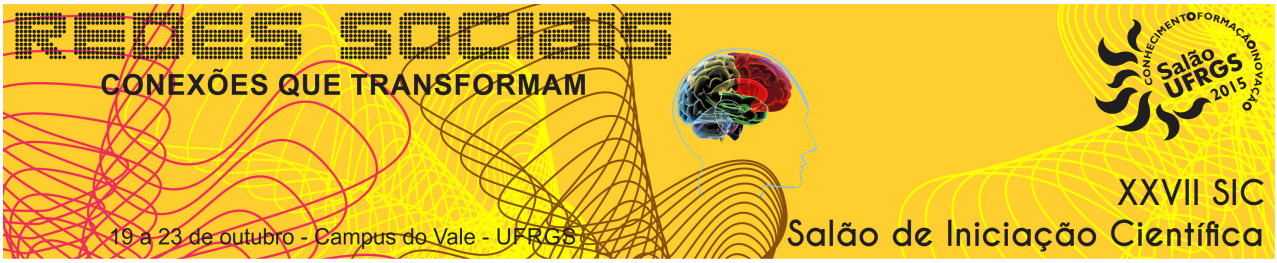

Evento	Salão UFRGS 2015: SIC - XXVII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2015
Local	Porto Alegre - RS
Título	Teoria da informação aplicada aos idiomas
Autor	ANA CAROLINA OSEROW
Orientador	DAVID RENATO CARRETA DOMINGUEZ

TEORIA MATEMÁTICA DA COMUNICAÇÃO APLICADA AOS IDIOMAS

Bolsista: Ana Carolina Oserow
Orientador: David Renato Carreta Dominguez
Universidade Federal do Rio Grande do Sul

A pesquisa realizada ao longo do ano letivo tem o objetivo de caracterizar os principais idiomas que utilizam o alfabeto latino quanto à informação que cada texto transfere a seus leitores, relacionando-os com seu grau de liberdade (entropia) e mostrando com que exatidão os símbolos de comunicação podem ser transmitidos, levando em conta inicialmente apenas aspectos da engenharia e, por conta disso, considerando aspectos semânticos (significado do grupo) e aspectos efetivos (eficiência das transmissões) como problemas secundários que podem ser resolvidos através do desenvolvimento técnico da teoria e de suas aplicações.

A Teoria Matemática da Comunicação desenvolvida por Claude Shannon expõe ideias gerais da comunicação que podem ser descritas pelo seguinte processo: a fonte de informação seleciona a mensagem desejada entre um grupo de mensagens possíveis. O transmissor transforma o recado em um sinal, que é enviado através de um canal de comunicação existente entre o transmissor e o receptor, e, em seguida, enviado ao destino após o processo de transformação do sinal em mensagem. Além de prever como funciona o processo de comunicação, a teoria se estende para conceitos como quantidade de informação, capacidade de um canal e propagação efetiva dos conjuntos de símbolos. Cada idioma tem suas peculiaridades e através dessa teoria é possível inferir seu comportamento e suas características.

Inicialmente, introduzimos o estudo para a língua alemã, espanhola, francesa, inglesa, italiana e portuguesa, que são os principais idiomas que utilizam o alfabeto latino. Para cada idioma, foram utilizadas 100 amostras de poemas para dedução da informação média, que mede a quantidade de informação que a mensagem inicial (enviada pelo transmissor) transmite e é calculada através da expressão: $I(X) = \Sigma p(x).\log2(1/p(x))$, também conhecida como entropia. Esta pode ser relacionada com a entropia de Boltzmann ($H = Kb.\log W$). A informação mútua mede a quantidade de informação comum que é recebida pelo destino após o processo de transmissão e é dada por: $I(X:Y) = H(X)+H(Y)-H(X,Y) = \Sigma p(x).p(y|x)\log2(p(y|x)/p(y))$; onde, $H(X)$ é a entropia de X, $H(Y)$ é a entropia de Y, $H(X,Y)$ é a entropia conjunta de x e y, $H(X|Y)$ é a entropia condicional. É natural que a informação seja medida pela entropia, pois é análoga ao volume de liberdade de escolha que temos para construir uma mensagem. Então, temos que a entropia pode ser calculada como: $H(x) = \Sigma p(x).\log2(1/p(x))$ conhecida também como informação média, e a entropia conjunta: $H(X,Y) = \Sigma p(x,y).\log2(1/p(x,y)$, onde $p(x,y)$ é a probabilidade conjunta dada por: $P(x,y) = P(x)P(y|x)$, onde $P(y|x) = P(x\cap y)/P(y)$ é a probabilidade condicional. Através destes cálculos, obtemos para língua portuguesa a informação média dos caracteres, que é aproximadamente 3, 92109135; podendo chegar a uma informação máxima média de 9, 81946223, mostrando que a língua portuguesa mostra uma informação média de 39,93%. Isso significa que esta fonte em sua escolha de símbolos estará mais ou menos 39,93% tão livre quanto possivelmente poderíamos presumir. O mesmo foi feito para duplas (sequências que podem ser consideradas sílabas), encontrando uma informação média de 6,71376291.

Futuramente, ao decorrer da pesquisa ainda será desenvolvida a informação média, mútua e mútua média para palavras; será estabelecida uma possível relação entre um texto X e sua tradução; a ampliação da pesquisa para textos informais e reportagens; e possíveis aplicações da teoria, como no desenvolvimento de corretores ortográficos avançados.