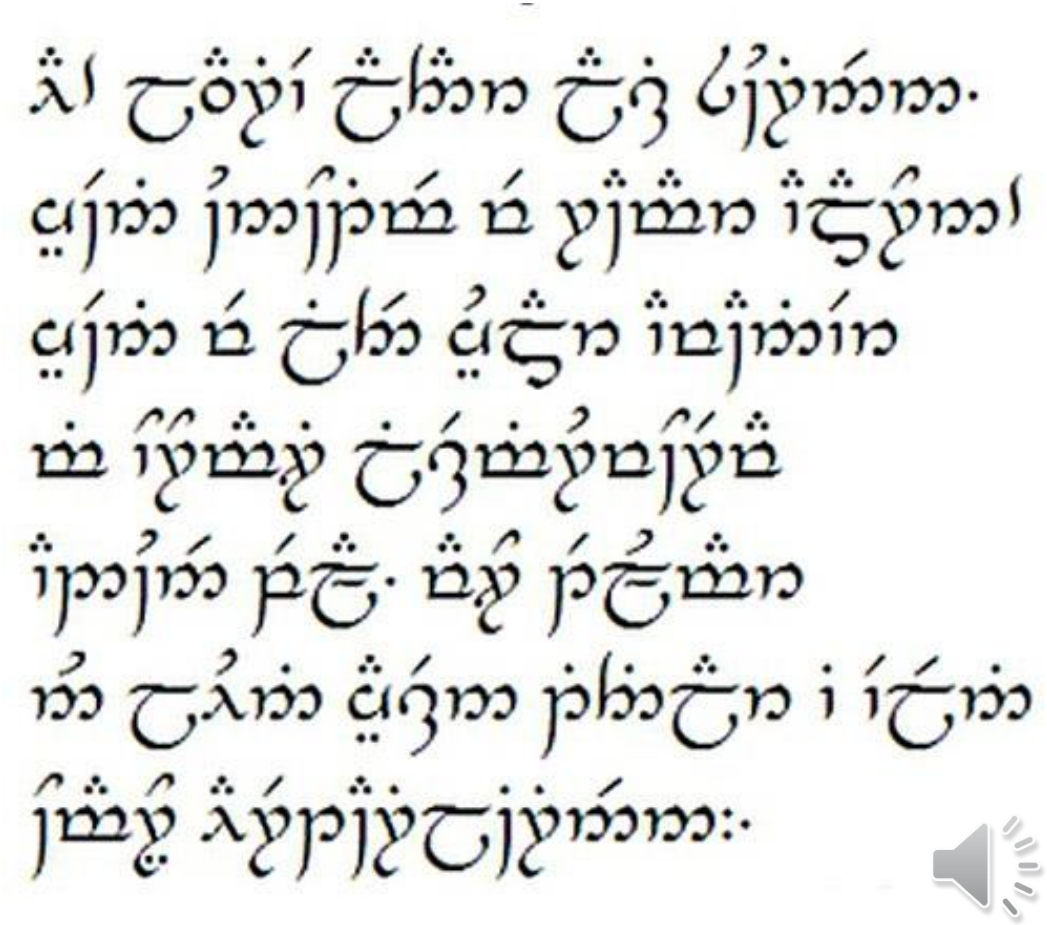
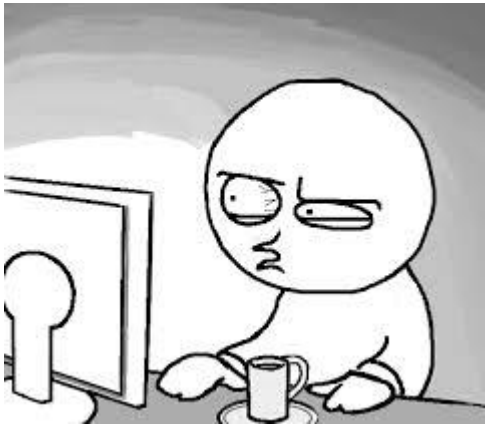
# SCC0633/5908 Processamento de Linguagem Natural





# SCC0633/5908 Processamento de Linguagem Natural

Primeira estrofe do poema “Namárië” (Adeus), também conhecido como “O Lamento de Galadriel”, maior texto escrito em quenya/élfico.

*Ah! Como ouro caem as folhas ao vento,*
*Longos anos inumeráveis como as asas das árvores!*
*Os longos anos se passaram como goles rápidos do doce hidromel*
*Em salões altos além do oeste,*
*Sob as abóbadas azuis de Varda*
*Onde as estrelas tremem na canção*
*De sua voz de Santa e Rainha.*
*Quem agora há de encher-me a taça outra vez?*
*Pois agora a Inflamadora, Varda, a Rainha das Estrelas,*
*do Monte Semprebranco, ergueu suas mãos como nuvens*
*E todos os caminhos mergulharam fundo nas trevas;*
*E de uma terra cinzenta a escuridão se deita*
*sobre as ondas espumantes entre nós*
*E a névoa cobre as jóias de Calacirya para sempre.*
*Agora perdida, perdida para aqueles do Leste está Valimar!*
*Adeus! Talvez hajas de encontrar Valimar.*
*Talvez tu mesmo hajas de encontrá-la. Adeus!*

# Para aquecer

# Retomando uma pergunta que já respondemos: o que é necessário para aprender uma língua?

- Conhecer as palavras e como elas são formadas
- Saber o significado das palavras
- Como compor frases
- Como referenciar entidades do mundo
- Como conectar frases
- Protocolos de comunicação na língua/cultura
- Etc.

Como ensinar isso às máquinas?

# Córpus

*SCC5908 Introdução ao Processamento de Língua Natural*
*SCC0633 Processamento de Linguagem Natural*

**Prof. Thiago A. S. Pardo**

# Córpus

- Ou, tradicionalmente, “corpus” e “corpora”
- Alinhado à tradição empirista
- Essencial na área hoje
  - Linguística de córpus e o estudo dos fenômenos da língua e sua ocorrência
    - Proposta de teorias e validação
  - Estatísticas da língua e modelagem linguístico-computacional
  - Aprendizado de máquina
    - Córpus & datasets
  - Avaliação de sistemas

# Definições e história

# Um pouco da história (Sardinha, 2000)

- Brown Corpus
  - *Brown University Standard Corpus of Present-Day American English*
  - Lançado em 1964, com uma quantidade invejável de dados na época: 1 milhão de palavras
  - Considerado o primeiro córpus eletrônico
    - Desafios tecnológicos: textos em cartões perfurados!
    - Desafios científicos: tal empreitada era tratada com incredulidade e até hostilidade
      - Época da visão predominante da obra de Chomsky: para que córpus, se tudo está na mente e pode ser estudado via introspecção?

# Um pouco da história (Sardinha, 2000)

- Mas havia córpus antes do computador
  - O termo original é “corpo” ou “conjunto de documentos”
    - Na Grécia antiga, Alexandre, o Grande, definiu o Corpus Helenístico
    - Na antiguidade, produziram-se córpus de citações da Bíblia
- Antes do computador
  - Córpus coletados, mantidos e analisados manualmente!
  - Ênfase no ensino de línguas
    - Por exemplo, crianças devem aprender primeiro as palavras mais frequentes da língua

# Um pouco da história (Sardinha, 2000)

- Córpus SEU (*Survey of English Usage*), com criação a partir de 1953
  - Inspiração para o Brown Corpus
  - Planejado para ter 1 milhão de palavras
  - Organizado em fichas de papel, com cada uma contendo uma palavra do corpus inserida em 17 linhas de texto e classificada gramaticalmente
  - Foi a base para uma das gramáticas mais famosas do inglês
  - Transformado em córpus eletrônico apenas em 1989

- As dificuldades desse tipo de trabalho
  - Pouca confiança na capacidade humana para analisar tanto material (lembrem-se, não havia computador pessoal)
  - Necessidade de muita mão de obra
  - Erros e inconsistências
    - Mas, se menos material para viabilizar o processo, perdia-se credibilidade

# Um pouco da história

- Com a advento do computador pessoal, tudo mudou
- Linguística de Córpus se fortalece
- Com o passar do tempo, as práticas da Linguística de Córpus e de PLN começam a se integrar
  - Mas ainda são vistas como disciplinas separadas
- Há o movimento de “web como córpus”
  - Google e o envolvimento de grandes empresas
  - Mas com ressalvas: *Googleology is Bad Science*
- O interesse multimodal
  - Não apenas texto, mas fala e vídeo também

# Um pouco da história

- O português também avança
  - Córpus NILC e o fortalecimento do PLN no Brasil
  - Corpus Brasileiro, com 1 bilhão de palavras
  - Tycho-Brahe e o português histórico (autores nascidos entre 1380 e 1845)
  - C-Oral-Brasil, com fala espontânea
  - Floresta Sintá(c)tica, com anotação morfossintática e sintática
  - CSTNews, com diversas camadas de anotação
  - brWaC, com material da web
  - E muitos outros!
    - AC/DC, projeto pioneiro da Linguateca, e outros portais de córpus
- A importância da anotação de córpus
  - *Towards a 'science' of corpus annotation: A new methodological challenge for Corpus Linguistics* (Hovy e Lavid, 2010)
- *Big data*, a ciência de dados e o aprendizado de máquina
  - *Linguistic Data Consortium* (LDC)
  - Kaggle
  - Etc.

# Definição

- Não é simples definir o que se entende por córpus, mas a definição abaixo é bastante aceita

  - *"Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise."* (Sanchez, 1995)

  - De maneira geral, pontos a destacar
    - Autenticidade do material
    - Propósito
    - Critérios para composição
    - Formatação e legibilidade por máquinas
    - Representatividade
    - Extensão

# TIPOLOGIA (SARDINHA, 2000)

- Modo
  - Falado: composto por falas transcritas
  - Escrito: composto por textos escritos, impressos ou não

- Tempo
  - Sincrônico: compreende um período de tempo
  - Diacrônico: compreende vários períodos de tempo
  - Contemporâneo: representa o período de tempo corrente
  - Histórico: representa um período de tempo passado

# TIPOLOGIA (SARDINHA, 2000)

- Seleção
  - De amostragem (*sample corpus*): composto por porções de textos ou de variedades textuais, planejado para ser uma amostra finita da linguagem como um todo
  - Monitor: a composição é reciclada para refletir o estado atual de uma língua, opondo-se a córpus de amostragem
  - Dinâmico ou orgânico: o crescimento e diminuição são permitidos, qualifica o córpus monitor
  - Estático: oposto de dinâmico, caracteriza o córpus de amostragem
  - Equilibrado (*balanced*): os componentes (gêneros, textos, etc.) são distribuídos em quantidades semelhantes (por exemplo, mesmo número de textos por gênero)

# Tipologia (Sardinha, 2000)

- Conteúdo
  - Especializado: os textos são de tipos específicos (em geral gêneros ou registros definidos)
  - Regional ou dialetal: os textos são provenientes de uma ou mais variedades sociolinguísticas específicas
  - Multilíngue: inclui idiomas diferentes

- Autoria
  - De aprendiz: os autores dos textos não são falantes nativos
  - De língua nativa: os autores são falantes nativos

# Tipologia (Sardinha, 2000)

- Disposição interna
  - Paralelo: os textos são comparáveis (por exemplo, original e tradução)
  - Alinhado: as traduções aparecem abaixo de cada linha do original
    - Mas, como já comentado, esses conceitos já foram estendidos para outras situações

- Finalidade
  - De estudo: o córpus que se pretende descrever
  - De referência: usado para fins de contraste com o córpus de estudo
  - De treinamento ou teste: construído para permitir o desenvolvimento de aplicações e ferramentas de análise
    - Muito comum em PLN

# TIPOLOGIA (SARDINHA, 2000)

- Também há outros critérios possíveis

  - Pluralidade de autoria: os textos foram produzidos por um autor apenas ou mais?

  - Integralidade: os elementos do córpus são textos integrais ou fragmentos?

  - Plurilinguísmo: o córpus possui só textos originais ou também as traduções destes textos para uma ou mais línguas?
    - Intercalação: as traduções dos textos são incorporadas a cada linha do texto original ou vêm em textos separados?

# Representatividade

- Questão difícil
  - Representativo do quê?
  - Representativo para quem?

- Critério usual: extensão
  - Pois a língua é um sistema probabilístico (que “dita” as ocorrências léxicas, de sentidos, etc.)
  - Relação com tamanho da comunidade falante, mas difícil de mensurar
  - Dimensões envolvidas
    - Número de palavras
    - Número de textos
    - Número de gêneros/tipos textuais

- Mas há aprendizados empíricos (às vezes, dependentes da tarefa em vista)
  - Em modelos de língua, fala-se em “bilhões”
    - Por exemplo, o BERTimbau foi treinado com o córpus brWaC, que tem 2,7 bilhões de tokens

# Modelagem da língua e seus fenômenos

# Córpus

- Diversas aplicações
  - Modelos matemáticos e estatísticos
    - Espaço vetorial, Bayes, Markov, modelos distribucionais, etc.
      - Raízes em estudos de frequência e “leis” de **distribuição de palavras/n-gramas**

# Frequência em PLN

- Exemplo: livros de Tom Sawyer (de Mark Twain)

| Word | Freq. | Use |
|---|---|---|
| the | 3332 | determiner (article) |
| and | 2972 | conjunction |
| a | 1775 | determiner |
| to | 1725 | preposition, verbal infinitive marker |
| of | 1440 | preposition |
| was | 1161 | auxiliary verb |
| it | 1027 | (personal/expletive) pronoun |
| in | 906 | preposition |
| that | 877 | complementizer, demonstrative |
| he | 877 | (personal) pronoun |
| I | 783 | (personal) pronoun |
| his | 772 | (possessive) pronoun |
| you | 686 | (personal) pronoun |
| Tom | 679 | proper noun |
| with | 642 | preposition |

Tokens = 71.370
Types = 8.018 (poucas para um texto tão grande)
→ para crianças

Taxa type/token = 0,11 (11%)

Em geral, quanto maior o córpus, menor a taxa

# Frequência em PLN

1902-1950

- Distribuição de palavras
  - Lei de **Zipf**
    - *George Kingsley Zipf*
    - Baseada em trabalho de Estoup (1916)
    - Proveniente do “Princípio do Mínimo Esforço”, publicado no livro *Human Behavior and the Principle of Least Effort* (1949)

# Frequência em PLN

- Distribuição de palavras
  - Lei de Zipf
    - Contam-se quantas vezes cada palavra ocorre em um córpus grande, montando-se um ranque em função da frequência delas
      - Há uma relação entre a frequência e a posição da palavra no ranque
        - **Frequência x posição no ranque = constante k**
        - Palavra na posição 50 deve ocorrer 3 vezes mais do que palavra na posição 150

# Frequência em PLN

- Exemplo: livros de Tom Sawyer
  - Há distorções, comuns na lei de Zipf

| Word | Freq. ($f$) | Rank ($r$) | $f \cdot r$ |
|---|---|---|---|
| the | 3332 | 1 | 3332 |
| and | 2972 | 2 | 5944 |
| a | 1775 | 3 | 5235 |
| he | 877 | 10 | 8770 |
| but | 410 | 20 | 8400 |
| be | 294 | 30 | 8820 |
| there | 222 | 40 | 8880 |
| one | 172 | 50 | 8600 |
| about | 158 | 60 | 9480 |
| more | 138 | 70 | 9660 |
| never | 124 | 80 | 9920 |
| Oh | 116 | 90 | 10440 |
| two | 104 | 100 | 10400 |

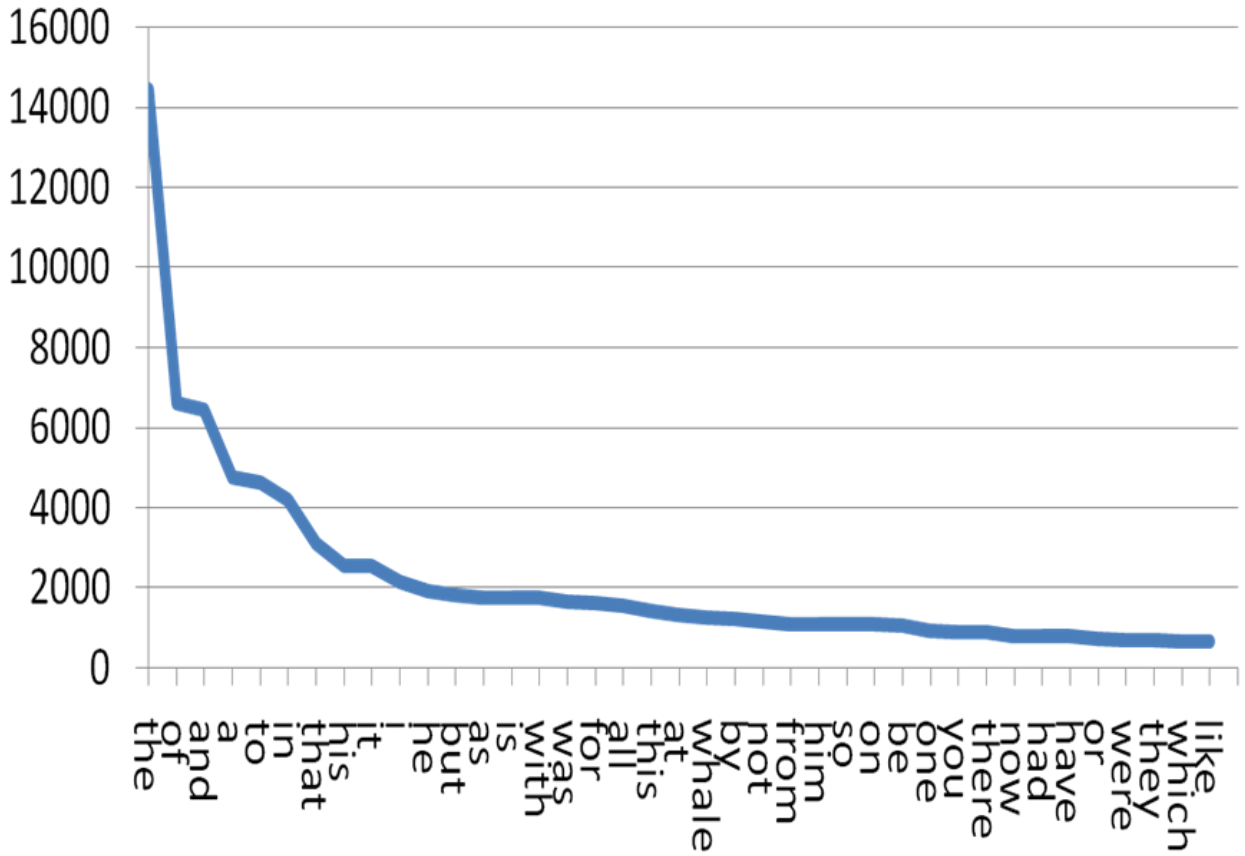
| Word | Freq. ($f$) | Rank ($r$) | $f \cdot r$ |
|---|---|---|---|
| turned | 51 | 200 | 10200 |
| you'll | 30 | 300 | 9000 |
| name | 21 | 400 | 8400 |
| comes | 16 | 500 | 8000 |
| group | 13 | 600 | 7800 |
| lead | 11 | 700 | 7700 |
| friends | 10 | 800 | 8000 |
| begin | 9 | 900 | 8100 |
| family | 8 | 1000 | 8000 |
| brushed | 4 | 2000 | 8000 |
| sins | 2 | 3000 | 6000 |
| Could | 2 | 4000 | 8000 |
| Applausive | 1 | 8000 | 8000 |

# Frequência em PLN

- Distribuição de palavras
  - Lei de Zipf
    - Poucas palavras muito frequentes
    - Número significativo de palavras de frequência média
    - Muitas palavras de frequência baixa
      - É possível plotar um gráfico

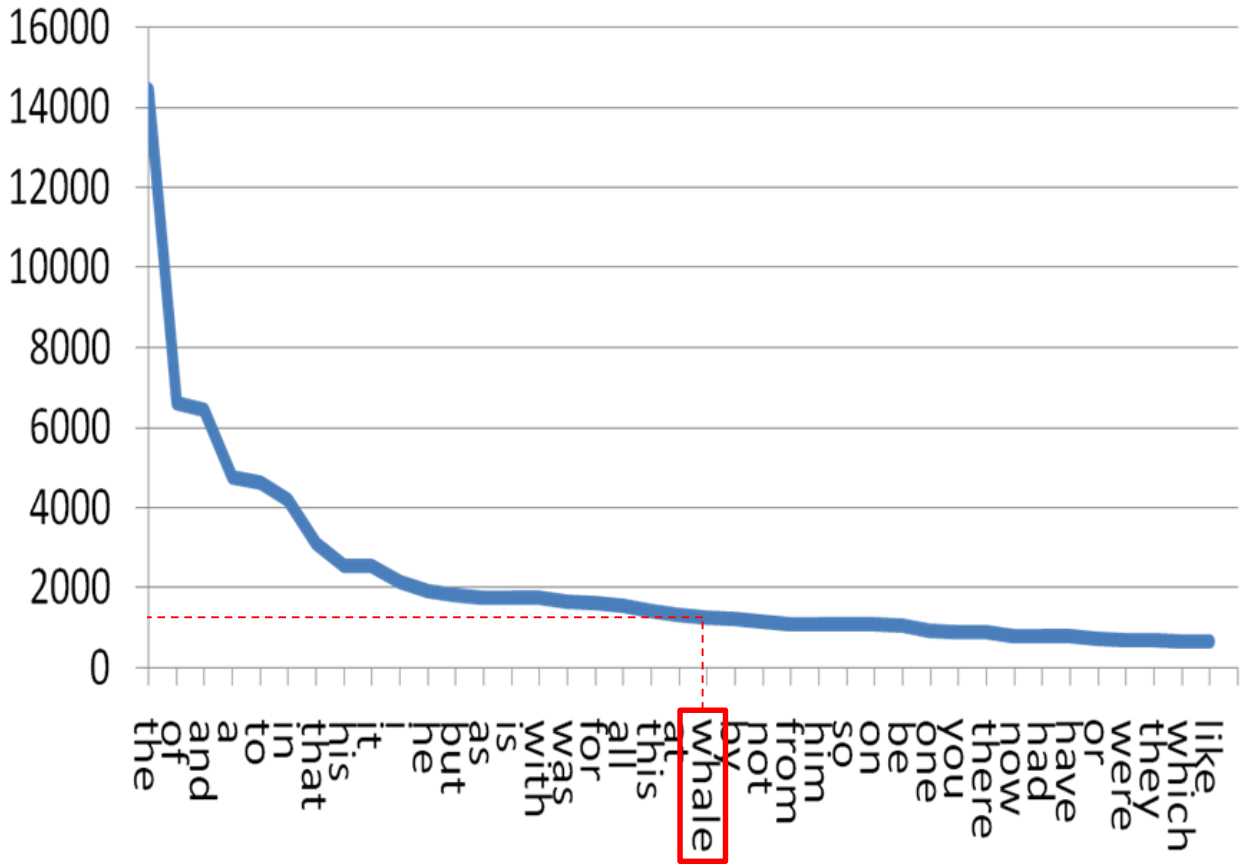
# Frequência em PLN

- Exemplo: parte inicial da curva de Zipf para Moby Dick



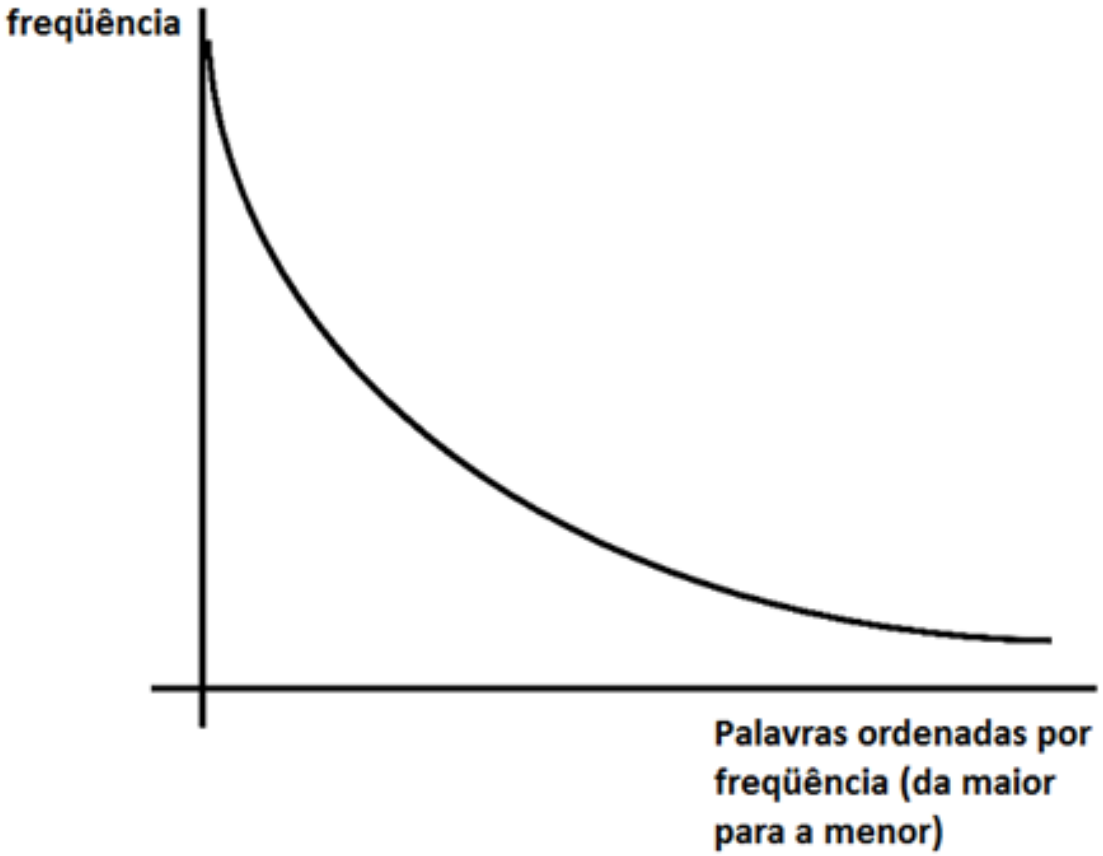
# Frequência em PLN

- Exemplo: parte inicial da curva de Zipf para Moby Dick



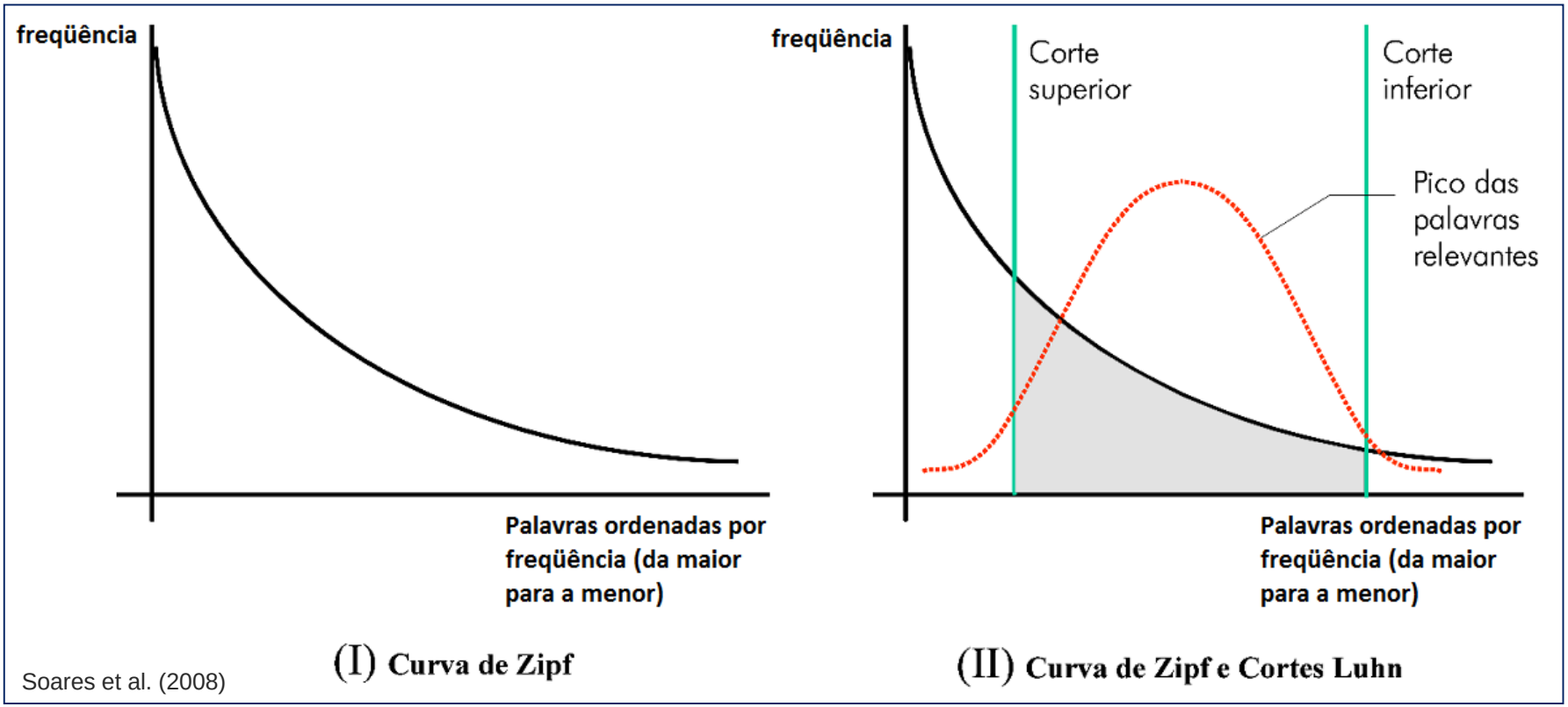
# Frequência em PLN

- Curva de Zipf



# Frequência em PLN

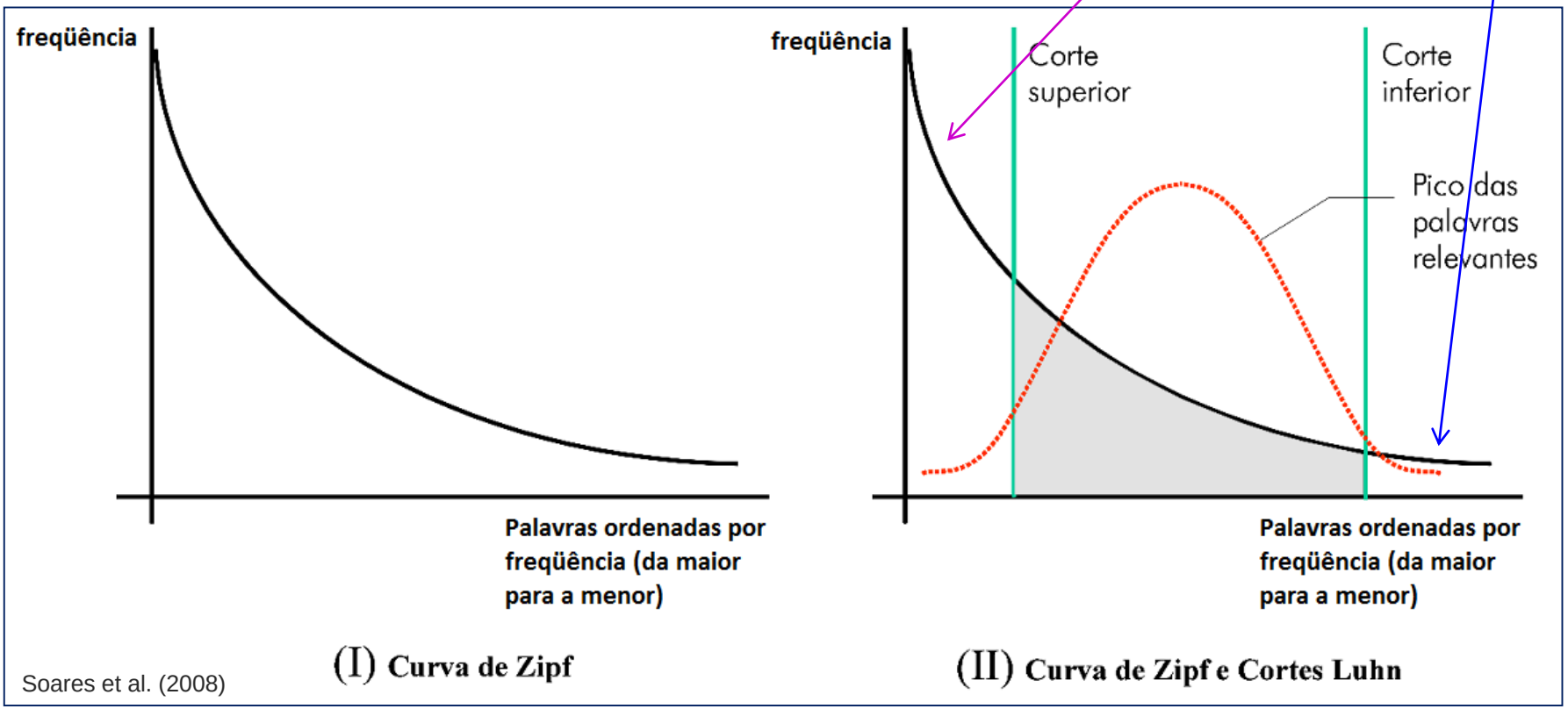
- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes




Soares et al. (2008)

# Frequência em PLN

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes



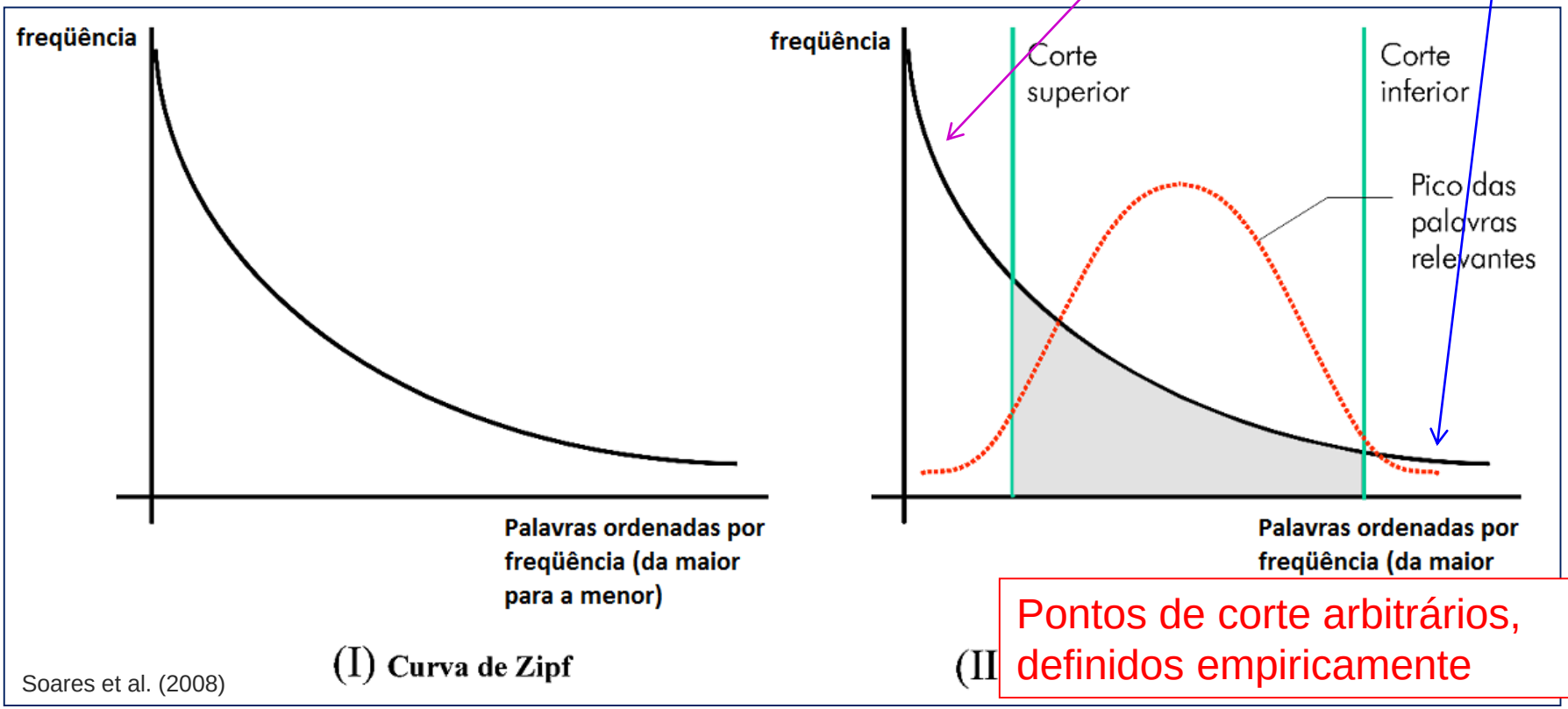

Soares et al. (2008)

# Frequência em PLN

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes

preposições, conjunções, etc.

termos raros




Soares et al. (2008)

Pontos de corte arbitrários, definidos empiricamente

# Frequência em PLN

- Distribuição de palavras
  - Outra lei de Zipf
    - O número de significados de uma palavra é correlacionado com sua frequência
      - Palavra com 10.000 ocorrências → 2.1 significados
      - Palavra com 5.000 ocorrências → 3 significados
      - Palavra com 2.000 ocorrências → 4.6 significados

# Frequência em PLN

- Distribuição de palavras
  - Ainda outras leis de Zipf
    - Uma palavra de conteúdo tende a ocorrer próxima a outra ocorrência sua
    - A frequência de uma palavra é inversamente proporcional ao seu tamanho
    - Quanto maior a frequência de uma palavra, mais “variação” há (em seus componentes morfológicos)

# Frequência em PLN

- Leis de Zipf
  - Exageradamente valorizadas
    - Não deveriam ser “leis”, mas “observações” aproximadas
  - Até alguns *eventos aleatórios* obedecem essas leis
    - Forma de gerar os dados, de construir a curva

# Frequência e limitações

- Frequência fornece informações interessantes, mas, isoladamente, é limitada em alguns aspectos, podendo esconder fatos e correlações, ser tendenciosa ou não suficientemente discriminativa

  - Alternativa?

# Por que estudar probabilidade?

- Probabilidade e língua

???

# Por que estudar probabilidade?

- Probabilidade e língua
  - Probabilidade dos fenômenos linguísticos
    - Descrição, caracterização
      - Caracterização de discursos políticos, detecção de mudanças históricas, contraste de discurso oral vs. textual, estudo de fenômenos sintáticos, probabilidades das colocações, etc.
    - Previsão
      - Que traduções são possíveis, qual a palavra correta mais provável dada uma palavra com ortografia errada, qual a chance de uma sentença ser importante no texto, etc.

# Por que estudar probabilidade?

- Probabilidade e língua
  - Probabilidade dos fenômenos linguísticos
    - Às vezes, esses “números mágicos” são intuitivos
      - Calculados naturalmente por nós
    - Às vezes, exigem raciocínio mais sofisticado

# Exemplo

- Exemplo (125 palavras)

  Foi controlado o incêndio que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
  De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
  As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



Qual a probabilidade da palavra “que” ocorrer?

*“chance”*

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



Qual a probabilidade da palavra “que” ocorrer? 3/125 = 0.024 = 2.4%

*“chance”*

# Exemplo

- Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E “incêndio”?

# Exemplo

- Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E “incêndio”? 4/125 = 0.032 = 3.2%

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E qualquer palavra do texto?

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E qualquer palavra do texto? 125/125 = 1 = 100%

# Exemplo

- Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O **fogo** atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E “incêndio” ou “fogo”?

# Exemplo

- Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O **fogo** atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E “incêndio” ou “fogo”? 4/125 + 1/125 = 5/125 = 0.04 = 4%

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E de um verbo seguir “que”?

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E de um verbo seguir “que”? 2/3 = 0.666 = 66.6%

# Exemplo

- Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos.
De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas.
As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.



E de um verbo seguir “que”? 2/3 = 0.666 = 66.6%
Reescrevendo: **P(palavra$_i$=verbo | palavra$_{i-1}$=“que”)**

# Exemplo

- Exemplo (125 palavras)

Foi **controlado** o incêndio que **atingiu** uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros **permaneciam** no local às 10h para o trabalho de rescaldo. Não **há** informação de feridos.
De acordo com a corporação, o incêndio **teve** início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros **foram encaminhados** para o local. O fogo **atingiu** vários barracos, mas as equipes ainda não **tinham** o número exato de propriedades **atingidas**.
As causas do incêndio ainda **serão investigadas**. Apesar do incêndio, a Infraero (estatal que **administra** os aeroportos no país) **informou** que a fumaça não **comprometeu** os pousos e decolagens no aeroporto de Congonhas.



E de “que” preceder um verbo?

# Exemplo

- Exemplo (125 palavras)

Foi **controlado** o incêndio que **atingiu** uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros **permaneciam** no local às 10h para o trabalho de rescaldo. Não **há** informação de feridos.
De acordo com a corporação, o incêndio **teve** início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros **foram encaminhados** para o local. O fogo **atingiu** vários barracos, mas as equipes ainda não **tinham** o número exato de propriedades **atingidas**.
As causas do incêndio ainda **serão investigadas**. Apesar do incêndio, a Infraero (estatal que **administra** os aeroportos no país) **informou** que a fumaça não **comprometeu** os pousos e decolagens no aeroporto de Congonhas.



E de “que” preceder um verbo? 2/15 = 0.133 = 13.3%

# Probabilidades

- Probabilidade: resultado entre 0 e 1, ou 0 e 100%
- P(evento impossível) = 0
- P(evento certo, ou qualquer coisa) = 1 (ou 100%)
- P(A) ou P(B) = P(A) + P(B)
  - P(qualquer coisa) = P(uma coisa) + P(segunda coisa) + ... + P(enésima coisa)
- Probabilidade condicional $P(A|B) = P(A \cap B) / P(B)$
- $P(A \cap B) = P(B) * P(A|B) = P(A) * P(B|A)$
  - $P(A \cap B) = P(A) * P(B)$, se eventos independentes

# Bayes

- Teorema de Bayes
  - P(A|B) = P(B|A) * P(A) / P(B)
    - Pode-se inverter: usar P(B|A) em vez de P(A|B)
      - Por que isso é interessante?

# Bayes

- Teorema de Bayes
  - P(A|B) = P(B|A) * P(A) / P(B)
    - Útil quando não se tem, é difícil ou ilógico calcular P(A|B) → pode-se usar o inverso
      - Por exemplo, o que é melhor?
        - P(doença|sintoma)
        - P(sintoma|doença)

# Bayes

- Teorema de Bayes
  - P(A|B) = P(B|A) * P(A) / P(B)
    - Útil quando não se tem, é difícil ou ilógico calcular P(A|B) → pode-se usar o inverso

P(doença|sintoma) = P(sintoma|doença) * P(doença) / P(sintoma)

- o sintoma é o que se observa, e a doença é o que se quer descobrir
  - P(doença|sintoma)
- ... mas quem causa o sintoma é a doença, e não o inverso
  - P(sintoma|doença)
- P(doença|sintoma) pode ser "tendencioso" e "temporal"

# Bayes

- Teorema de Bayes
  - Exemplo: “chegou com dor de cabeça no médico”
    - P(sarampo|dor de cabeça) = P(dor de cabeça|sarampo) * P(sarampo) / P(dor de cabeça)
    - P(malária|dor de cabeça) = P(dor de cabeça|malária) * P(malária) / P(dor de cabeça)
      - A maior probabilidade ganha e indica o diagnóstico final!
        - Atenção: P(dor de cabeça) é constante. Faz diferença no resultado?

# Bayes

- Teorema de Bayes
  - Exemplo: “chegou com dor de cabeça no médico”
    - P(sarampo|dor de cabeça) = P(dor de cabeça|sarampo) * P(sarampo) ~~/ P(dor de cabeça)~~
    - P(malária|dor de cabeça) = P(dor de cabeça|malária) * P(malária) ~~/ P(dor de cabeça)~~
      - A maior probabilidade ganha e indica o diagnóstico final!
        - Atenção: P(dor de cabeça) é constante. Faz diferença no resultado?
          - Ao comparar hipóteses, pode-se usar P(A|B) = P(B|A) * P(A)

# Exercício

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm catáforas

- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos

- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

# Exercício

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm catáforas

- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos

- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução: P(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim|catáfora) = 0.98 P(não|catáfora) = 0.02

P(sim|sem catáfora) = 0.03 P(não|sem catáfora) = 0.97

P(catáfora|sim) = P(sim|catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora|sim) = P(sim|sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

O que aconteceu?

# Exercício

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm catáforas

- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos

- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução: P(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim|catáfora) = 0.98 P(não|catáfora) = 0.02

P(sim|sem catáfora) = 0.03 P(não|sem catáfora) = 0.97

P(catáfora|sim) = P(sim|catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora|sim) = P(sim|sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

# Exercício

E se a probabilidade de ocorrência de catáforas fosse **uniforme** no córpus?

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm catáforas

- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos

- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução: P(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim|catáfora) = 0.98 P(não|catáfora) = 0.02

P(sim|sem catáfora) = 0.03 P(não|sem catáfora) = 0.97

P(catáfora|sim) = P(sim|catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora|sim) = P(sim|sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

# Exercício

E a probabilidade da sentença ser catafórica? Já conseguimos?

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm catáforas

- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos

- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução: P(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim|catáfora) = 0.98 P(não|catáfora) = 0.02

P(sim|sem catáfora) = 0.03 P(não|sem catáfora) = 0.97

P(catáfora|sim) = P(sim|catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora|sim) = P(sim|sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

# Exercício

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm catáforas

- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos

- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

**Normalizando...**

P(catáfora|sim) = 0.0078 ➔ 0.0078 / (0.0078+0.0298) = **0.21 = 21%**

P(sem catáfora|sim) = 0.0298 ➔ 0.0298 / (0.0078+0.0298) = **0.79 = 79%**

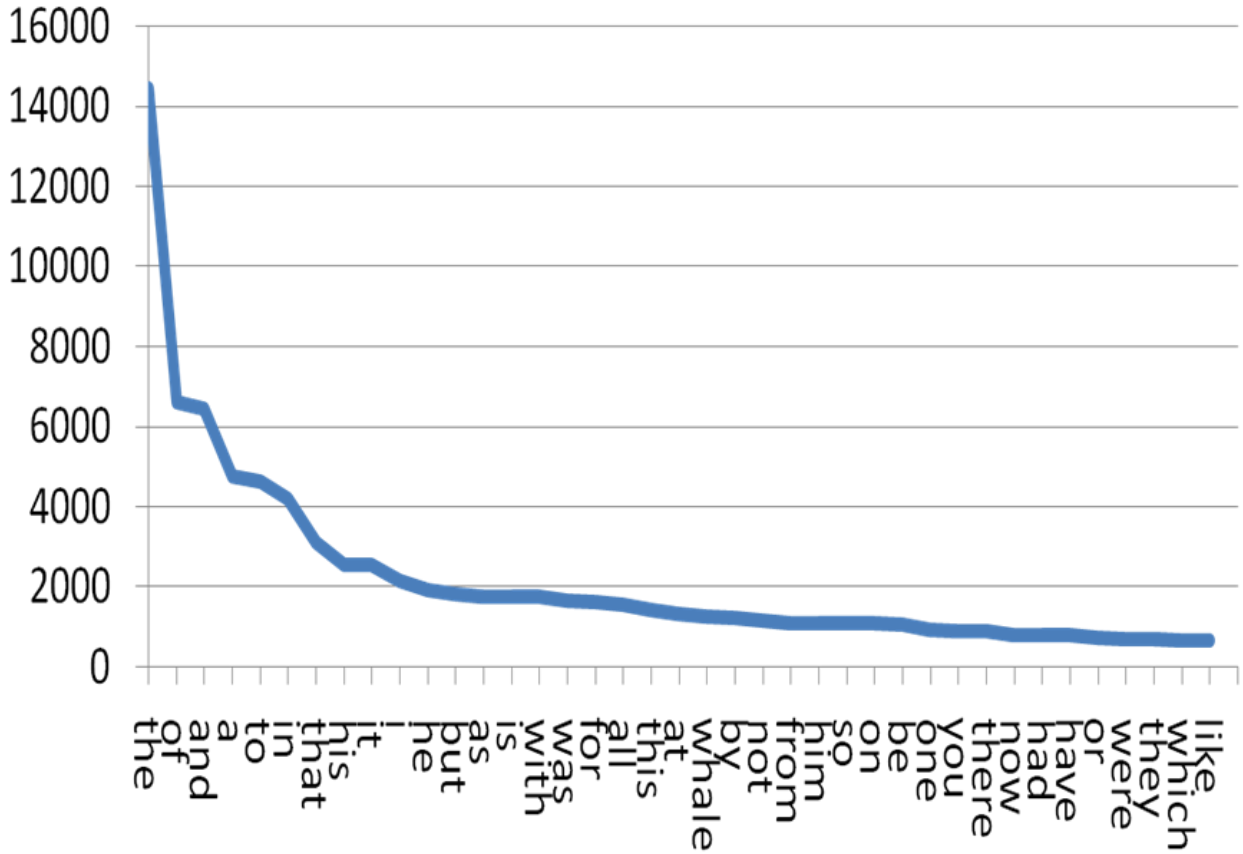
# Distribuições

- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Exemplo?

# Distribuições

- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Por que conhecer esses comportamentos é importante?

Lei de Zipf

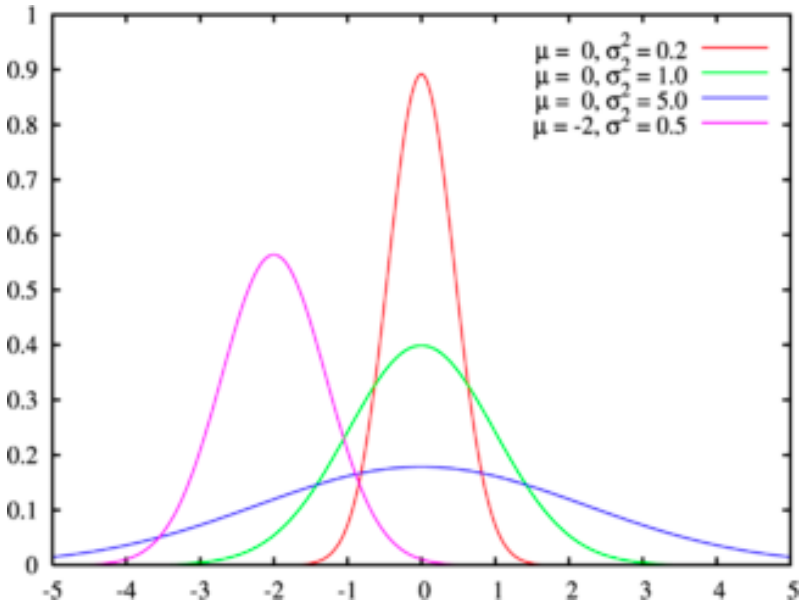


# Distribuições

- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Com alguns parâmetros, podemos descrever ou prever número médio, variações e onde encontrar os fenômenos modelados
      - Em geral, parâmetros são **média** ($\mu$), **frequência**, **desvio padrão** ($\sigma$) ou **variância** ($\sigma^2$)

# Distribuições

- Exemplos
  - Poisson
  - Geométrica
  - Uniforme
  - **Normal (ou gaussiana)**
  - Weibull
  - Pareto
  - Etc.



# Exemplo de uso de estatística

- Análise sintática automática

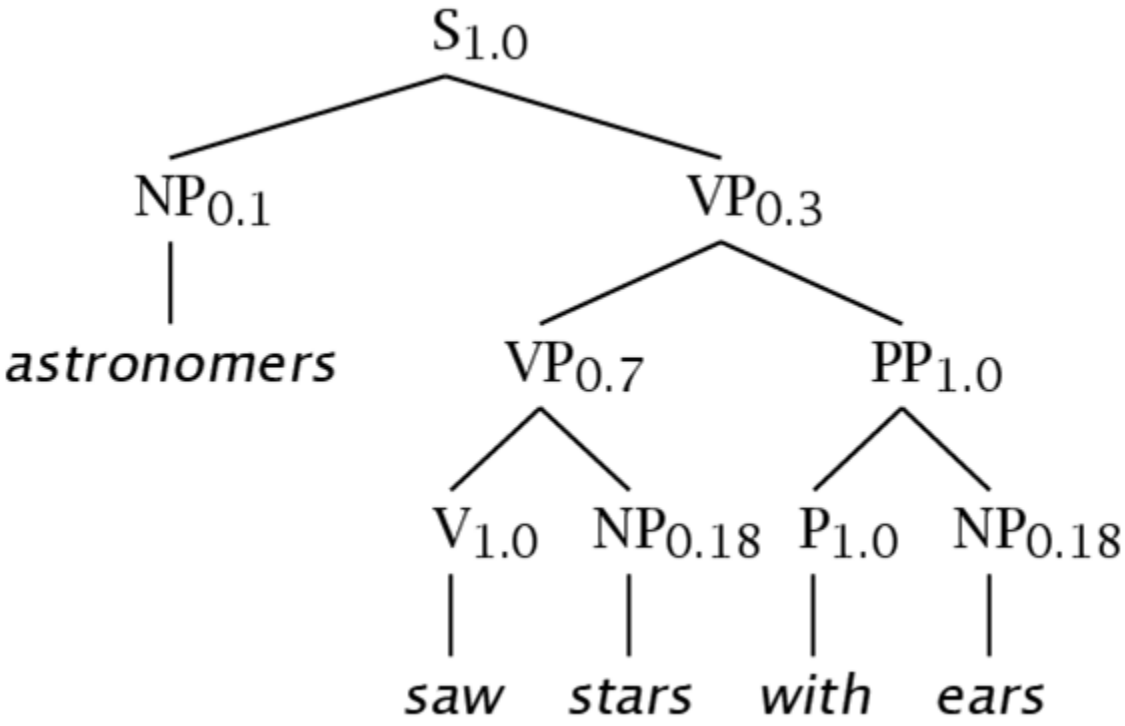
  - Gramáticas probabilísticas

| | | | |
|---|---|---|---|
| S → NP VP | 1.0 | V → *saw* | 1.0 |
| PP → P NP | 1.0 | NP → *astronomers* | 0.1 |
| VP → V NP | 0.7 | NP → *ears* | 0.18 |
| VP → VP PP | 0.3 | NP → *saw* | 0.04 |
| NP → NP PP | 0.4 | NP → *stars* | 0.18 |
| P → *with* | 1.0 | NP → *telescopes* | 0.1 |

  - De onde se conseguem as probabilidades de cada regra?

# Exemplo de uso de estatística

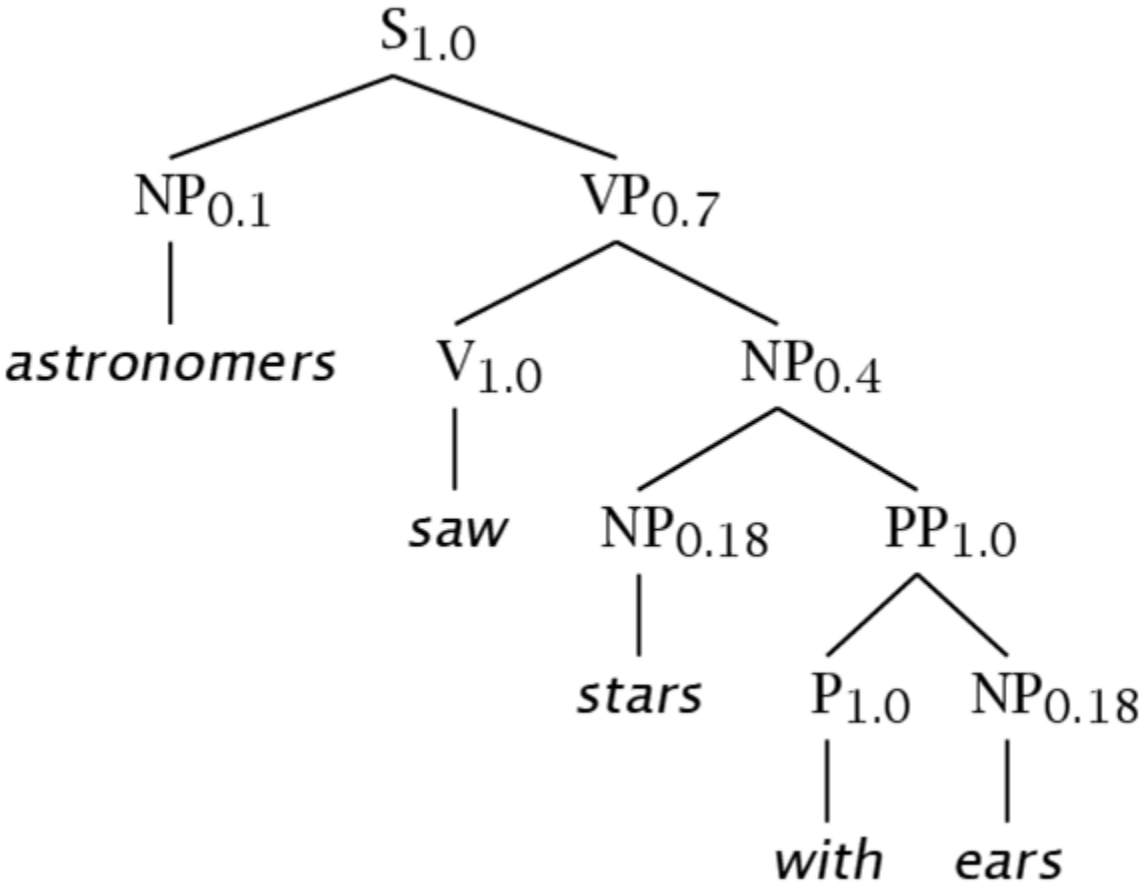
- Possibilidade 1




P(1) = 1.0 * 0.1 * 0.3 * 0.7 * 1.0 * 0.18 * 1.0 * 1.0 * 0.18 = 0.0006804

# Exemplo de uso de estatística

- Possibilidade 2

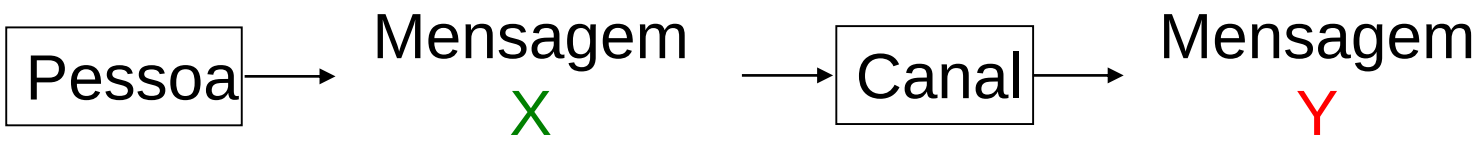



P(2) = 1.0 * 0.1 * 0.7 * 1.0 * 0.4 * 0.18 * 1.0 * 1.0 * 0.18 = **0.0009072** (>anterior)

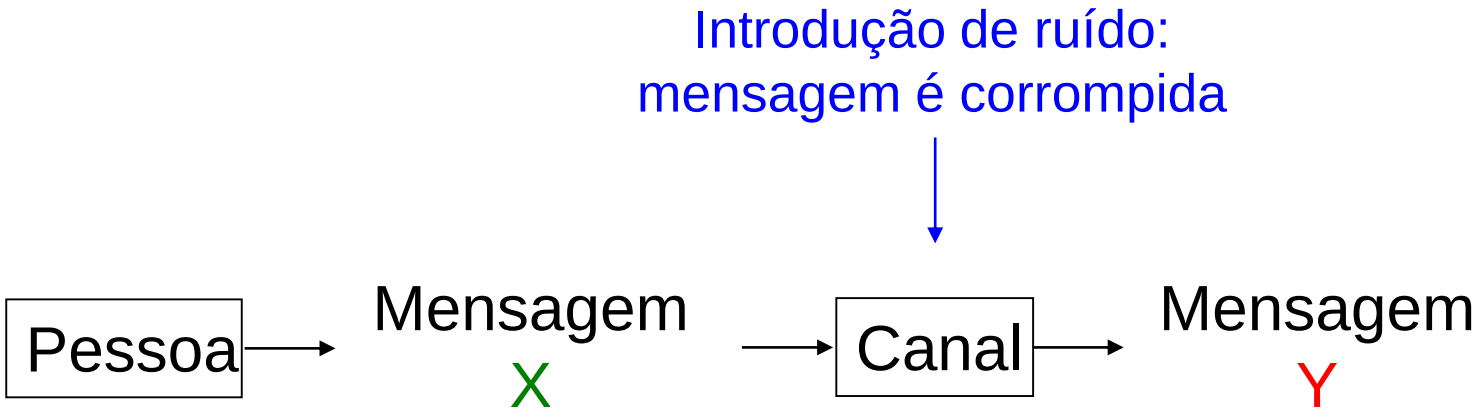
# Modelo *Noisy-Channel*

- Shannon, 1948
  - Teoria da Informação
- Modelo probabilístico
  - No coração do renascimento da estatística no PLN na década de 70
- Transmissão de mensagens pela linha telefônica
  - Capacidade de transmissão por um canal (*channel*)
  - Ocorrência de ruídos (*noise*)
  - Quantidade de informação necessária para recuperação da mensagem original

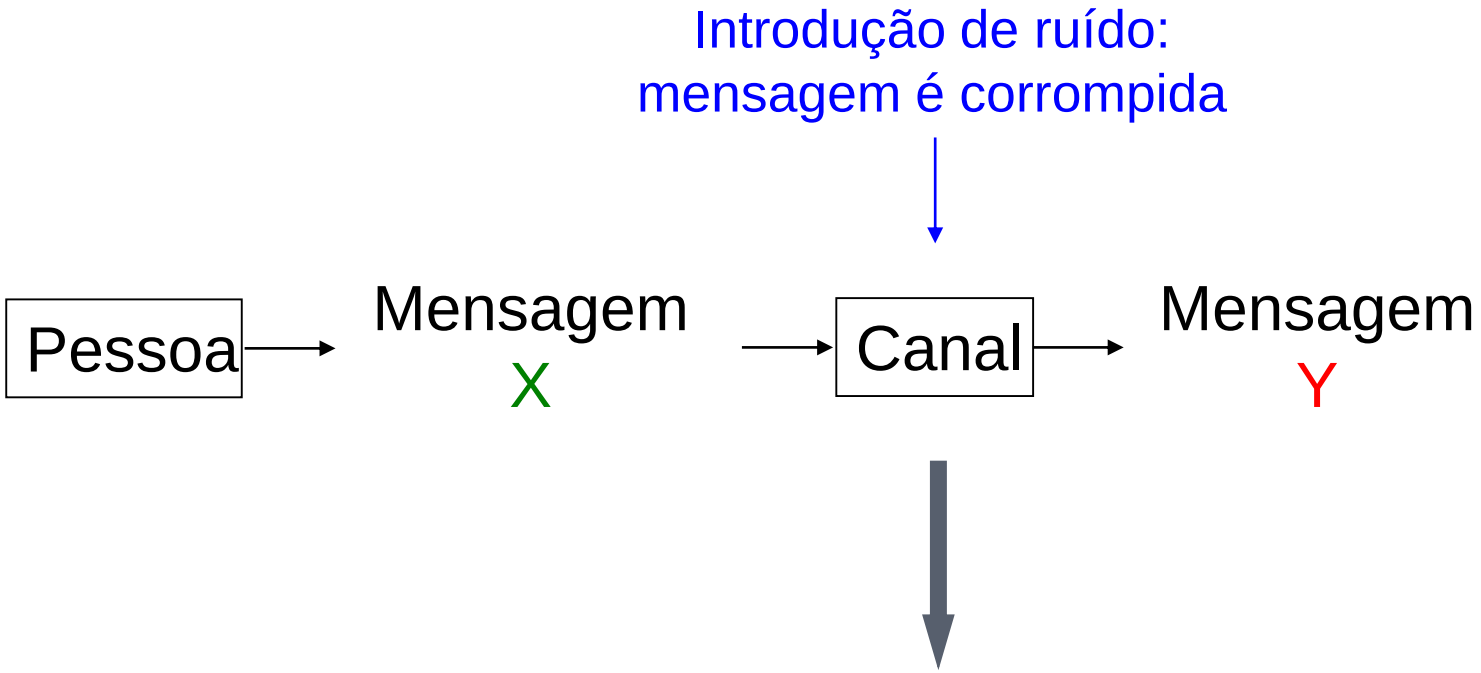
# MODELO *NOISY-CHANNEL*



# Modelo *Noisy-Channel*

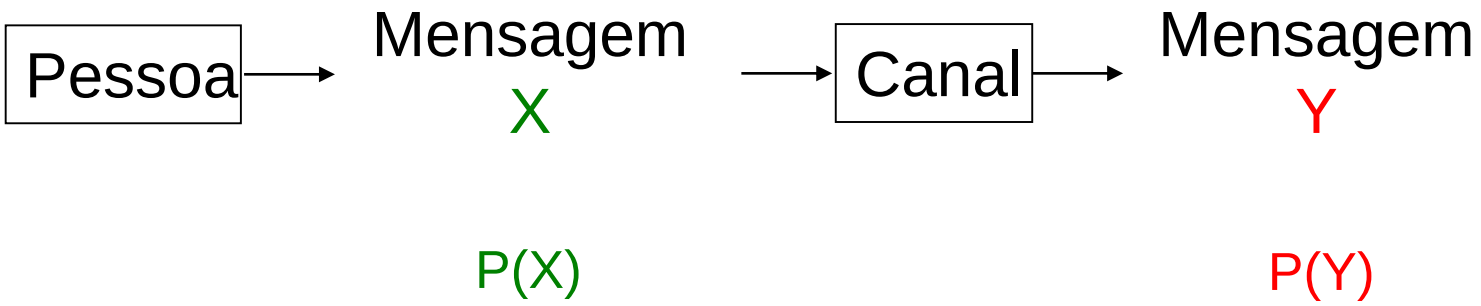


# Modelo *Noisy-Channel*

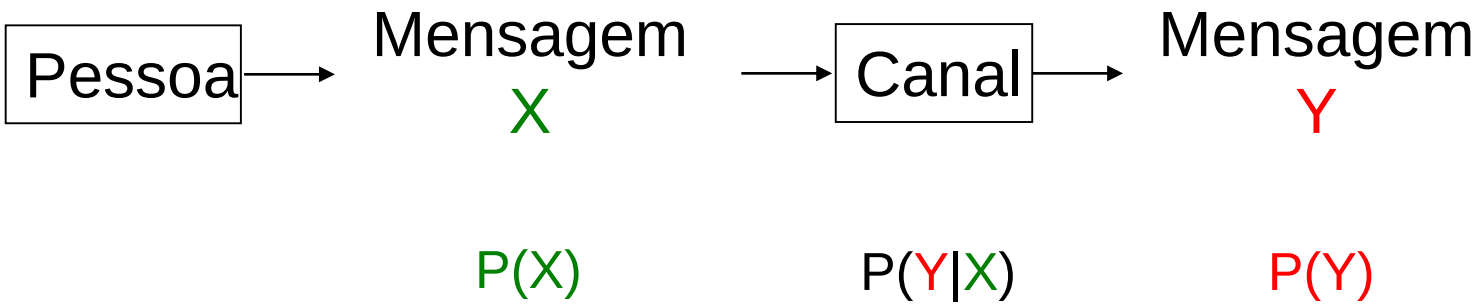



Quanta informação pode ser transmitida?
Quanta informação pode ser transmitida para minimizar a ocorrência de ruídos?
Como e que tipos de ruído ocorrem?
Como recuperar a mensagem original X a partir de Y?

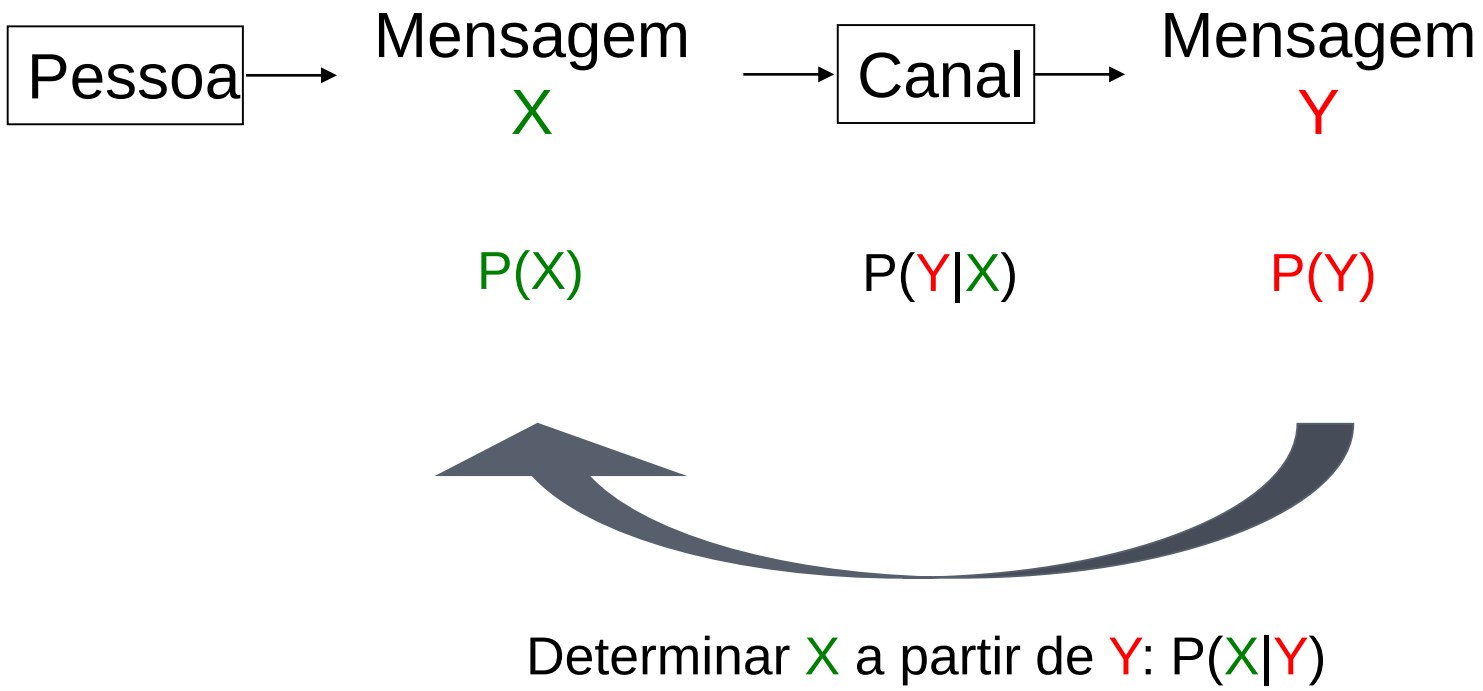
# Modelo *Noisy-Channel*



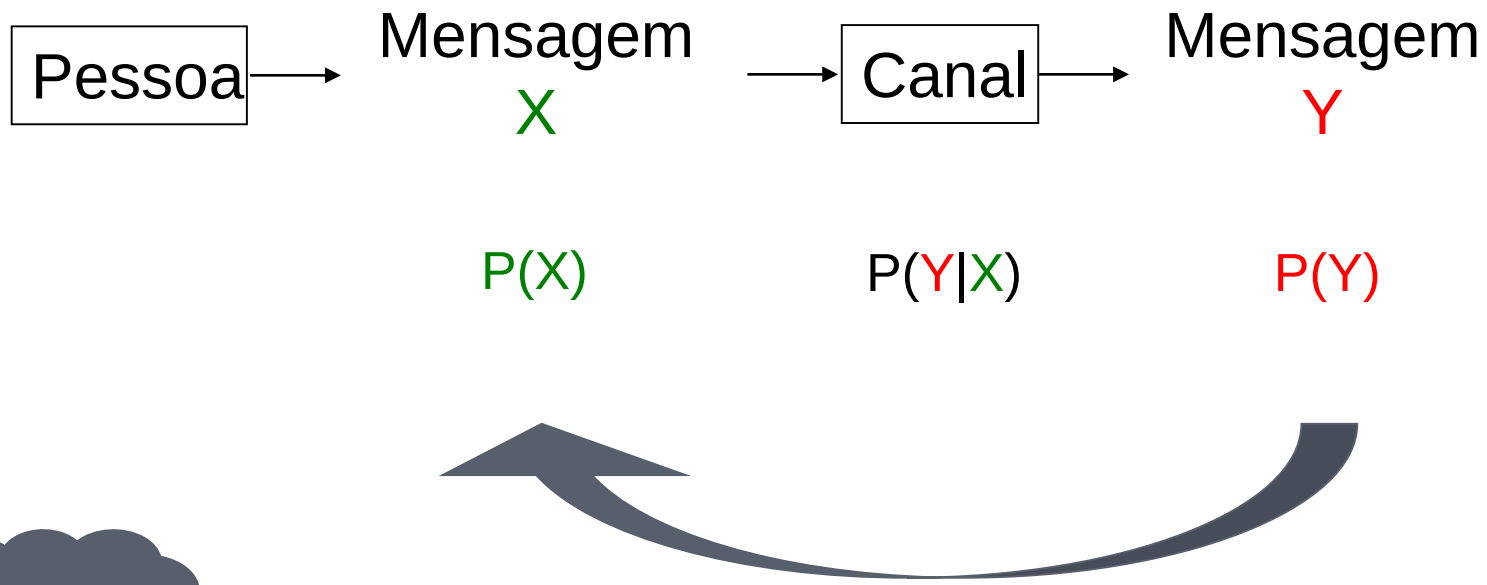
# Modelo *Noisy-Channel*



# Modelo *Noisy-Channel*



# Modelo *Noisy-Channel*




$$P(X|Y) = P(Y|X) \times P(X) / P(Y)$$

# Tarefas

- Leitura no e-Disciplinas
  - Sardinha, T.B. (2000). Linguística de Corpus: Histórico e Problemática. Delta, Vol. 16, N. 2, pp. 323-367.

- Provinha 4 disponível à tarde