Seminários do PPG-Inf (05/06/2020)

# Ligados pela linguagem: diálogos possíveis (e necessários) entre linguistas e cientistas de computação

Adelaide H.P. Silva
Supervisor: prof. dr. Fabiano Silva
(LIAMF/DInf/UFPR)

# Estrutura da apresentação

- Como a Linguística aborda a linguagem
- Convergências entre Linguística e Computação
  - arquitetura da linguagem;
  - processamento da linguagem;
  - análise de dados;
  - modelos de análise e ferramentas tecnológicas
    - jogo para ensino de ortografia
- Considerações finais

# Pra começo de conversa...

- Em linhas gerais, Linguística e Computação tomam um mesmo objeto: linguagem;

- Linguística: estudo científico da linguagem humana (cf. Akmajian *et al*, 1995);

- Computação: análise de dados através de linguagens artificiais

# Como a Linguística aborda a linguagem

- Não prescreve “formas certas”;
- Procura depreender padrões de estrutura e funcionamento das línguas, a partir da descrição e análise de dados;
- Foco sobre uso;
- Procura descrever e explicar construções como

  “Ele$_i$ viu ele$_j$ ontem no cinema.” ==> ((Ele)(viu ele)((ontem)(no cinema)))

# Como a Linguística aborda a linguagem

- Ou esta:


- Possibilidade 1: ((Museu Luiz de Queiroz)(recebe mostra)(sobre a origem (do universo em Piracicaba)))
- Possibilidade 2: ((Museu Luiz de Queiroz)(recebe mostra)(sobre a origem (do universo)) (em Piracicaba))

# Como a Linguística aborda a linguagem

- Nota

- Possibilidade 1: ((Museu Luiz de Queiroz)(recebe mostra)(sobre a origem (do universo em Piracicaba)))

- Possibilidade 2: ((Museu Luiz de Queiroz)(recebe mostra)(sobre a origem (do universo)) (em Piracicaba))

- Diferentes relações de precedência (sintaxe) se refletem sobre a semântica dos enunciados

# Como a Linguística aborda a linguagem

- Nota adicional:

- Para algumas abordagens da Linguística, sintaxe é independente da semântica (e.g. “*Colorless green ideas sleep furiously.”)*

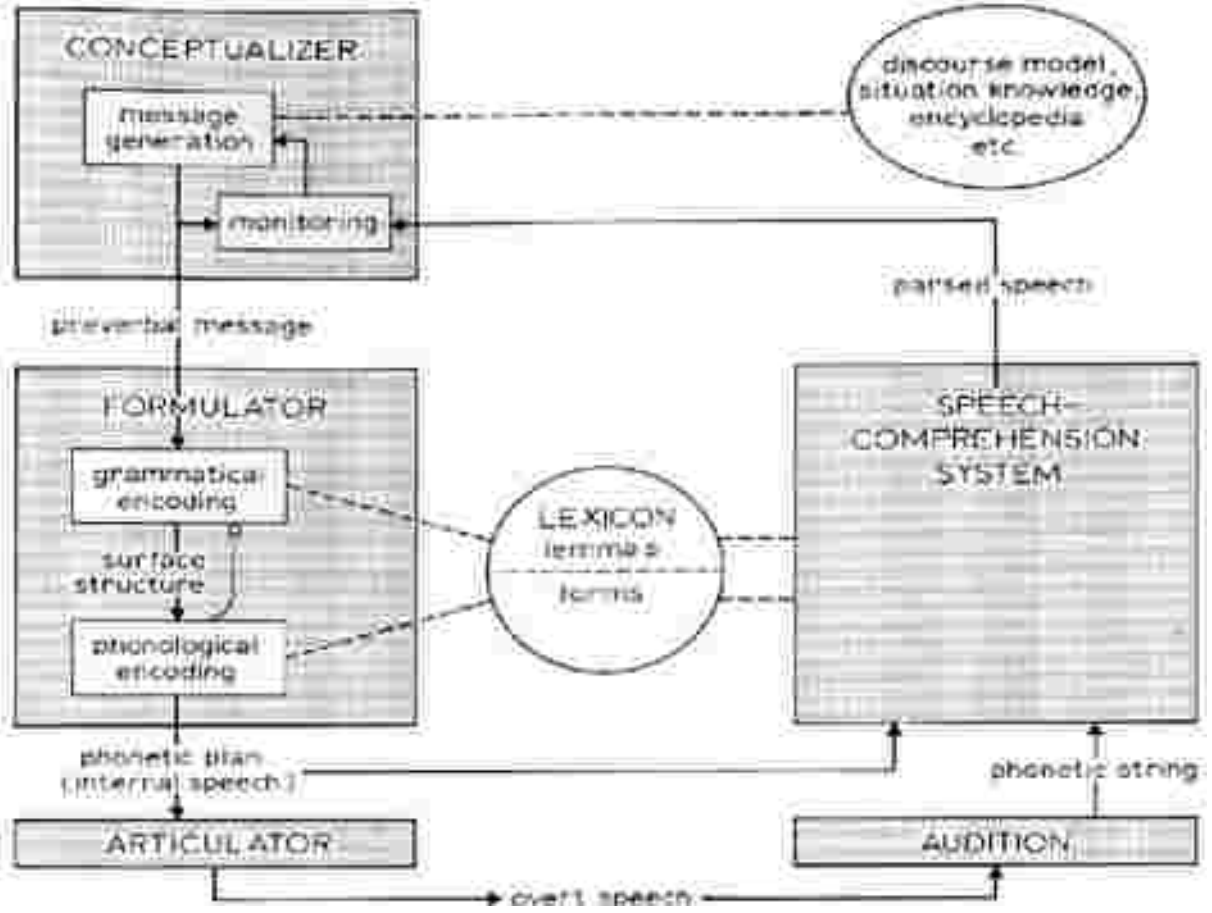
- Possível analogia ==> código com sintaxe correta, mas com estrutura lógica (semântica) problemática

# Convergência entre Linguística e Computação: arquitetura da linguagem

- Linguística: linguagem constituída de “partes/módulos”; compreensão da estrutura e do funcionamento de cada parte contribui para a compreensão sobre estrutura e funcionamento do todo;

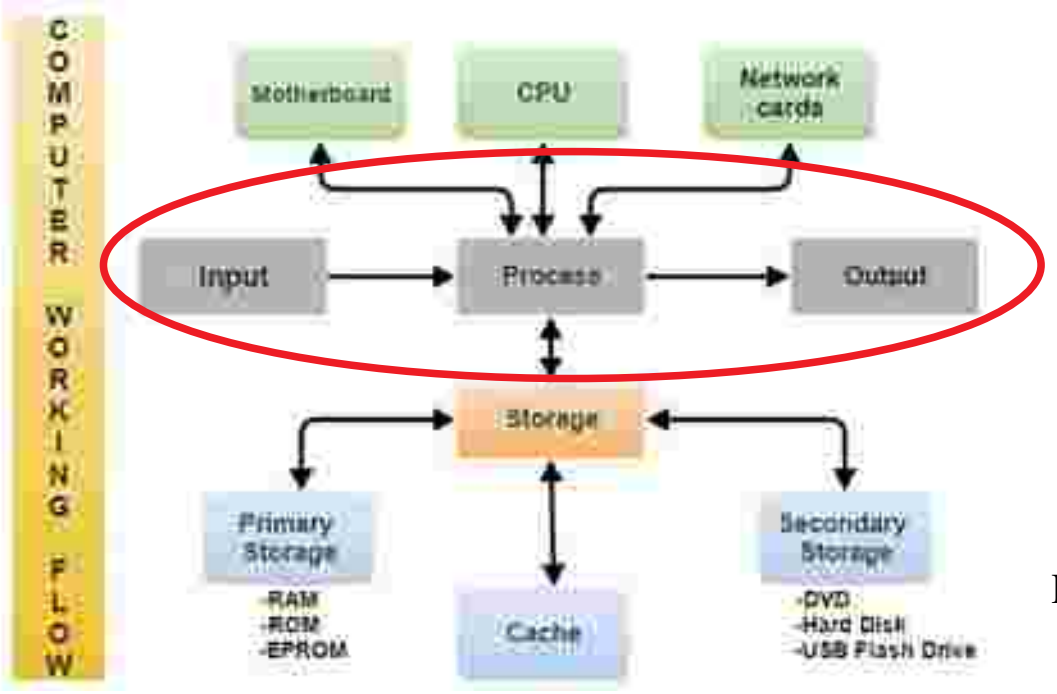
- Programação: “just a few building blocks suffice to enable us to write programs that can help solve all sorts of fascinating, but otherwise unapproachable, problems” (cf. Sedgewick; Wayne, 2017)

# Convergência entre Linguística e Computação: processamento da linguagem




Esquema de um falante, ou como um indivíduo gera a fala, desde sua intenção até sua articulação, cf. Levelt (1989)

# Convergência entre Linguística e Computação: processamento da linguagem




https://i.pinimg.com/736x/50/8b/f2/508bf298870b221d105cdfc120cb9965--device-computers.jpg

Processamento de dados semelhante ao processamento de linguagem humana. Memória envolvida igualmente em ambos os processos.

# Convergência entre Linguística e Computação: análise de dados

- Linguística: recorre à coleta e análise de dados para testar hipóteses e propor representações (modelos);

- Computação: recorre à coleta e análise de dados para propor representações (modelos);

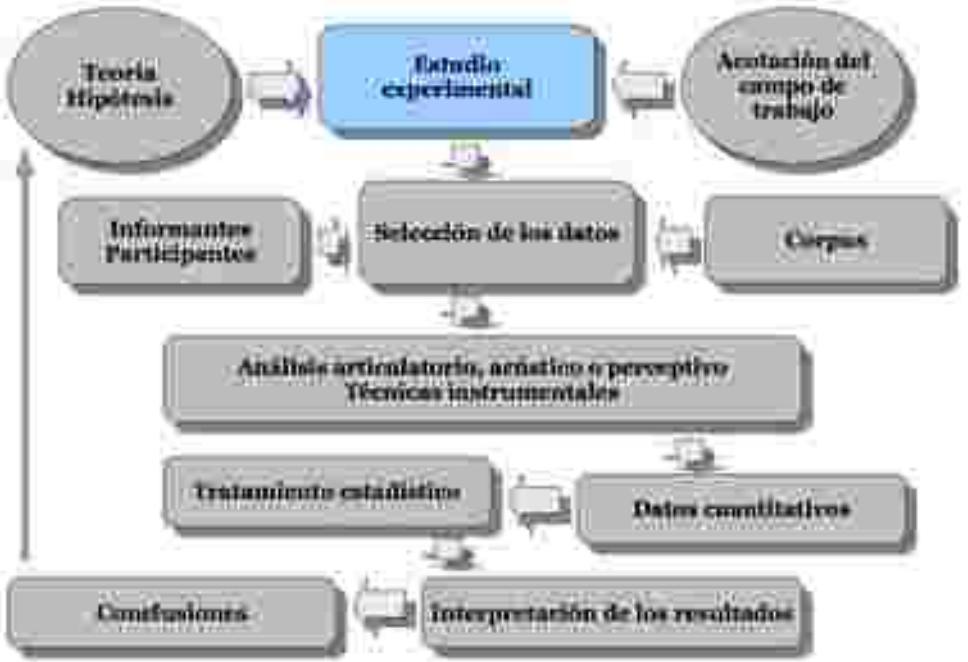
- Processamento de linguagem (Linguística) e de informação (Computação) concebidos de modo semelhante.

# Convergência entre Linguística e Computação: análise de dados

- Convergência entre as duas ciências nem sempre é direta (nem tem de ser);

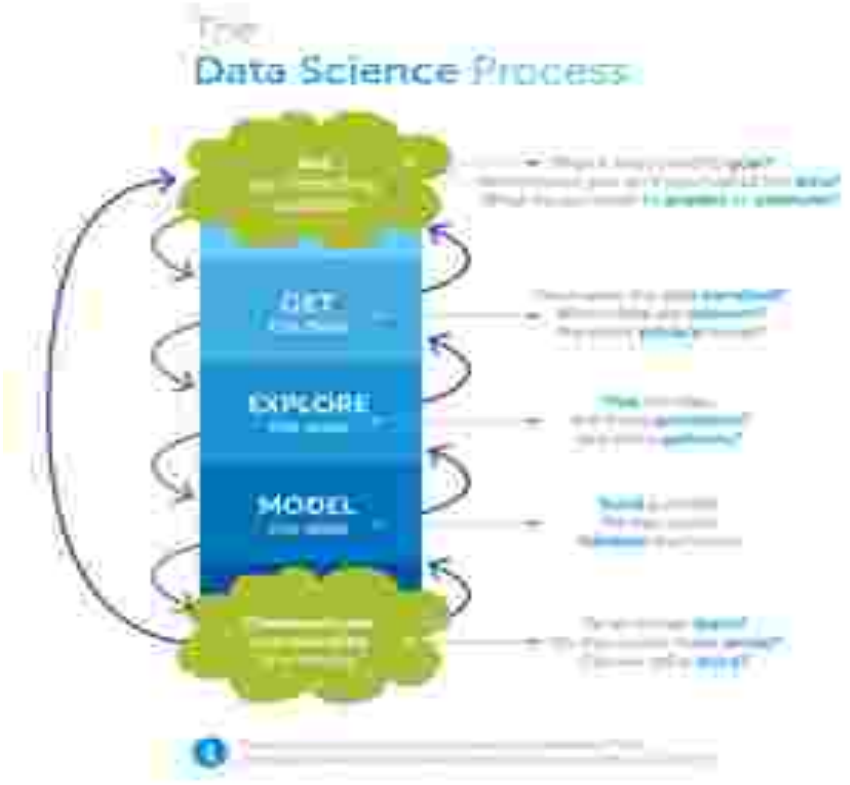
- Uma ciência pode se valer da outra para lidar melhor com dados

# Convergência entre Linguística e Computação: análise de dados




Llisterri, J. El método experimental en fonética (http://liceu.uab.cat/~joaquim/phonetics/fon_met_exper/met_exp_fon.html)

# Convergência entre Linguística e Computação: análise de dados




http://www.cherhan.net/data-scientist-really/

# Convergência entre Linguística e Computação: análise de dados

http://www.ilc.it/img-mescolanze/dialettinpuzzle.gif

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Interação entre Computação e Linguística

1) permite testagem de hipóteses sobre estrutura e funcionamento das línguas naturais;

2) possibilita elaboração de ferramentas que utilizam línguas naturais para diferentes finalidades

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Primeiras "conversas" entre Computação e Linguística tomando linguagem humana como objeto: desenvolvimento de "máquinas falantes"

  - síntese automática de fala ("Voder" - *Voice Operating Demonstrator*, 1938);

  - tradução automática;

  - simulação de comunicação entre humano e máquina através da combinação de padrões (Eliza, 1964 – https://en.wikipedia.org/wiki/ELIZA);

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

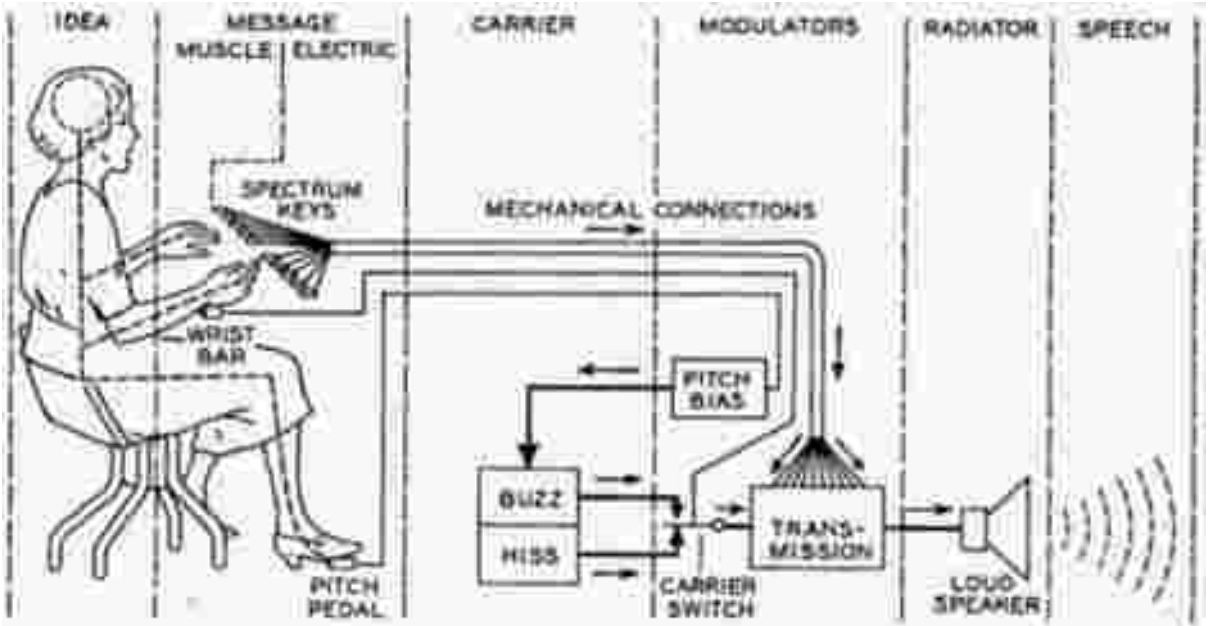



Fig. 8—Schematic circuit of the voder.

https://frank-lovisolo.fr/WordPress/wp-content/uploads/2036/06/Schematic-Circuit-of-the-VODER-560x321.jpeg

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Linguística: sintaxe gerativa (Chomsky, 1957);

- Concepção de que linguagem humana pode ser reduzida a uma estrutura lógico-formal do tipo

  S((SN(Det,N)),(SV(V(SN(Det,N)))));

- Linguagem humana não é sistema fechado (criatividade) ==> composicionalidade e recursividade;

- Montague (década de 1970): aplicação de técnicas desenvolvidas para linguagens formais (e.g. cálculo lambda) à descrição do significado nas línguas naturais;

- Na proposta de Montague, “constituintes formados pelas regra sintáticas são simultaneamente interpretados pelas regras semântica correspondentes” (Borges Neto *et al.*, 2012:131).

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Processamento de Linguagem Natural: maior proximidade com linguística;

  - utilização de inferência bayesiana, por exemplo, para detecção de padrões empregados no aprendizado de máquina

- Envolve conjunto de métodos para “tornar linguagem humana acessível aos computadores” (cf. Eisenstein, 2019).

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Desenvolvimento de ferramentas largamente utilizadas ("sofisticação linguística"):

  - tradução automática (e.g. Google Translator; DeepL; Reverso Context);

  - reconhecimento de fala/sistemas de diálogo (em assistentes pessoais, como Alexa, Siri, Google Assistant);

  - síntese de fala (Natural Readers, From Text to Speech, Google);

  - classificação de textos (e.g. em servidores de e-mail);

  - *voice-to-text* (Simon Says);

  - sistemas de busca na internet;

  - produção de textos, como resumos ou narrativas.

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Linguística Probabilística: efeitos probabilísticos da ocorrência de sequências de sons (gradiente) se refletem, de alguma maneira, sobre a representação mental da fonologia (discreta), cf. Pierrehumbert (no prelo);

- Efeitos probabilísticos são obtidos pela aplicação de modelos bayesianos ao nível sonoro das línguas e a probabilidade de ocorrência de um som, ou uma sequência deles, é associada a descritores fonológicos (e.g., consoantes e vogais);

- Estudos experimentais em psicolinguística e fonologia de laboratório (anos 90 e 2000): grande correlação entre probabilidades fonológicas, estimadas a partir de grandes *corpora*, e entre vários tipos de comportamentos linguísticos.

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Exemplo da correlação entre probabilidades fonológicas e tipos de comportamentos linguísticos: julgamento de aceitabilidade de sequências de sons em pseudopalavras x tamanho da palavra (cf. Pierrehumbert, no prelo)

  - Julgamento de boa formação de pseudopalavras reflete os *scores* estatísticos das formas e é determinado a partir de uma análise do léxico;

  - *Score* fonotático mais simples e mais largamente utilizado assume que as palavras são geradas por um modelo bigrama, definido a partir de fonemas (não de letras), com *scores* calculados pela multiplicação das probabilidades de transições, assumindo-se que cada transição é independente daquela que a precede e que o conjunto de probabilidade de dois eventos independentes é o produto das suas probabilidades individuais, obtido pela multiplicação das probabilidades de transições (modelo de Markov);

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Exemplo da correlação entre probabilidades fonológicas e tipos de comportamentos linguísticos: julgamento de aceitabilidade de sequências de sons em pseudopalavras x tamanho da palavra

  - Previsão (Miller, 1957): assumir um modelo tipo “n-grama” significa que as palavras longas constituídas de uma quantidade maior de sequências prováveis terão *scores* de julgamento de aceitabilidade semelhantes ao de palavras curtas constituídas de sequências menos prováveis;

  - Previsão sobre os *scores* se reflete nas avaliações que os humanos fazem sobre aceitabilidade de sequências de sons;

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Exemplo da correlação entre probabilidades fonológicas e tipos de comportamentos linguísticos: julgamento de aceitabilidade de sequências de sons em pseudopalavras x tamanho da palavra

- Frisch, Large, & Pisoni (2000): coleta de avaliações sobre aceitabilidade de sequências sonoras em pseudopalavras de 2 a 4 sílabas, contendo tanto sílabas CV frequentes como sílabas CV raras;

  - Itens dissílabos constituídos de sílabas raras tiveram avaliação semelhante a pseudopalavras de quatro sílabas, contendo sílabas frequentes;

  - contagem fonotática mais simples assume que palavras são geradas por um modelo de bigramas definido sobre fonemas, com *scores* calculados a partir de testes de julgamento de boa formação de pseudopalavras;

  - consequência: resultados sugerem que probabilidades são cognitivamente relevantes e que informação probabilística se combina de modo cumulativo.

# Convergência entre Linguística e Computação: modelos de análise e ferramentas tecnológicas

- Plano inicial: corretor ortográfico plurilíngue para aplicativos de texto (tempo real);

- Plano atual: “jogo” para fixação de normas ortográficas (linguística probabilística + *machine learning*);

- Motivação para mudança de plano: pandemia e necessidade de atividades educacionais remotas;

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Por que ortografia?

- Aspecto mais “visível” de falta de domínio de norma padrão em textos;

- Indivíduo “não sabe escrever” x hipóteses que o falante constrói sobre correspondência entre fala e escrita;

# Convergência entre Linguística e Computação: jogo para ensino de ortografia


https://www.soportugues.com.br/secoes/maltratando/objetos.jpg

- transposição da fala (como o indivíduo percebe o que produz) para a escrita;
- generalização da "regra" de que o som [ʒ] é grafado com <g> antes de <e>

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Como ensinar/aprender ortografia?

  - pouca generalização possível, como

    <m> anota a nasalidade da vogal que o antecede, na mesma sílaba se, na sílaba seguinte, o som inicial é grafado com <p> ou <b>, como em “campo”; “bambu”;

    <n> anota a nasalidade da vogal que o antecede, na mesma sílaba se, na sílaba seguinte, ocorrem outras consoantes que não sejam <p> nem <b>.

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Outra possibilidade de generalização:

  - <rr> utilizado no início de sílaba, meio de palavra, entre vogais, pra escrever o “r forte”, que é o som que encontramos em “carro”;

  - <r> utilizado nos demais casos, isto é , em final de sílaba/palavra; grupo consonantal; início de sílaba e de palavra (onde representa o som como o do início de “rato”); início de sílaba, meio de palavra, entre vogais (onde representa o som da consoante que ocorre em “arara”); em início de sílaba, meio de palavra, depois das consoantes <l>, <n> (como “guelra”; “honra”).

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Outros casos: ortografia guarda “informação histórica” da língua, como <ch> (em “macho”, por exemplo) ou <x> no prefixo “ex-”;

- Falta de correspondência entre fala e escrita/ escolhas (aparentemente) arbitrárias;

- Decorre: ensino/aprendizagem de ortografia envolve memória visual.

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Objetivo do projeto: criar um programa, associado a um jogo, que possibilite aos usuários aprender ortografia de maneira contextualizada;

- Público visado: alunos do Fundamental II (6º ao 9º ano);

- O que se espera que saibam nesse nível: correspondências letra-som "transparentes", como as estabelecidas pelos grafemas <p, b, t, d, f, v>. (Ausência de dificuldades decorre da inexistência de outro grafema para anotar o som inicial de "bola" a não ser <b>.

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- O que já existe?


Assis *et al.* (2017) - Grapphia

Criança deve escolher entre duas alternativas oferecidas (50% de chance de acerto). Narrativa criada pelos autores do *software*. Faixa etária visada: 08 a 10 anos.

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Desafio para a elaboração do *software*:

  - quais dificuldades ortográficas frisar em cada momento escolar?

  - BNCC (2018)– diretrizes para o Ensino Fundamental

  "Fono-ortografia

  Conhecer e analisar as relações regulares e irregulares entre fonemas e grafemas na escrita do português do Brasil. Conhecer e analisar as possibilidades de estruturação da sílaba na escrita do português do Brasil."

  (http://basenacionalcomum.mec.gov.br/abase/#fundamental/lingua-portuguesa)

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Desafio para a elaboração do *software:*

  *"... como a ortografia é tratada entre nós mais como tema de verificação que de ensino sistemático, a maioria das escolas do país funciona sem planejar o que espera conseguir na promoção da competência ortográfica de seus alunos a cada série. E como quem não tem metas não antevê aonde quer chegar, não planifica sua ação... pode não conseguir progressos significativas no rendimento que seus alunos expressam ao escrever." (Morais, 2000:66)*

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Decisão: gradação na dificuldade envolvida na grafia das palavras, partindo das relações mais transparentes para as menos transparentes;

- Hipóteses a verificar:

  - variável “grau de transparência”: deve haver “erros” mais recorrentes relacionados ao grau de transparência na correspondência grafema/fonema;

  - variável “frequência de uso”: quanto maior o uso de um item lexical, menor a probabilidade de ser grafado “errado” ;

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Requisitos:

  - sistema deve poder aprender as principais dificuldades encontradas por um usuário ao resolver o teste (jogo);

  - sistema deve poder verificar se um grupo de indivíduos de um mesmo nível de educação formal apresenta as mesmas dificuldades;

  - banco de textos e de palavras a serem testadas;

  - jogo deve ser “atraente”.

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Arquitetura do jogo:

  - palavras devem preencher lacunas de um texto (e.g. autores com obra em domínio público);

  - usuário escuta a palavra para, em seguida, grafá-la;

  - número fixo de tentativas de grafia da mesma palavra (3);

  -resposta aberta (possibilita verificação de dificuldades além da contemplada no teste – e.g. “muxoxo”)

  - verificação de mais de um aspecto relacionado à correspondência letra/grafema (?)

# Convergência entre Linguística e Computação: jogo para ensino de ortografia

- Expectativas:

  - verificação das dificuldades de um indivíduo: necessário focalizar um aspecto específico da correspondência grafema/fonema? Ou mais de um aspecto?

  - verificação das dificuldades de um grupo de indivíduos de mesmo nível de instrução formal: existem dificuldades que se tornam menos frequentes à medida que os indivíduos progridem?

  - planejamento sobre o que ensinar em cada momento da instrução formal

# Considerações finais

- Relação simbiótica entre Computação e Linguística tem longa tradição;
- Linguística auxilia Computação a melhorar sistemas que processam linguagem humana, por exemplo;
- Computação auxilia Linguística a verificar hipóteses sobre estrutura e funcionamento de línguas;
- Elaboração de ferramentas que manipulam linguagem natural pode oferecer alternativas na relação ensino-aprendizagem e suprir lacunas de naturezas várias;
- Ortografia: nível mais visível (ponta do *iceberg*), mas produção de textos é outro grande desafio;
- Pra terminar esta conversa: relação entre as duas ciências tem de se expandir.

# Referências


- Akmajian, A.; Farmer, A.; Bickmore, L.; Demers, R.; Harnish, R. *Linguistics - An Introduction to Language and Communication,* Cambridge (MA): The MIT Press, 1995.
- Assis, L.; Bodolay, A.; Gregório, L.O.; Santos, M.J.; Vivas, A.; Pitangui, C.; Perry, D. Grapphia: Aplicativo para Dispositivos Móveis para Auxiliar o Ensino da Ortografia. In *Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação, 2017.* DOI: 10.5753/cbie.wcbie.2017.609
- Borges Neto, J.; Müller, A.; Pires de Oliveira, R. A semântica formal das línguas naturais: história e desafios. In Revista Estudos Linguísticos, Belo Horizonte, v. 20, n. 1, p. 119-148, 2012.
- Chomsky, N. *Syntactic Structures*. The Hague: Mouton,1957.
- Eisenstein, J. *Introduction to Natural Language Processing.* Cambridge (MA): The MIT Press, 2019.
- Ilari, R.; Basso, R. O português da gente: a língua que estudamos, a língua que falamos. São Paulo: Editora Contexto, 2007.
- Morais, A.G. Ortografia: ensinar e aprender. São Paulo: Editora Ática, 2000.
- Pierrehumbert, J. 70+ years of probabilistic phonology. In B. Elan Dresher and Harry van der Hulst (Eds.) *Oxford Handbook on the History of Phonology*, Oxford University Press (no prelo)
- Sedgewick; Wayne, K. Introduction to programming in Java – an interdisciplinary approach. Pearson Education Inc., 2017.