satisfazer, afixar, vocacionar

expor, comprometer, associar, interessar, concentrar, preocupar, dispor, afastar, parar, representar, encerrar, implicar, ameaçar, isolar, condenar, sentar, submeter, prever, esconder

Verbos mais frequentes com a passiva com *ficar*

chatear, decepcionar, confinar, saturar, impressionar, desiludir, submergir, surpreender, satisfazer, deprimir, alojar, danificar, descansar, hospedar, chocar, sossegar, congelar, sediar, normalizar, reter, paralisar, zangar, imobilizar, magoar, concluir, ferir…

isolar, instalar, comprometer, adiar, parar, esclarecer, definir, preocupar, afastar, resolver, prejudicar, reduzir, sentar, decidir, provar, expor, afectar, suspender, esquecer, destruir, situar

Verbos mais frequentes com a passiva com *estar* e *ficar*

concluir, sujeitar, chatear, confinar, saturar, decepcionar, descansar, normalizar, relacionar, vedar, satisfazer, sossegar, arredar, impressionar, fartar, deprimir, submergir, sintonizar, desiludir, lesionar, alojar, zangar, hospitalizar, ancorar, vincular, paralisar,

obcecar, congelar, sediar, hospedar, circunscrever, danificar, imobilizar, surpreender, ligar, comprometer, isolar, preparar, envolver, expor, preocupar, parar, ferir, afastar, concentrar, associar, instalar, interessar, dispor, resolver, sentar, encerrar, definir, representar, adiar

Verbos mais frequentes com a passiva com *estar* e *ficar*

concluir, sujeitar, chatear, confinar, saturar, decepcionar, descansar, normalizar, relacionar, vedar, satisfazer, sossegar, arredar, impressionar, fartar, deprimir, submergir, sintonizar, desiludir, lesionar, alojar, zangar, hospitalizar, ancorar, vincular, paralisar,

obcecar, congelar, sediar, hospedar, circunscrever, danificar, imobilizar, surpreender, ligar, comprometer, isolar, preparar, envolver, expor, preocupar, parar, ferir, afastar, concentrar, associar, instalar, interessar, dispor, resolver, sentar, encerrar, definir, representar, adiar

Vale a pena fazer estas explorações?

De um ponto de vista pedagógico, é certamente aproveitável

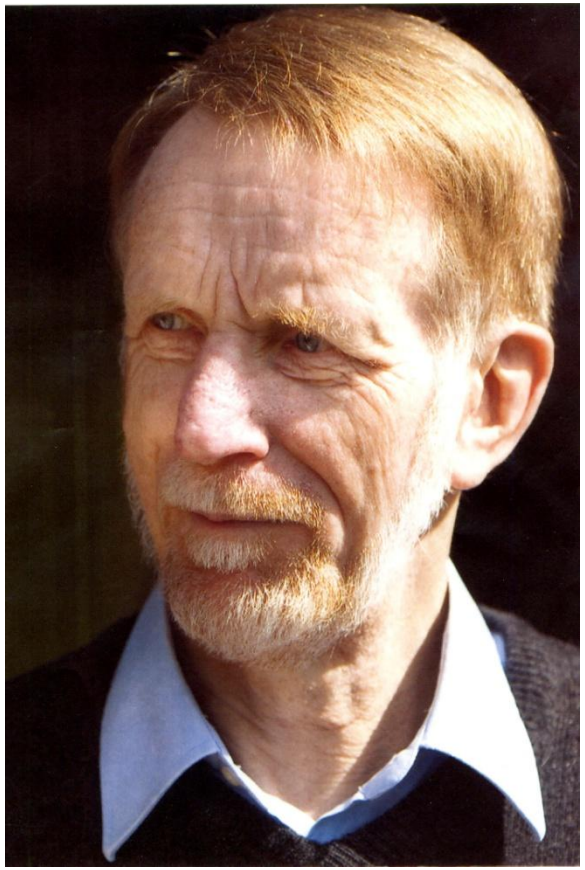
De um ponto de vista prático (por exemplo tradução automática) também

Mas: Ficamos a saber mais sobre a língua?

Depende da interpretação

Muito obrigada pela vossa atenção!

Gostava de dedicar esta palestra ao Stig



Stig Johansson