# Crosswords and Codebreaking

**Luís Felipe Ramos Ferreira**[1], **Gabriel Fialho**[2], **Diego Pereira**[3]

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

## 1. Introdução

A teoria da informação trata da quantificação, armazenamento e comunicação de informações. Este relatório explora duas aplicações: palavras cruzadas e quebras de códigos, analisando a importância da entropia e da redundância na linguagem. As palavras cruzadas exemplificam como esses conceitos influenciam sua criação e resolução. Na quebra de códigos, o trabalho ilustra como a entropia e a redundância ajudaram os analistas de Bletchley Park a decifrar mensagens criptografadas durante a Segunda Guerra Mundial.

## 2. Palavras cruzadas

Um passatempo clássico comumente encontrado em jornais impressos era o jogo de palavras cruzadas, em que o objetivo é preencher todas os espaços em brancos com palavras, na horizontal ou vertical, que possuem letras em comum, cruzando-se entre si. Por causa disso, a possibilidade da criação de palavras cruzadas depende da existência de palavras com caracteres em comum, uma característica relacionada ao grau de redundância da língua.

### 2.1. Entropia e Redundância da linguagem nas palavras cruzadas

Entropia e redundância são conceitos fundamentais para o entendimento da complexidade das palavras cruzadas. No contexto de palavras cruzadas, a redundância de uma linguagem está relacionada com a facilidade ou dificuldade de criar palavras cruzadas. Uma alta entropia indica uma grande capacidade de comunicação do canal, portanto, mais palavras cruzadas podem existir caso haja uma alta entropia.

Redundância se refere à presença de informações que ajudam a inferir uma mensagem. Nas palavras cruzadas, a redundância ajuda a inferir uma determinada palavra com base no cruzamento de letras. Se uma linguagem não tiver nenhuma redundância, então qualquer combinação de letras em uma grade formam palavras cruzadas válidas. Por outro lado, caso haja muita redundância, é difícil criar palavras cruzadas, já que menos combinações serão possíveis.

#### 2.1.1. Diferenças entre Tipos de Palavras Cruzadas

[MacKay 2005] distingue dois diferentes tipos de palavras cruzadas, o tipo A (americano), em que todos os espaços fazem parte de duas palavras, e o tipo B (britânico), em que não existe essa restrição, e em média metade dos caracteres estão presentes em duas palavras simultaneamente. Por ser mais restritivo, o tipo A é mais difícil de se montar, porém mais fácil de se resolver, pois todas as letras descobertas fornecem uma dica para outra palavra.

**Figura 1. Palavras cruzadas dos tipos A (americano) e B (britânico).**

### 2.1.2. Estudo de Caso

Utilizando métricas da Teoria da Informação, é possível calcular condições para a existências de palavras cruzadas para um certo idioma. [MacKay 2005] faz este cálculo para um modelo chamado "Wenglish". Neste trabalho, usaremos esta simplificação, mas ao invés de ter como base o inglês, que possui um alfabeto com 26 letras, utilizamos o chinês, considerando 2000 ideogramas que são mais comuns no idioma. Chamamos esta língua imaginária de "palavro-chinês", que consiste em $W$ palavras com $L$ ideogramas cada, com entropia $H_W \equiv \frac{\log_2 W}{L+1}$, as $W$ palavras que compõem este idioma são originadas de uma seleção de caracteres aleatória feita por um canal fonte. Se assumirmos uma distribuição com probabilidades iguais para cada ideograma, teríamos uma entropia de $\log_2 2000 = 10,97$, mas vamos considerar na verdade uma fonte com entropia $H_0 = 10$, levando em conta uma distribuição mais factível em que alguns ideogramas são mais comuns que outros e portanto a incerteza é levemente reduzida. Para uma instância de palavras cruzadas com $S$ quadrados, seja $w = f_w S$ o número de palavras e $l = f_1 S$ o número de espaços ocupados com caracteres. A Tabela 1 contém fórmulas para os valores aproximados de $f_w$ e $f_1$, de acordo com [MacKay 2005].

| | A | B |
|---|---|---|
| $f_w$ | $\frac{2}{L+1}$ | $\frac{1}{L+1}$ |
| $f_1$ | $\frac{L}{L+1}$ | $\frac{3}{4}\frac{L}{L+1}$ |

**Tabela 1. Tabela com valores de $f_w$ e $f_1$ para os tipos A e B.**

O próximo passo é calcular o número de palavas cruzadas com $S$ quadrados. Para tanto, calculamos de quantas formas é possível preencher aleatoriamente o quadro: $|T| = 2^{lH_0}$, e a probabilidade de que uma dessas palavras seja válida é $\beta = \frac{W}{2^{LH_0}}$, e $\beta^w$ é a chance de que todas as palavras estejam preenchidas de forma válida. Desse modo, o número total de palavras cruzadas é $\log \beta^w |T| = w(L+1)H_W + H_0(l - Lw) = S[(f_1 - f_w L)H_0 + f_w(L+1)H_W]$, que é uma função crescente de S quando $(f_1 - f_w L)H_0 + f_w(L+1)H_W > 0$. Substituindo com os valores da Tabela 1, con-

cluí-se que é possível formar palavras cruzadas do tipo B quando $H_W > \frac{1}{4}\frac{L}{L+1}H_0$, já para formação no tipo A a condição $H_W > \frac{1}{2}\frac{L}{L+1}H_0$ deve ser satisfeita. Neste modelo, considerando que o "palavro-chinês" tem $W = 4000$ palavras de tamanho $L = 4$, temos $H_W = 2,4$, enquanto $K = \frac{L}{L+1}H_0 = 8$, então é possível criar palavras cruzadas do tipo B, já que $2,4 > \frac{1}{4}K = 2$, mas não existem palavras cruzadas do tipo A, pois $2,4 < \frac{1}{2}K = 4$. Este resultado permite levantar a hipótese de que é extremamente difícil criar palavras cruzadas do tipo A no idioma chinês. De fato, ao fazer uma busca por imagens da palavra "crosswords" (palavras cruzadas em inglês), é possível encontrar facilmente uma instância do tipo A, já ao pesquisar pelo termo em chinês ("填字游戏"), eu não consegui localizar nenhum exemplo do tipo A em que o conteúdo estivesse de fato em chinês.

## 3. Quebra de códigos

### 3.1. Teoria da Informação e Criptografia

No contexto da criptografia, os conceitos sobre Teoria da Informação também podem ser extremamente úteis, principalmente na arte da decifrar mensagens. Em termos gerais, cifrar uma mensagem significa converter o conteúdo textual desta mensagem para um diferente, de modo que reverter este processo (decifrar a mensagem criptografada) seja difícil sem conhecimento dos métodos e/ou chaves utilizadas no momento da cifragem [Singh 1999]. Tal dificuldade pode ser mitigada por um profissional da criptoanálise ao analisar as falhas do processo de cifragem e também as redundâncias do idioma em que o texto original foi escrito.

### 3.2. História e Contexto da Máquina Enigma

Em [MacKay 2005], é citado um exemplo do uso dos conceitos de entropia e informação para a quebra de códigos. Durante a Segunda Guerra Mundial, os exércitos alemães utilizaram uma máquina chamada Enigma para cifrar suas mensagens. A máquina era composta de diversas partes internas, com fiações e rotores, que permitia uma vasta gama de configurações iniciais, um número em torno de $8 \times 10^{12}$ e, a cada dia, uma nova configuração inicial era escolhida. O processo era simples, uma vez que a Enigma se parecia com uma máquina de escrever. O usuário digitava a mensagem e a máquina imprimia a mensagem cifrada. A cada letra selecionada, as configurações da máquina mudavam de maneira determinística, de modo a tornar ainda mais difícil o trabalho de quebrar o código.

### 3.3. Aplicações na Segunda Guerra Mundial

#### 3.3.1. Métodos de Quebra de Código

Os quebradores de códigos de Bletchley Park, onde Alan Turing trabalhou, se aproveitaram de diferentes vulnerabilidades para quebrar os códigos alemães. Como o número de mensagens trocadas entre os exércitos era gigantesca, era natural suspeitar que duas máquinas Enigma, em algum momento, estariam na mesma configuração inicial ao cifrar uma mensagem. Desse modo, os textos originais de ambas mensagens seriam cifrados utilizando as mesmas configurações. A correlação entre a saída de ambas essas

máquinas permitiu que os analistas de cifras de Bletchley Park extraíssem a quantidade de informação necessária para descobrir qual era a configuração inicial da máquina Enigma daquele dia.

Nesse cenário, dadas duas máquinas suspeitas de estarem no mesmo estado, duas hipóteses deveriam ser consideradas. A hipótese nula, $\mathcal{H}_0$, que assume que as máquinas estão em estados diferentes, e a hipótese de que elas estão no mesmo estado, $\mathcal{H}_1$. Para realizar os cálculos necessários, assume-se que os dados disponíveis são dois textos $x$ e $y$ de comprimento $T$, originados de um alfabeto com $A$ caracteres. Assumindo que as máquinas enigma geram permutações distribuídas uniformemente e, portanto, todo texto cifrado possui a mesma probabilidade de ocorrer, é fácil enxergar que, para todas mensagens $x$ e $y$ de comprimento $T$: $P(x, y \mid \mathcal{H}_0) = \left(\frac{1}{A}\right)^{2T}$

### 3.3.2. Análise de Redundância em Mensagens Criptografadas

A hipótese $\mathcal{H}_1$, por sua vez, possui algumas sutilezas que devem ser consideradas. Se o texto original das mensagens cifradas fosse gerado por uma fonte aleatória, $P(x, y \mid \mathcal{H}_1)$ seria igual a $P(x, y \mid \mathcal{H}_0)$ e ambas hipóteses seriam igualmente prováveis. No entanto, sabia-se que os textos originais foram gerados por um idioma e, como sabe-se, idiomas possuem redundâncias, como letras frequentes e digramas ou trigramas que ocorrem constantemente ao fim de palavras e sentenças, as quais foram exploradas pelos quebradores de códigos.

Os quebradores de códigos procuravam, então, por diversos pares de mensagens cifradas que eram muito parecidas umas com as outras, em um sentido de compartilharem digramas, trigramas, etc. O primeiro a se utilizar desse tipo de método foi o matemático polonês Rejewski.

O uso dessa técnicas foi melhorado ainda mais ao longo dos anos, se aproveitando ainda mais de redundâncias contidas nas linguagens humanas e se utilizando de máquinas que podiam automatizar e acelerar diversas etapas do processo. De acordo com I. J. Good, matemática que trabalhou em Bletchley Park, uma sequência de 8 caracteres consecutivos idênticos entre dois textos cifrados foi o maior falso positivo encontrado pela equipe. Apesar de ser raro, se tratavam de dois textos cifrados por máquinas em estados sem relação alguma.

## 4. Conclusão

A teoria da informação é fundamental para estudar comunicação e criptografia. Nas palavras cruzadas, entropia e redundância influenciam na complexidade de criação e resolução. A alta entropia aumenta a diversidade de palavras cruzadas, enquanto a redundância pode facilitar ou dificultar a criação. Na quebra de códigos, a análise da máquina Enigma mostra como a teoria da informação ajudou analistas de Bletchley Park a descriptografar mensagens, alterando o curso da história. O trabalho destaca a aplicação prática da teoria da informação em passatempos e segurança nacional, beneficiando diversas áreas da ciência da comunicação.

## Referências


Alvim, M. S., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., and Smith, G. (2020). *The Science of Quantitative Information Flow*. Springer.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Hinsley, F. H. and Stripp, A. (2001). *Codebreakers: the inside story of Bletchley Park*. Oxford University Press, USA.

MacKay, D. J. C. (2005). *Information Theory, Inference and Learning Algorithms*. 7th edition.

Singh, S. (1999). *The code book*, volume 7. Doubleday New York.