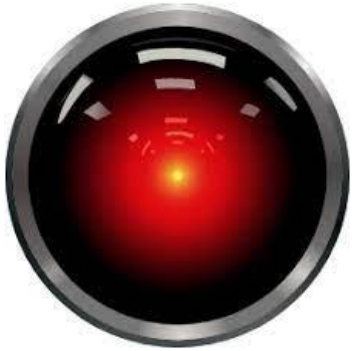
PRG0018 Processamento de Linguagem Natural

Prof. Thiago A. S. Pardo

ICMC USP SÃO CARLOS
50 anos 1971 - 2021
Instituto de Ciências Matemáticas e de Computação
| Universidade de São Paulo |

Retomando a partir da última aula

- O mundo de PLN
  - De produtos “relativamente simples” (como corretores ortográficos) a sistemas “altamente sofisticados” (como os assistentes virtuais gerativos)
    - Atenção para as aspas ☺
  - HAL e as ambições clássicas da IA e do PLN



Questões existenciais da área

- O que significa "processar" a língua humana?
  - Diferentes perspectivas: áreas diferentes, expectativas diferentes, hardwares diferentes
- O que significa ser inteligente?
  - Processar a língua é um tipo de inteligência?
- Podemos ou devemos nos inspirar no ser humano?
  - De onde vem a língua?
  - E se passar no Teste de Turing?
  - IA fraca é pior do que IA forte?

Luger (2013) em seu livro clássico sobre inteligência artificial

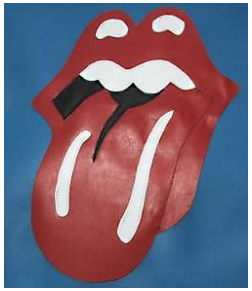
- *O problema de definir o campo inteiro da inteligência artificial é semelhante ao de definir a própria inteligência: ela é uma única faculdade ou é apenas um nome para a coleção de capacidades distintas e não relacionadas? Até que ponto a inteligência é aprendida e não existe desde o nascimento? O que acontece exatamente quando ocorre o aprendizado? O que é criatividade? O que é intuição? A inteligência pode ser deduzida do comportamento observável ou ela requer evidências de um mecanismo interno em particular? Como o conhecimento é representado no tecido nervoso de um ser humano e que lições isso traz para o projeto de máquinas inteligentes? O que é autopercepção? Que papel ela desempenha na inteligência? Além disso, o conhecimento sobre a inteligência humana é necessário para construir um programa inteligente, ou uma técnica estritamente de "engenharia" é suficiente para tratar o problema? É possível conseguir inteligência em um computador, ou uma entidade inteligente requer a riqueza de sensações e experiências que só poderiam ser encontradas em uma existência biológica?*

Muitos desafios e questões difíceis

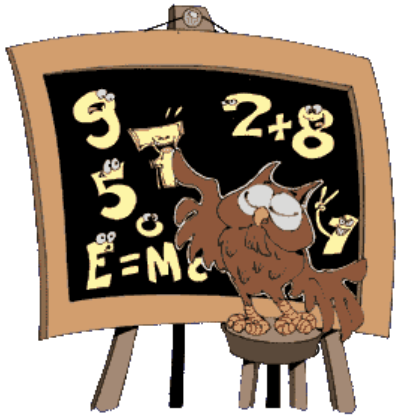
- Vamos começar do básico

Língua Natural

- Língua humana



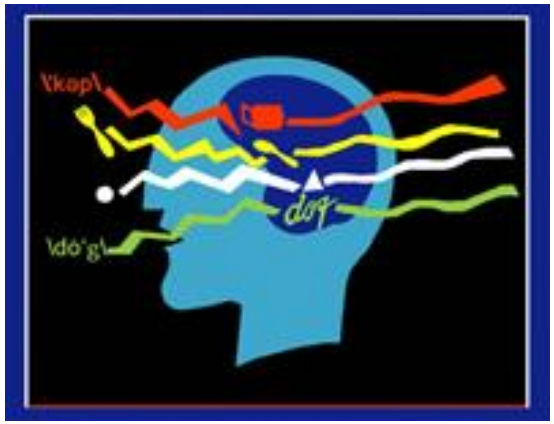
- Em oposição às linguagens artificiais
  - Matemática, lógica, linguagens de programação de computadores



PLN

- Processamento de Língua Natural
  - Linguística Computacional
  - Processamento de Linguagem Natural
  - Engenharia das Línguas Naturais

- No Brasil, tradicionalmente visto como subárea da Inteligência Artificial & Computação
  - Habilidade linguística é vista um tipo de inteligência



Questão

- Qual a diferença entre “língua” e “linguagem”?
  - É Processamento de Linguagem Natural ou Processamento de Línguas Naturais?

Linguagem & língua

- Linguagem: capacidade humana de comunicação e suas manifestações, de forma verbal ou não
  - Fala, gestos, música, dança, pintura, *um sorriso*
  - Envolve nosso aparato físico e mental/cognição

- Língua: código de comunicação utilizado por uma comunidade, com suas regras específicas
  - Português, Inglês, LIBRAS, etc.

PLN: um pouco de história

- Nascimento na 2ª guerra mundial
  - Tradução automática

- Possíveis nomes
  - *Computational Linguistics*
  - *Mechanolinguistics*
  - *Automatic Language Data Processing*
  - *Natural Language Processing*

PLN: um pouco de história

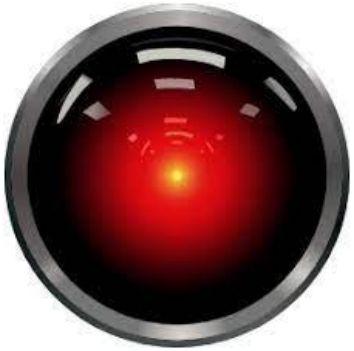
- Trajetória da Inteligência Artificial
  - Primeiros trabalhos → grande expectativa → resultados pobres → desilusão e hibernação da área → novos horizontes
    - Relatório da ALPAC (*Automatic Language Processing Advisory Committee*): *Languages and Machines – computers in translation and linguistics*

PLN: um pouco de história

- Globalização, internet, tecnologia da informação, Google e demais *big techs*
  - Mais recentemente, smartphones, redes sociais, modelos distribucionais, aprendizado profundo, *big data* e ciência de dados, internet das coisas

Questão

- Já conseguimos fazer um computador como o HAL? Do que precisamos?



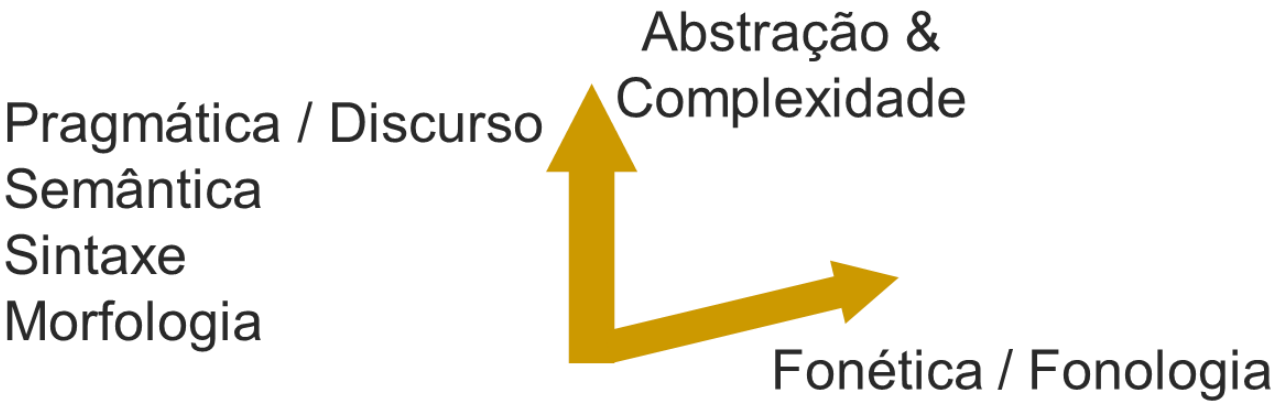
Para construir um computador como o HAL (Jurafsky e Martin, 2008)

- Requer um volume enorme de conhecimento de uma dada língua
  - Reconhecimento (faz até leitura labial) e síntese de fala (fonética e fonologia)
  - Conhecimento das palavras envolvidas (morfologia e vocabulário)
    - Significado (semântica) e como combinam (uso das palavras)
  - Como grupos de palavras de juntam (gramática)
  - Manter um diálogo (discurso)
    - É educado responder... mesmo que você queira matar alguém (HAL)
    - É educado ser cooperativo... mesmo que esteja fingindo (HAL)
- O uso de língua natural também pressupõe conhecimento do mundo e de senso comum

Como esse "volume enorme de conhecimento" pode ser adquirido, representado e utilizado?

A estrutura da língua

- Níveis de representação e processamento linguístico



PLN & IA

- Classificações de abordagens... nem sempre triviais

Critérios	**Paradigmas**
Uso de conhecimento linguístico	Superficial, profundo e híbrido
Representação do conhecimento	Simbólico, sub/não-simbólico e híbrido
Obtenção do conhecimento	Manual, automática e híbrida

Superficial vs. profundo

- Superficial
  - Normalmente, mais simples aplicação e desenvolvimento, mais robusto
  - Resultados piores, normalmente
- Profundo
  - De mais difícil modelagem e aquisição
  - Resultados melhores, para domínios limitados, muitas vezes
- Híbrido: como fazer?
- Métodos profundos “explicam” a língua, mas alguns métodos superficiais são muito bons
  - Por exemplo, sumarização de notícias jornalísticas
- “Métodos cada vez mais sofisticados para fazer a mesma coisa”
  - Dilema da sumarização automática

Simbolismo vs. estatística/matemática

- Regras são muito “rígidas” para a fluidez e flexibilidade da língua
  - Por exemplo, regras gramaticais para boa formação de sentenças
- Padrões mais frequentes de organização da língua podem ser aprendidos (estatisticamente)
- Mas alguns tipos de regras são muito bons
  - Regras de formação de sintagmas nominais
  - Explicitam o conhecimento
    - Dedução e indução
- Tendência atual: rumo aos modelos neurossimbólicos

Abordagens conflitantes

- Simbolismo/profundidade e a **validação de teorias e modelos**
  - Explicitação do conhecimento

- Grande **utilidade** dos números
  - O conhecimento está lá... “codificado” (controverso)
    - Dilemas da TA estatística/neural
      - Funciona melhor que outras abordagens, codifica conhecimento, conhecimento pode estar errado (quem se importa?)

Abordagens: PLN

- *The key to automatically processing human languages lies in the appropriate combination of symbolic [rationalist] and non-symbolic [empiricist] techniques*

  (Robert Dale, 2000)

História do PLN

- Direcionada por correntes filosófico-linguísticas
  - Às vezes complementares
  - Às vezes “rivais até a morte”

Racionalismo

- 1960-1985: racionalismo entre linguistas, informatas, etc.
  - Racionalismo: crença de que parte significativa do conhecimento humano não vem dos sentidos, mas é herdada geneticamente
- Noam Chomsky
  - Linguagem inata
    - Argumento: *muito pouco estímulo para um aprendizado muito eficiente de algo complexo*
      - Como é possível aprender tanto a partir de tão pouca evidência linguística?
- IA: sistemas com muito conhecimento manualmente fornecido e com mecanismos de inferência

Para ler em casa

- *Por que somos o único bicho com linguagem?*
  https://super.abril.com.br/ciencia/por-que-somos-o-unico-bicho-com-linguagem/

  - *Porque só a gente é capaz de se expressar como em tantos poemas que conhecemos. Bem… em termos. Na verdade, poesia assim é para poucos, como Carlos Drummond de Andrade, mas os seres humanos se destacam entre outras espécies consideradas inteligentes, como chimpanzés e golfinhos, porque, entre outras coisas, são capazes de encaixar uma ideia na outra, formando frases quilométricas, sem fim. Esse componente, presente apenas na linguagem da nossa espécie, é chamado de recursividade.*
  - *Para o linguista americano Noam Chomsky, que há mais de 5 décadas estuda esse assunto, o que nos torna diferentes é que temos uma espécie de "órgão da linguagem" no cérebro, que talvez nem tenha surgido com esse fim, mas para realizar cálculos combinatórios. Daí a ideia de que a recursividade seja o fato que torna a linguagem humana única...*

Empirismo

- 1920-1960: empirismo
  - Mente não vem com princípios e procedimentos pré-determinados
  - Mas vem com operações gerais de associação, reconhecimento de padrões e generalizações
    - Importância do estímulo sensorial para o aprendizado da língua

- Linha dominante na atualidade
  - Aprendizado automático

Empirismo

- Não temos como observar uma quantidade muito grande de uso da língua em seu contexto no mundo
- Alternativa: textos
  - *Corpus* e *corpora*
    - Ou **córpus**, simplesmente
- Firth (1957): *You shall know a word by the company it keeps*
- *Como é possível aprender tão pouco a partir de tanta evidência linguística?*
  - Questão importante para a área de Aprendizado de Máquina

Racionalismo vs. empirismo

- Racionalismo
  - Linguística a la Chomsky (*gerativismo*)
    - Descrição do módulo linguístico da mente humana, sendo córpus somente evidência indireta
      - "Regras" e "princípios" que regem/geram a linguagem

- Empirismo
  - Descrição da língua em uso, representada em córpus

Racionalismo vs. empirismo

- Distinção importante de Chomsky (1965)
  - Competência linguística: conhecimento da língua pelo falante
    - Foco do racionalismo/gerativismo
      - Argumentam que é possível isolar esse componente para estudo e formalização
  - Desempenho linguístico: afetado por vários fatores, como memória disponível, distrações do ambiente, etc.
    - Foco do empirismo

Racionalismo vs. empirismo

- Linguística a la Chomsky
  - Princípios categóricos
    - Sentenças satisfazem ou não

- Empirismo
  - Usual e “não usual”
    - Preferências, padrões mais comuns, convenções

Argumento contra princípios categóricos

- Exemplos no inglês
  - *Near*: adjetivo ou preposição?
    - Adjetivo: *We will review that decision in the near future.*
      - Evidências: entre determinante e nome, pode formar um advérbio pela adição de *-ly*
    - Preposição: *He lives near the station.*
      - Evidências: componente principal da frase locativa que complementa o verbo *live* (papel clássico de preposições), pode ser modificado por *right*
    - Adjetivo e preposição: *We live nearer the water than you thought.*
      - Evidências: forma comparativa (*-er*) é marca registrada de adjetivos, age como preposição ao ser o componente principal da frase locativa

Abordagens: PLN

- Domínio atual: empirismo
  - **Córpus** para estudo e formalização de fenômenos, verificação e validação de hipóteses, evidências linguísticas, aprendizado de máquina
- Tratamento de exceções
  - Modelos simplistas vs. sofisticados
    - Modelos simplistas → má impressão original da área
- Atenção aos “erros”

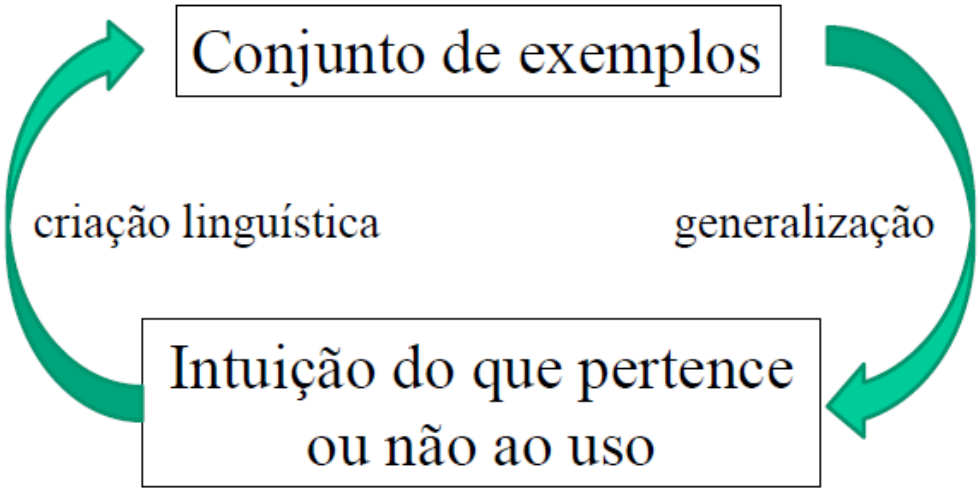
Abordagens: PLN

- Eric Laporte (2012) - *linguista*
  - As diferenças já não são evidentes
    - “Todo gerativista usa o Google escondido”
    - “Todo empiricista usa seu conhecimento e intuição”

Ufa! Assunto encerrado! Ou não?

Abordagens: PLN

- Eric Laporte (2012) - *linguista*
  - Dualidade córpus/introspecção

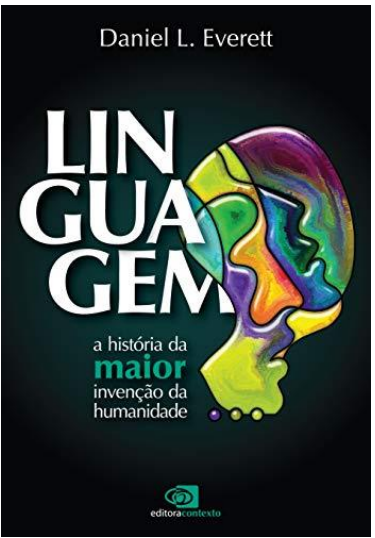


Gerativismo vs empirismo

- Daniel Everett
  (entrevista para a revista Veja, em 2012)

“Durante cinco décadas, os linguistas seguiram a teoria da gramática universal, concebida por Noam Chomsky. De acordo com essa teoria, a gramática e a linguagem são inatas ao ser humano e já vêm programadas no cérebro. Acho essa ideia ridícula. Nunca houve provas de que existem estruturas em nosso cérebro ou em nosso DNA que nos autorizem a dizer que a linguagem é hereditária. O célebre gene FOXP 2, que por um tempo foi classificado como o gene da linguagem e prova da gramática universal, tem na verdade múltiplas funções. Ele atua no desenvolvimento dos pulmões, dos controles dos músculos da face e define mais uma dezena de funções no organismo. O FOXP 2 tampouco é exclusivo do homem. Os ratos, alguns pássaros e outros animais têm esse mesmo gene.”

2019



Gerativismo vs empirismo

- Noam Chomsky, Ian Roberts e Jeffrey Watumull
  (entrevista para a Folha de São Paulo, em 2023)

  - A falsa promessa do ChatGPT
    Versão mais proeminente de Inteligência Artificial codifica uma concepção errônea de linguagem e conhecimento

    “Observe, apesar de todo o raciocínio e linguagem aparentemente sofisticados, a indiferença moral originária da falta de inteligência. Aqui, o ChatGPT exibe algo como a banalidade do mal: plágio, apatia e obviação. Ele resume os argumentos padrão da literatura por uma espécie de "superautocompletar", recusa-se a assumir posição sobre qualquer coisa, alega não apenas ignorância, mas falta de inteligência e, finalmente, apresenta uma defesa de "apenas seguir ordens", transferindo a responsabilidade para seus criadores. Resumindo, o ChatGPT e seus irmãos são constitucionalmente incapazes de equilibrar criatividade com restrição. Eles supergeram (ao mesmo tempo produzindo verdades e falsidades, endossando decisões éticas e antiéticas) ou subgeram (demonstrando falta de compromisso com quaisquer decisões e indiferença com as consequências). Dada a amoralidade, falsa ciência e incompetência linguística desses sistemas, podemos apenas rir ou chorar de sua popularidade.”

Resumo da história de PLN em mais detalhes

- Avanços da área no tempo
  - 1940-56: fundação da área
    - Máquinas de estados finitos, gramáticas e modelos probabilísticos
  - 1957-70: dois campos
    - Simbolismo vs. estatística e os primeiros córpus on-line
  - 1970-83: quatro paradigmas
    - Estocástico, lógico, interpretação textual, discurso

Resumo da história de PLN em mais detalhes

- Avanços da área no tempo
  - 1983-93: empirismo
    - Probabilidades, avaliação, geração textual
  - 1994-99: fortalecimento da área
    - Modelos baseados em dados, exploração comercial, web
  - 2000-atual: aprendizado de máquina
    - Semissupervisão e não supervisão, aprendizado sem fim, aprendizado profundo
    - Competições e grandes conjuntos de dados
    - Modelos distribucionais
    - Tendências: modelos multimodais, neurossimbólicos

PLN e áreas/tópicos correlatos

- PLN e mineração de textos têm tópicos em comuns (como representações textuais, mineração de opinião, etc.), mas têm focos um pouco diferentes
  - Em PLN, a língua em si (além da aplicação visada) é um objeto de interesse
  - A língua tem diferentes níveis de representação
- PLN e ciência de dados compartilham métodos e, muitas vezes, ciência de dados faz uso de PLN, mas não são a mesma área
- PLN vai muito além de aprendizado de máquina e do popular aprendizado profundo
  - Aprendizado de máquina é uma técnica possível em PLN
  - IA clássica tem grande peso em PLN, em especial, as técnicas de representação de conhecimento (como redes semânticas, frames e regras)
  - Há mais em PLN do que as aplicações computacionais, por exemplo, em PLN há interesse nos mecanismos de processamento mental da língua, na evolução das línguas, em modelos linguísticos formais, etc.

Leitura interessante: Manning, 2015

Tendências no mundo

- Tópicos de pesquisa
  - E-mails, mensagens, redes sociais e *User Generated Content* (UGC)
  - Mineração de opiniões
  - Assistentes/agentes inteligentes
  - Abordagens multimodais
- Entrada da indústria no cenário

Tendências no mundo

- Aplicações *cross-language*
  - Apesar de possíveis limitações de PLN
- Robustez, escalabilidade e independência de língua
  - “Deve funcionar para qualquer coisa na web”
- Atenção aos minoritários

Dilemas no Brasil (mas não só no Brasil)

- Multidisciplinaridade, mas...
  - Formação especializada e fragmentada
  - Ainda há desafios de interação
- Texto & fala
  - Comunidades ainda diferentes

Os desafios do português

Famílias de línguas

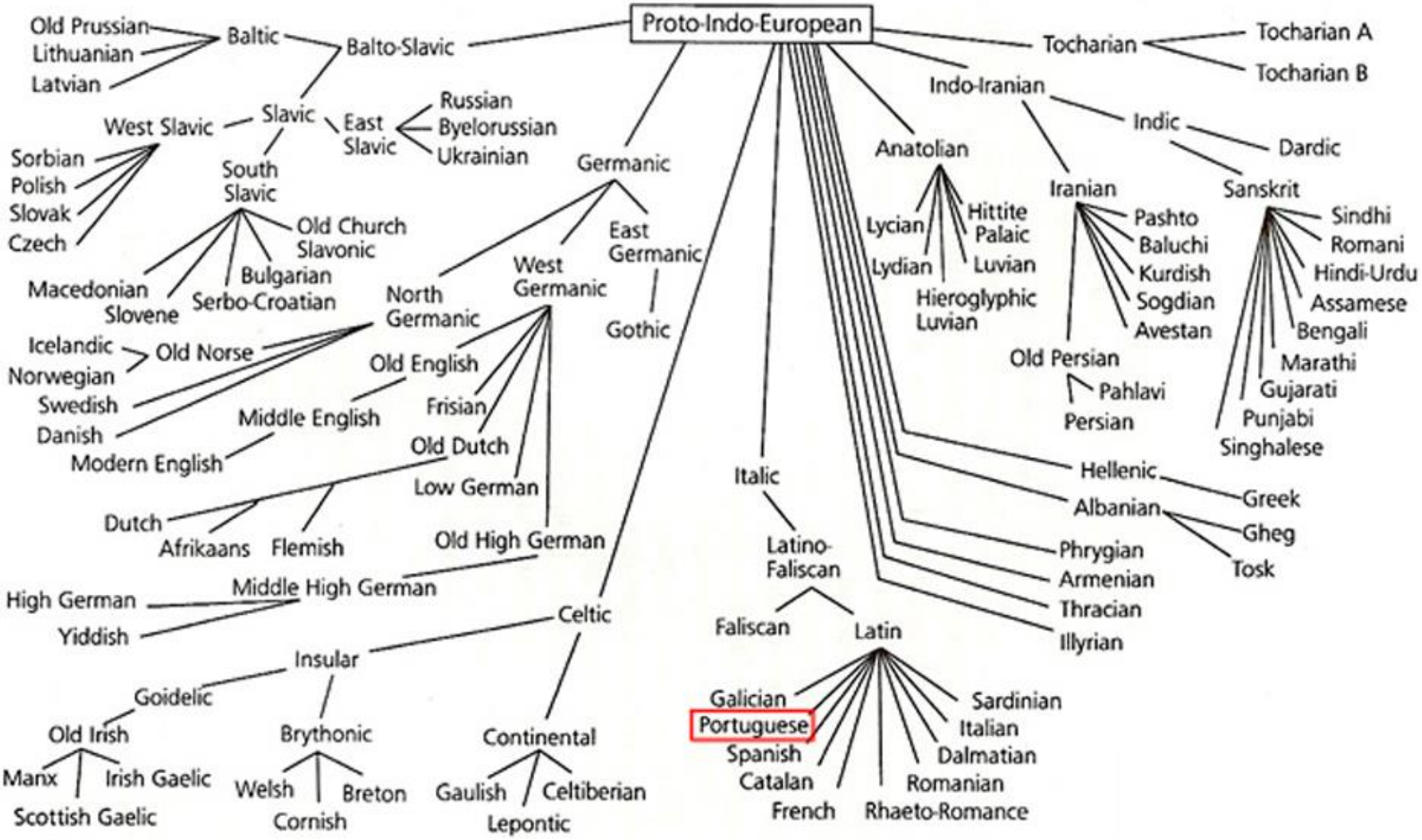
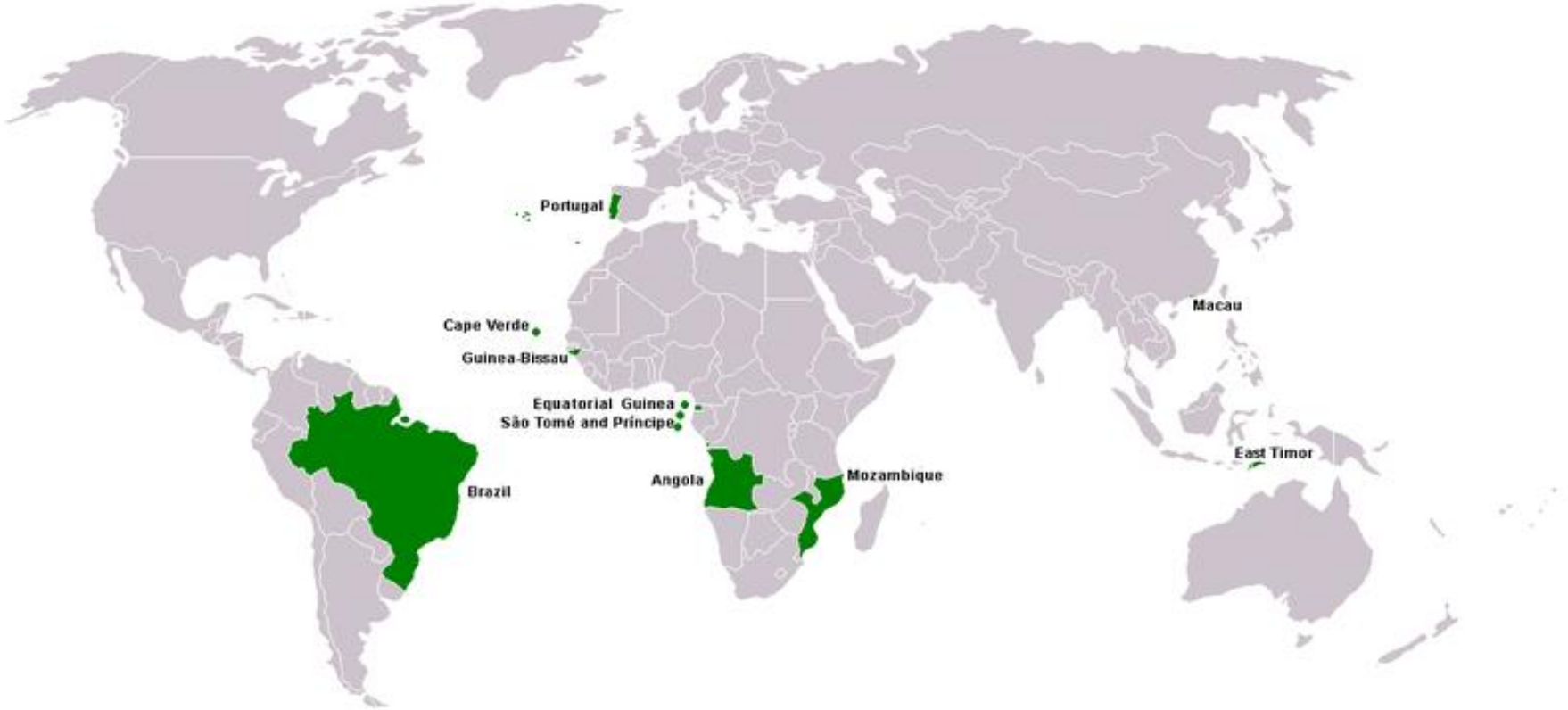


Imagem obtida no site: http://www.ime.usp.br/~tycho/participants/psousa/images/indo_europeu_700.jpg

O Português no Mundo

- Falado em 10 territórios



Dados sobre a língua

- Português
  - 6ª língua mais falada no mundo
  - Instituto Internacional da Língua Portuguesa (IILP)
    - Divulgação e promoção da língua portuguesa
  - Grande número de vocábulos
    - Dicionário Houaiss
      - 228.500 entradas
      - 376.500 acepções
      - 415.500 sinônimos
      - 26.400 antônimos
      - 57.000 palavras arcaicas
    - Academia Brasileira de Letras: mais de 356.000 palavras

Desafios para processamento

- **Variações** nos países em que é falado
  - Pronúncias variadas
  - Ortografias variadas, apesar do acordo ortográfico
  - Diferenças nas construções sintáticas mais usuais
  - Sentidos diferentes de palavras
  - “Perfis de usuários” diferentes: hábitos, expectativas e cultura variados

Desafios para processamento

- Dialetos (Branco et al., 2012)

> Em Portugal, a divisão geográfica dos dialetos [13] distingue os dialetos do Centro-Sul, os dialetos do Norte e os dialetos das ilhas atlânticas. Os dialetos do Norte podem ser identificados pela ausência da distinção fonológica entre /b/ e /v/, com prevalência do /b/, pela preservação de antigos ditongos, e pela existência de fricativas ápicoalveolares. As diferenças entre estes dialetos encontram-se sobretudo ao nível da fonética e fonologia e ao nível lexical, sendo todos eles mutuamente compreensíveis de forma imediata (possivelmente com a exceção de alguns dialetos das ilhas).

Desafios para processamento

- Dialetos (Branco et al., 2012)

> A situação das variedades africanas do português é variada: enquanto em Angola e Moçambique o número de falantes de português tem vindo a aumentar desde a independência destes países, noutros casos, como São Tomé e Princípe ou Cabo Verde, em muitas circunstâncias utiliza-se amplamente o crioulo e o português é adquirido como língua segunda.

> Quanto ao Brasil, dada a dimensão geográfica deste país, não é viável apresentar aqui as suas variedades linguísticas. Por razões geográficas, políticas e sociais, não é possível falar de uma variedade padrão do português do Brasil. Os especialistas tendem a mencionar "normas urbanas cultas".

Desafios para processamento

- Zuchini (2011)
  - Elevado número de vocábulos existentes
  - Elevado número de sinônimos entre vocábulos
  - Elevado número de flexões verbais
  - Diversas possibilidades de construção sintática
  - Elevado número de flexões em gênero, número e grau
  - Grande número de exceções para praticamente todas as regras

PLN para o português

- Perspectiva histórica
  - Fortalecimento a partir da década de 90
  - Alinhamento aos tópicos de trabalhos para outras línguas (inglês, principalmente), mesmo que com algum atraso
  - Valorização do tratamento "individualizado" do português

Tarefa da semana

- Leitura
  - Finger (2021): Inteligência Artificial e os rumos do processamento do português brasileiro
    - No e-Disciplinas

Desafios para casa

- Iniciativa da USP
  - *ChatGPT: Potencial, Limites e Implicações para a Universidade*
    - https://www.youtube.com/live/bY2aTBeCyJU?feature=share
- Iniciativa da SBC
  - *Assistentes Virtuais Inteligentes: ChatGPT em Foco*
    - https://www.youtube.com/watch?v=tqGQfFb0OhA