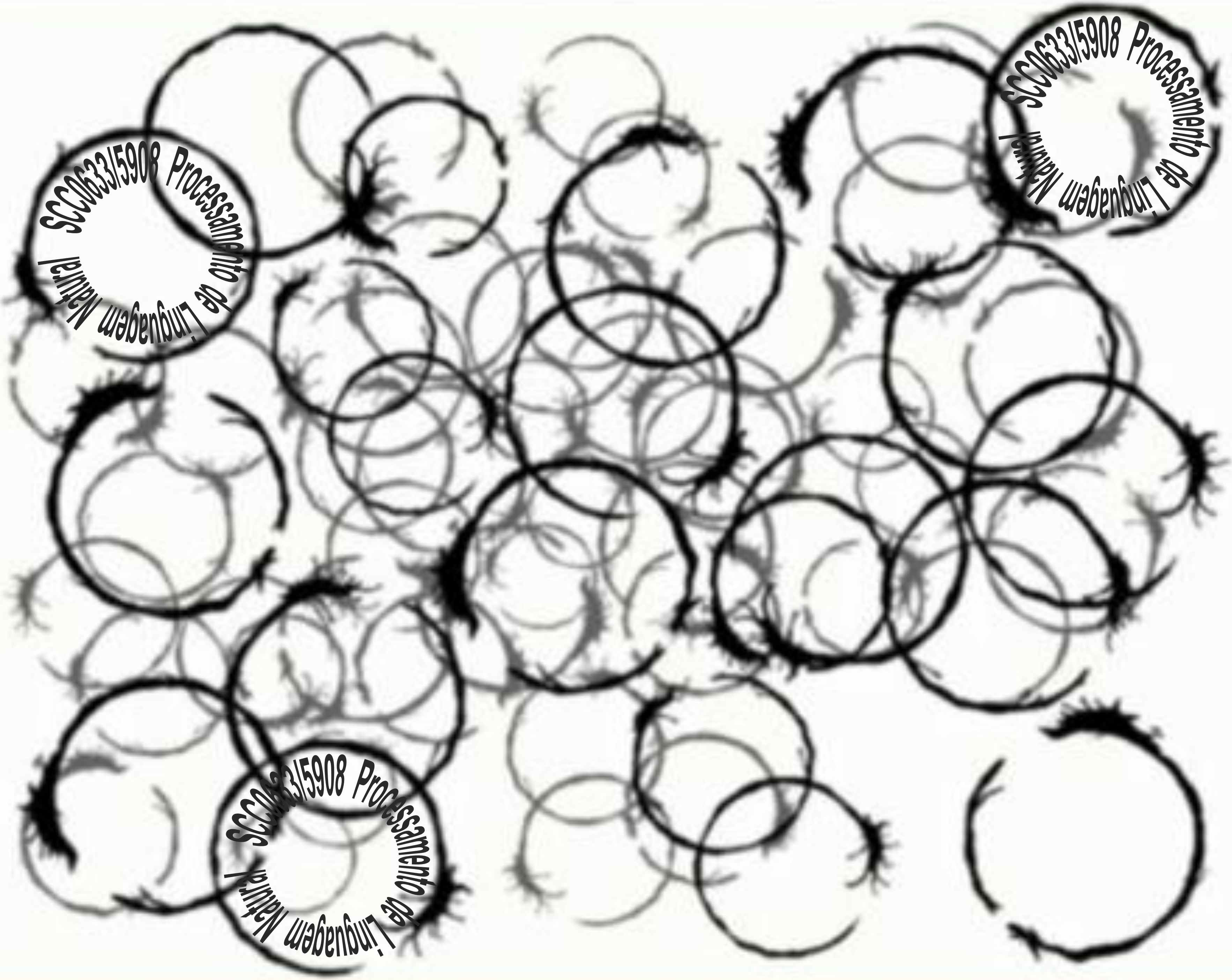
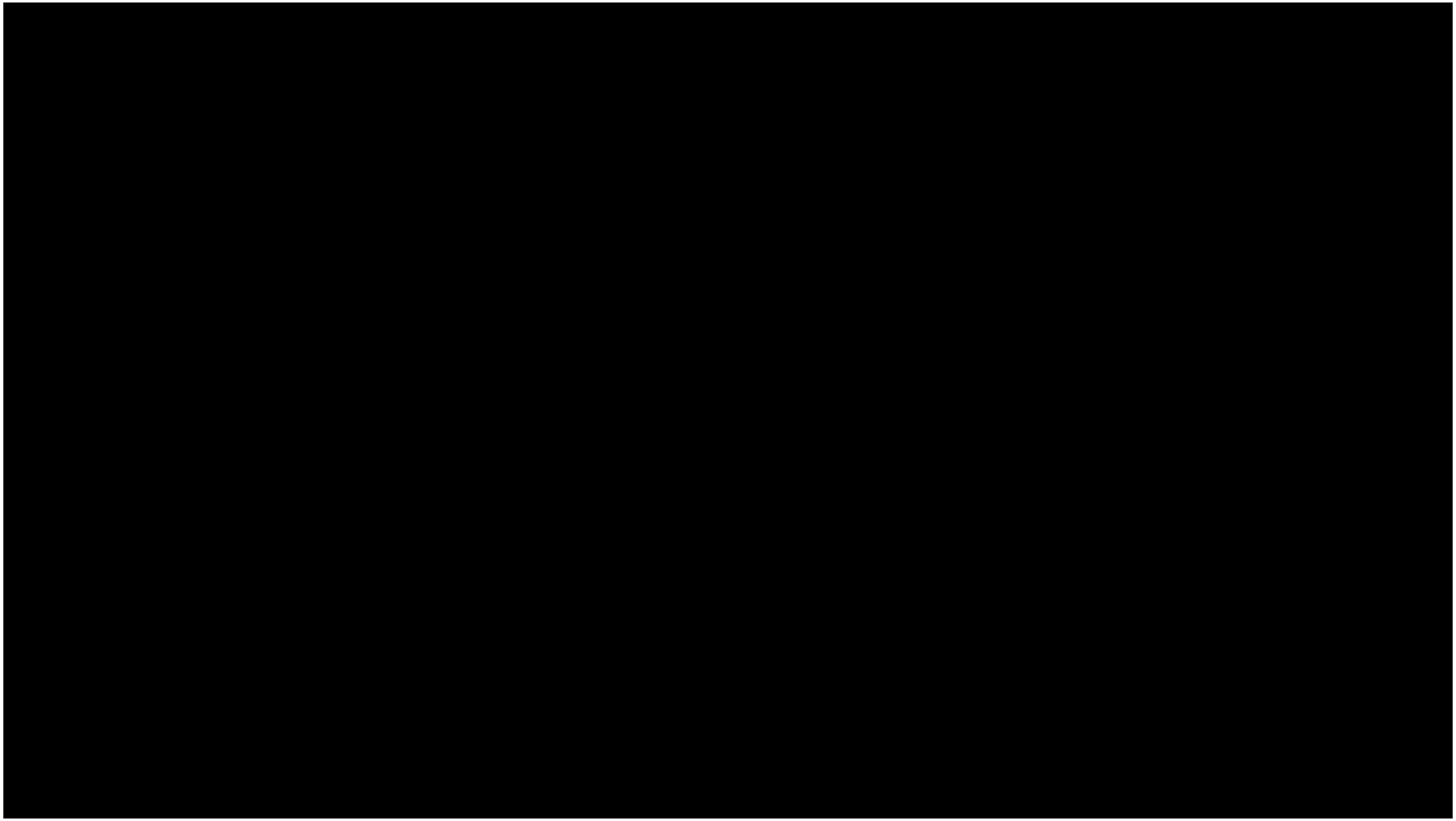

SCC0633/5908 Processamento de Linguagem Natural
SCC0633/5908 Processamento de Linguagem Natural
SCC0633/5908 Processamento de Linguagem Natural

# Representação nas telas

- A Chegada (2016)

# Primeiro contato



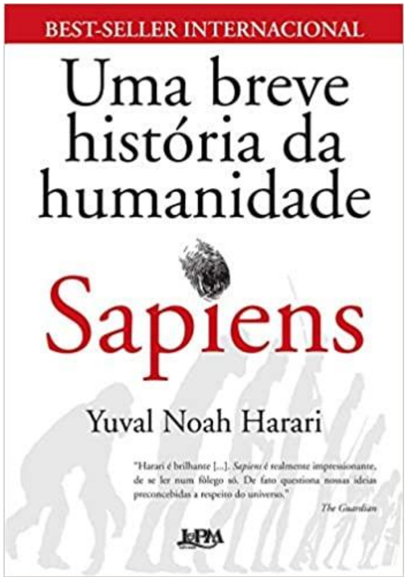
# A língua alienígena de "A Chegada"

# Linguagem, pensamento e humanidade

- Diversas questões complexas
  - *Nos tornamos “humanos” porque temos linguagem sofisticada?*
    - *Cognição sofisticada depende de linguagem?*
  - *Nossa capacidade de linguagem é “geneticamente” codificada?*
  - *A percepção e a atuação no mundo dependem da linguagem?*
    - *Em que língua nós pensamos?*

# Questão

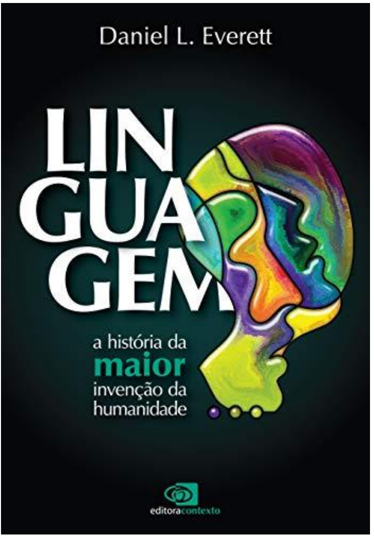
- Por que é tão difícil responder essas questões?
  - Dilema da História: como descrever e explicar algo de natureza abstrata, cujos rastros praticamente desapareceram?
    - Yuval Harari: revolução cognitiva há ~70.000 anos
      - “O Homo Sapiens conquistou o mundo, acima de tudo, graças à sua linguagem única... O surgimento de novas formas de pensar e se comunicar, entre 70 mil anos atrás a 30 mil anos atrás, constituiu a Revolução Cognitiva. O que a causou? Não sabemos ao certo. A teoria mais aceita afirma que mutações genéticas acidentais mudaram as conexões internas do cérebro dos sapiens, possibilitando que pensassem de uma maneira sem precedentes e se comunicassem usando um tipo de linguagem totalmente novo. Poderíamos chamá-las de mutações da árvore do conhecimento.”



# Questão

- Por que é tão difícil responder essas questões?
  - Dilema da Linguística e da Biologia: como e por que falamos?
    - Daniel Everett: se nossa espécie desaparecesse hoje, alguém conseguiria afirmar/evidenciar que nós tivemos linguagem sofisticada?
      - “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.”

(Fonte: entrevista à Revista Veja, em 2012)



# Questão

- Por que é tão difícil responder essas questões?
  - Dilema da IA: se nosso “cérebro” (mente?) fosse tão simples de explicar, nós seríamos tão “simples” que não conseguiríamos fazê-lo

# Introdução ao Processamento de Línguas Naturais

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

Thiago A. S. Pardo

# Na última aula

- Breve história do PLN
- Os desafios de PLN e sua transdisciplinaridade
- Aplicações avançadas e do dia a dia
- “Conversando” com uma máquina
  - Da Eliza ao GPT
  - A inspiração do HAL 9000: como fazer tudo que ele faz?

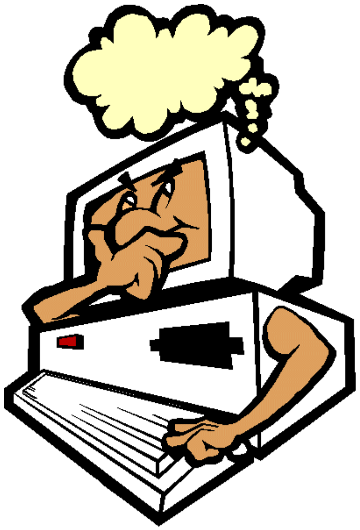
# LX-Suite

- Abrir LX-Suite, na página do grupo LX-Center (http://lxcenter.di.fc.ul.pt/) e testar as ferramentas abaixo
  - Syllabifier
  - Verbal Lemmatizer
  - POS Tagger
  - Constituency Parser
  - Dependency Parser
  - Named Entity Recognizer
  - Semantic Role Labeller
  - Semantic Similarity

- Questões
  - O que cada ferramenta faz? Alguma cometeu erro?

# PLN

- De que um computador necessita para ser capaz de entender uma fala humana e interagir adequadamente?
  - Como nós, humanos, fazemos isso?



Quem é Lula?

Que preguiça!

Está calor aqui.

Linguista: O que Chomsky disse?

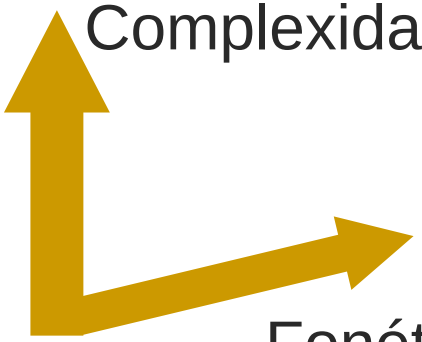
Informata: O que Chomsky disse?

Quem é Lula? Sei que não é o molusco.

# PLN

- Vários níveis de conhecimento
  - Tradicionalmente distinguidos em PLN, apesar dos limites entre eles serem nebulosos na maioria dos casos

Pragmática / Discurso
Semântica
Sintaxe
Morfologia



# PLN

- Fonética: estuda como os humanos produzem, transmitem e recebem sons, independente de língua; sistema físico

- Fonologia: estudo dos sons em uma língua específica, como os sons são construídos

  - Fones, fonemas, local (bilabial, palatal, etc.) e modo de articulação (pausa, nasal, fricativo, etc.), etc.

Pragmática / Discurso
Semântica
Sintaxe
Morfologia
**Fonética / Fonologia**

# PLN

International Phonetic Alphabet

| IPA | ASCII | examples |
|---|---|---|
| ʌ | ^ | cup, luck |
| ɑ: | a: | arm, father |
| æ | @ | cat, black |
| ə | .. | away, cinema |
| e | e | met, bed |
| ɜ:ʳ | e:(r) | turn, learn |
| ɪ | i | hit, sitting |
| i: | i: | see, heat |
| ɒ | o | hot, rock |
| ɔ: | o: | call, four |
| ʊ | u | put, could |
| u: | u: | blue, food |
| aɪ | ai | five, eye |
| aʊ | au | now, out |
| oʊ/əʊ | Ou | go, home |
| eəʳ | e..(r) | where, air |
| eɪ | ei | say, eight |
| ɪəʳ | i..(r) | near, here |
| oɪ | oi | boy, join |
| ʊəʳ | u..(r) | pure, tourist |

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.

Transcrição fonética

Pragmática / Discurso
Semântica
Sintaxe
Morfologia
**Fonética / Fonologia**

# PLN

- Palavra: construção, componentes de formação
  - Morfema, raiz, afixo (prefixo, sufixo, etc.), vogal temática, desinência

Pragmática / Discurso
Semântica
Sintaxe
**Morfologia**
Fonética / Fonologia

# PLN

- Interação entre morfologia e sintaxe: classes gramaticais ou etiquetas morfossintáticas

  - Substantivo/nome, verbo, adjetivo, advérbio, pronome, preposição, conjunção, interjeição, etc.

Pragmática / Discurso
Semântica
**Sintaxe**
**Morfologia**
Fonética / Fonologia

# PLN

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.

Pragmática / Discurso
Semântica
**Sintaxe**
**Morfologia**
Fonética / Fonologia

**Ele** [ele] **PERS** M 3S NOM
**queria** [querer] <fmc> **V** IMPF 3S IND VFIN
**jogar** [jogar] **V** INF
**tênis** [tênis] **N** M S/P
**com** [com] **PRP**
**Janete** [Janete] **PROP** M/F S
**,**
**mas** "mas" <co-vfin> <co-fmc> **KC**
**também** [também] **ADV**
**queria** [querer] <fmc> **V** IMPF 3S IND VFIN
**jantar** [jantar] **V** INF
**com** [com] **PRP**
**Suzana** [Suzana] **PROP** F S
**.**
**Sua** [seu] <poss 3S> **DET** F S
**indecisão** [indecisão] **N** F S
**o** [ele] **PERS** M 3S ACC
**deixou** [deixar] <fmc> **V** PS 3S IND VFIN
**louco** [louco] **ADJ** M S
**.**

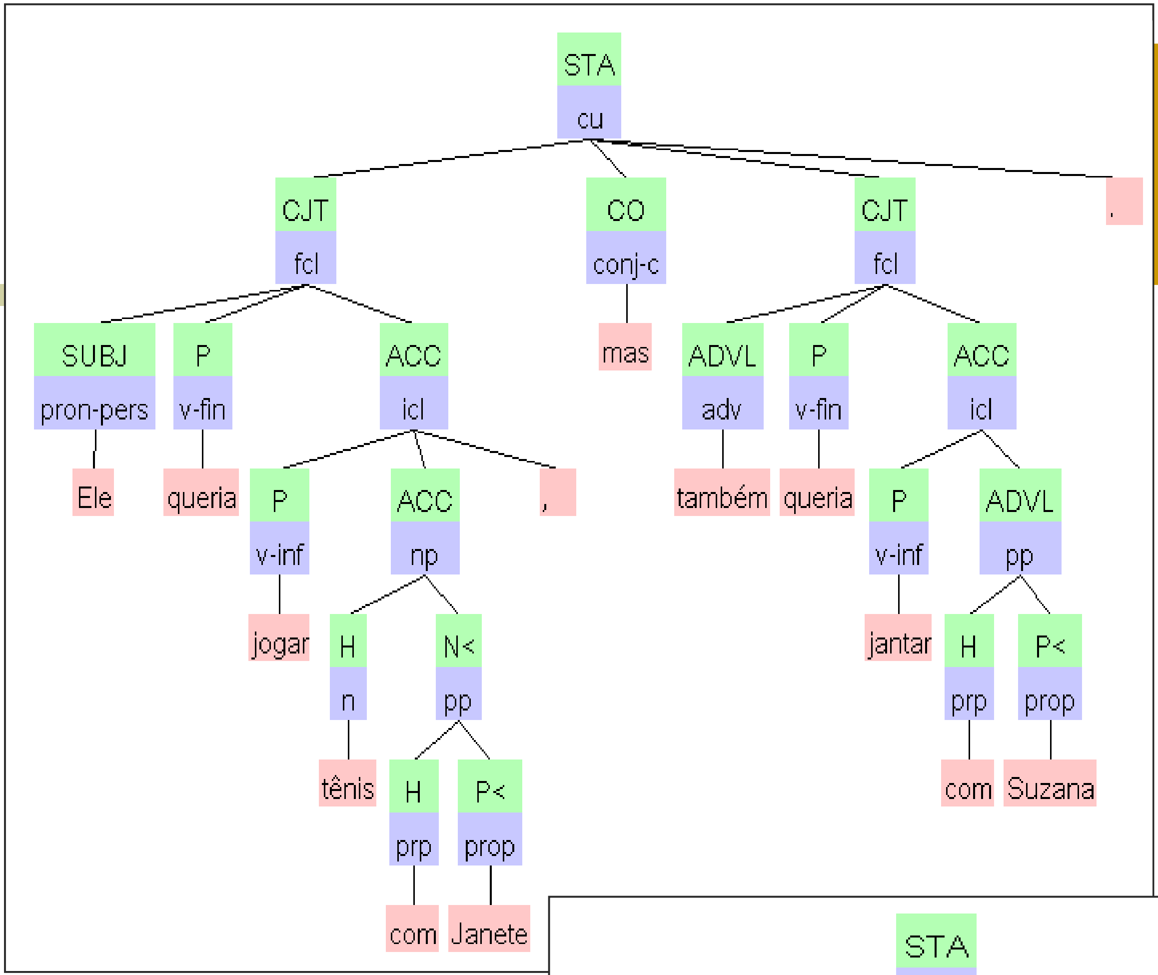
# PLN

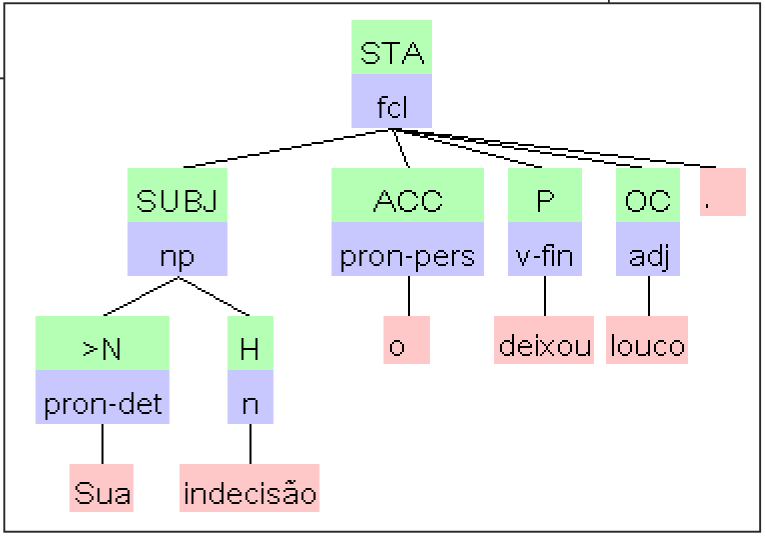
- Como as sentenças são formadas, como as palavras podem se combinar
  - Função: sujeito, predicado, objetos, predicativos, etc.
  - Estruturação/constituição: sintagma nominal, sintagma verbal, etc.

Pragmática / Discurso
Semântica
**Sintaxe**
Morfologia
Fonética / Fonologia

# PLN

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.




Pragmática / Discurso
Semântica
**Sintaxe**
Morfologia
Fonética / Fonologia

# PLN

- Significado
  - Palavras, expressões, orações, sentenças, textos
  - Lexical, composicional, textual

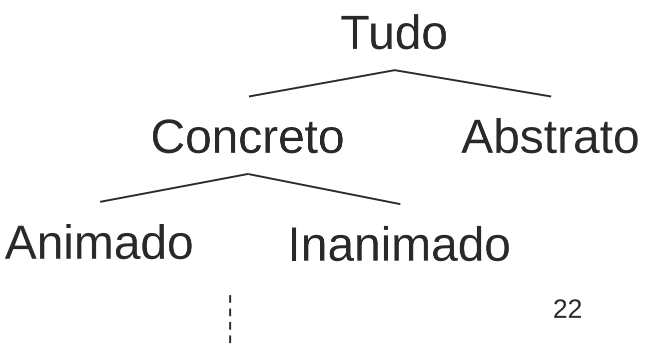
Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Traços semânticos, classificações ontológicas

| | Mesa | Cavalo | Garota | Mulher |
|---|---|---|---|---|
| Animado | - | + | + | + |
| Humano | - | - | + | + |
| Fêmea | - | - | + | + |
| Adulto | - | + | - | + |

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia



# PLN

- Papéis semânticos/temáticos
  - Agente, tema, instrumento, experienciador, fonte, etc.
    - [O menino]$_{\text{AGENTE}}$ chutou [a bola]$_{\text{TEMA}}$

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Classes/categorias/tipos semânticos
  - Humano, local, data, organização, etc.
    - O [menino]$_{\text{HUMANO}}$ chutou a bola
  - Entidades nomeadas

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Relações “lexicais”
  - Sinonímia, antonímia, hiperonímia/hiponímia, meronímia/holonímia, etc.

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Diversos fenômenos
  - Metáforas, expressões idiomáticas, polissemia
    - Qual a diferença entre polissemia e homonímia?
      - Banco (assento vs. instituição financeira) é polissêmico, mas manga (camisa vs. fruta) não é

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.

“Ele”, “Janete” e “Suzana” = humanos.

Jogar tênis = praticar o esporte tênis ≠ arremessar o calçado.

...

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.

queria(exper(ele),objetivo(jogar(tênis),comutativo(Janete)))...

Pragmática / Discurso
**Semântica**
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Discurso
  - Aquilo que está além da sentença
    - Relacionamento proposicional, correferência e expressões referenciais, intenções, tópicos/subtópicos, componentes retóricos, etc.

**Pragmática / Discurso**
Semântica
Sintaxe
Morfologia
Fonética / Fonologia

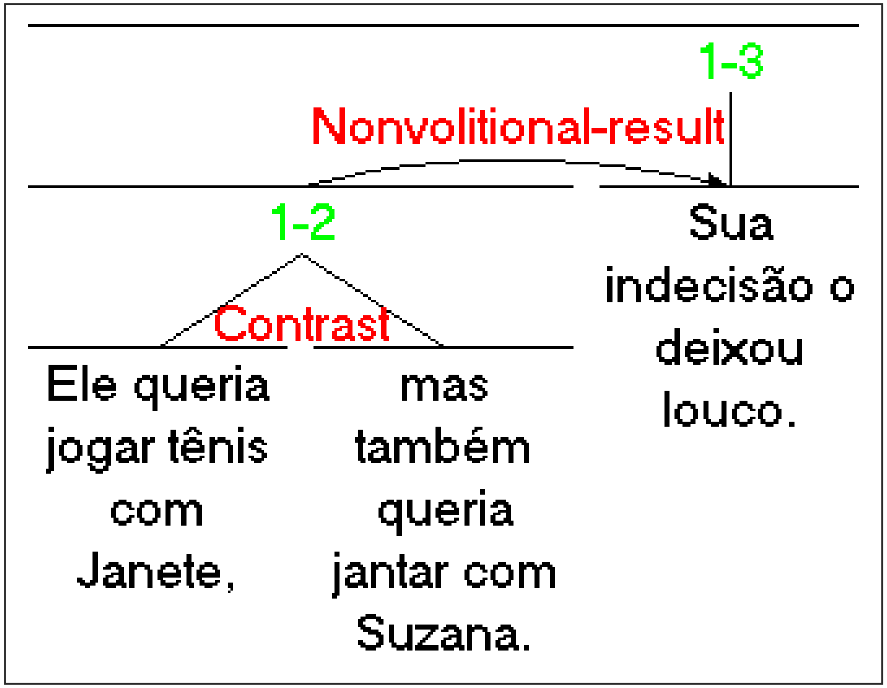
# PLN

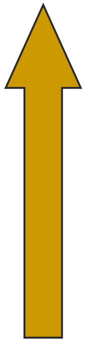
Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.

**Pragmática / Discurso**
Semântica
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.



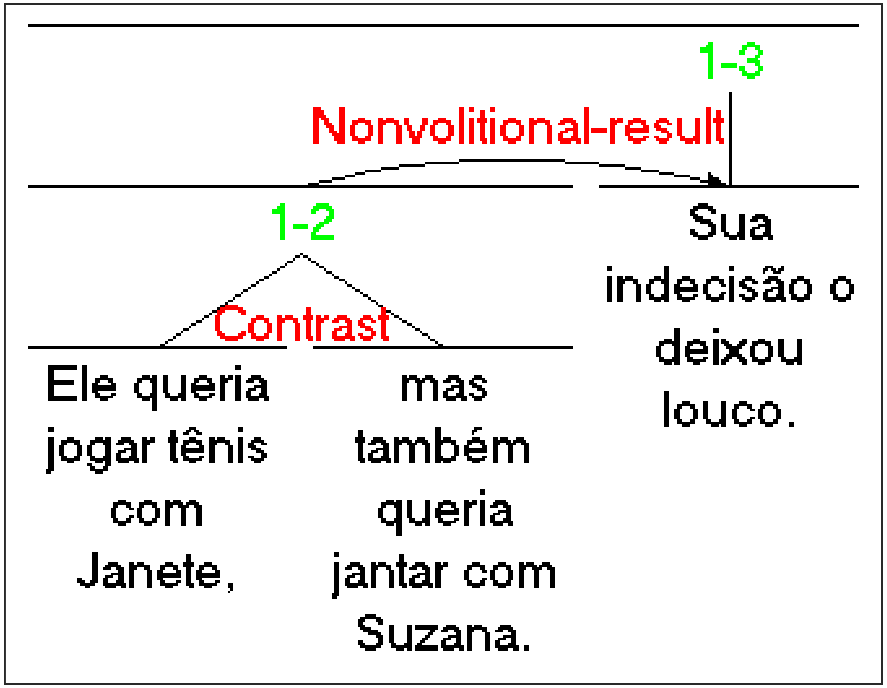



**Pragmática / Discurso**
Semântica
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

Ele queria jogar tênis com Janete, mas também queria jantar com Suzana. Sua indecisão o deixou louco.

**(*Intend* E (*Believe* L "o desejo de fazer duas coisas incompatíveis o deixou louco"))**




**Pragmática / Discurso**
Semântica
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Pragmática
  - Língua em uso, interação, contexto
    - Fatores como força, educação, hierarquia, crença, cooperação, atitude
    - Estilos de escrita e de fala
    - Suposições sobre produtor e receptor, nível de conhecimento, interesses
      - Modelagem do usuário

**Pragmática / Discurso**
Semântica
Sintaxe
Morfologia
Fonética / Fonologia

# PLN

- Considerações para uso por um computador
  - Os níveis de conhecimento precisam ser representados (formalizados) e manipulados automaticamente
  - Interação entre os níveis
    - Morfologia e sintaxe
    - Sintaxe e semântica
    - Semântica e discurso

# PLN

- Considerações para uso por um computador
  - Os níveis de conhecimento precisam ser representados (formalizados) e manipulados automaticamente
  - Interação entre níveis mais distantes
    - Morfologia e semântica (goleiro e porteiro vs. padeiro)
    - Morfologia e pragmática (são carlense vs. são carlino, laranjada e limonada vs. cajuada)
    - Sintaxe e discurso (subordinadas)

# PLN e humanos

- Processamento sequencial vs. paralelo
    - Arquiteturas em *pipeline* vs. integradas

# PLN e humanos

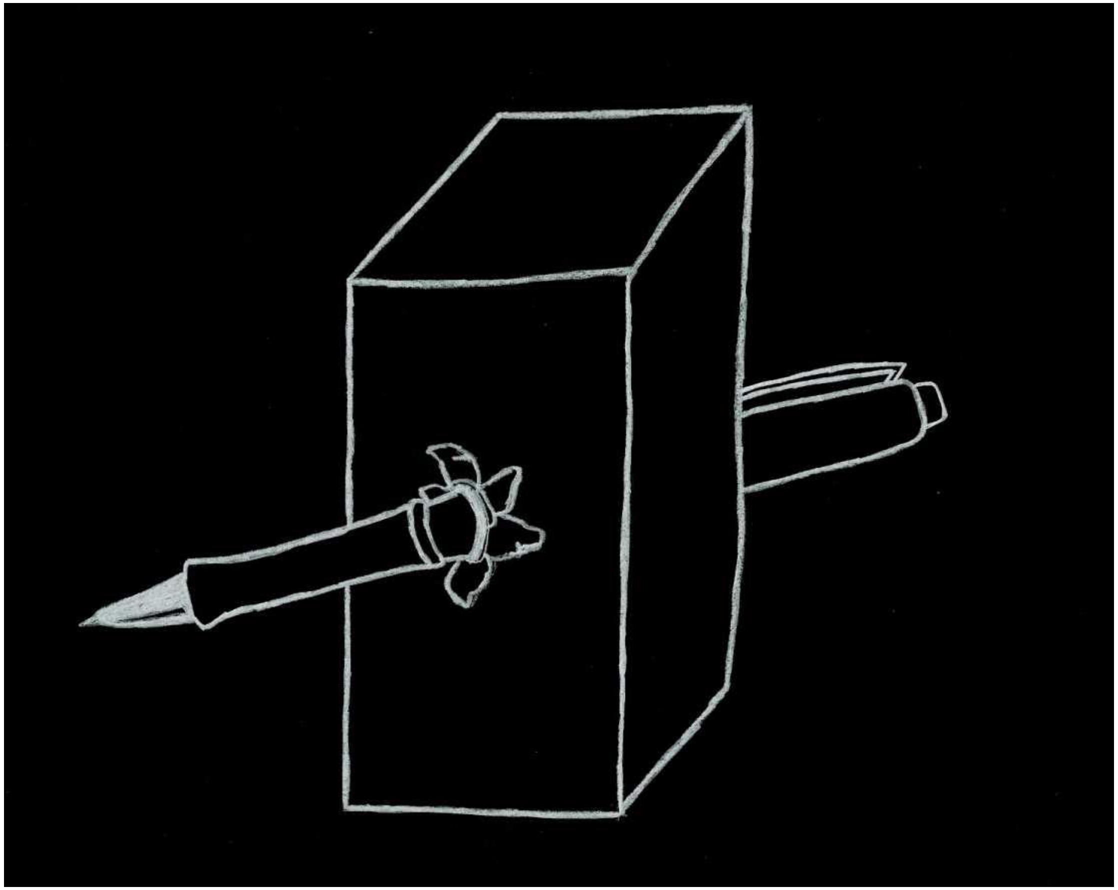
- Humanos lidam naturalmente com
  - Ambiguidade
  - Irregularidade
  - Vagueza
  - Variedade
  - Etc.

- ... máquinas (ainda) não!

# Exemplos de dificuldades

- O homem viu a mulher na montanha de binóculos
- Você sabe as horas?
- O coelho foi servido
- O homem foi servido
- A caneta está na caixa
- A caixa está na caneta

# Exemplos de dificuldades



# PLN

- Trabalho em PLN

Fase linguística ⇔ Fase representacional ⇔ Fase implementacional

# PLN

- Trabalho em PLN

| Fase linguística | Fase representacional | Fase implementacional |
| --- | --- | --- |
| Resumos de artigos de jornais | Formalização das regras para resumir | Desenvolvimento do sumarizador automático |
| Tradução espanhol-português | Regras de tradução, dicionários bilíngues | Desenvolvimento do tradutor automático |

# PLN

- Trabalho em PLN

| Fase linguística | ⟺ | Fase representacional | ⟺ | Fase implementacional |
|---|---|---|---|---|
| Teorias e dados linguísticos | | Formalismos/esquemas de representação | | Métodos e modelos computacionais |

- Aspectos da língua que são possíveis capturar e automatizar
  - Maioria das teorias linguísticas são sofisticadas demais para o PLN... alguns recursos também (exemplo?)

# PLN: Computação e Linguística

- Esforço inicial para interação entre diferentes áreas
  - Como na maioria das áreas multidisciplinares
    - Informata: sujeito, predicado, relações semânticas, vozes do texto, Saussure?
    - Linguista: scripts, usabilidade, autômato, Turing?

# PLN & IA

- Classificações... nem sempre triviais

| **Critérios** | **Paradigmas** |
| --- | --- |
| Uso de conhecimento linguístico | Superficial, profundo e híbrido |
| Representação do conhecimento | Simbólico, 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

# 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...*

SUPER INTERESSANTE

# Empirismo

- 1920-1960: empirismo
  - Mente não vem com princípios e procedimentos pré-determinados
  - Mas vem com <u>operações gerais</u> 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 <u>quantidade muito grande</u> 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 (*geratívismo*)
    - Descrição do <u>módulo linguístico da mente humana</u>, sendo <u>córpus somente evidência indireta</u>
      - “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

- <u>Linguística a la Chomsky</u>
  - Princípios categóricos
    - Sentenças satisfazem ou não

- <u>Empirismo</u>
  - 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

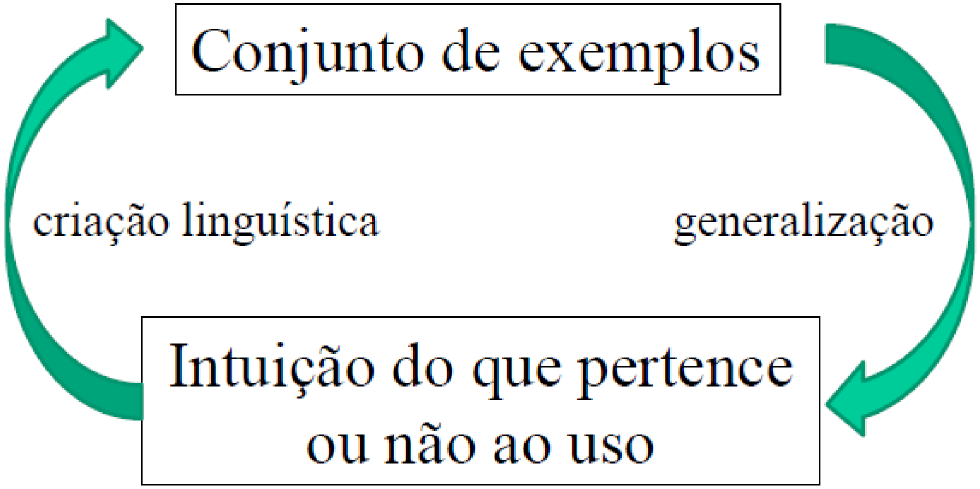
- Atenção aos “erros”

# 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”

# Abordagens: PLN

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



# 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

# PLN

- Classificação
  - Recursos
  - Ferramentas
  - Aplicações

# Recursos

- Córpus
  - Anotação: humana e/ou automática
    - XML, XCES, TEI, etc.
  - Paralelo, comparável, alinhado, etc.

- Dicionários monolíngues e bilíngues
  - *Machine readable* vs. *machine tractable*

- Léxicos
  - Vários paradigmas

# Ferramentas

- Segmentadores textuais: palavras (*tokenizador*), sentenças, parágrafos, tópicos
- Stemmers, lematizadores, nominalizadores
- Etiquetadores morfossintáticos (*taggers*)
- Analisadores sintáticos *shallow* (*chunkers*) e *deep* (*parsers*)
- Analisadores semânticos e discursivos
- Alinhadores textuais: lexicais, sentenciais, etc.
- Concordanceadores, *word counting, …*
- Classificadores de polaridade
- Etc.

# Aplicações

- Tradutores automáticos
- Revisores ortográficos e gramaticais
- Ferramentas de auxílio à escrita
- Sumarizadores automáticos
- Simplificadores textuais
- Minerador de opinião
- Etc.

# Recursos, ferramentas e aplicações

- Atenção
  - Classificação difusa, às vezes
  - Dependente do uso
    - Sumarizador como passo intermediário para recuperação da informação → ferramenta
    - Dicionário eletrônico para consulta → aplicação

# PLN e áreas correlatas

- Limites cada vez mais suaves entre PLN e outras áreas
  - Recuperação de informação
  - Banco de dados
  - Interação humano-computador
  - Mineração de textos
  - Linguística de córpus

# 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

# Tarefas

- Capítulo 1 do livro *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*
  - No e-Disciplinas

- Provinha 2 disponível à tarde