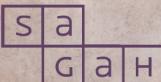


# PROCESSAMENTO DE LINGUAGEM NATURAL

Sofia Maria Amorim Falco Rodrigues



SOLUÇÕES  
EDUCACIONAIS  
INTEGRADAS

# Introdução ao processamento de linguagem natural

## Objetivos de aprendizagem

Ao final deste texto, você deve apresentar os seguintes aprendizados:

- Descrever o fluxo básico do processamento de linguagem natural.
- Reconhecer as áreas de conhecimento relacionadas ao processamento de linguagem natural.
- Comparar o papel do processamento de linguagem natural em diferentes problemas.

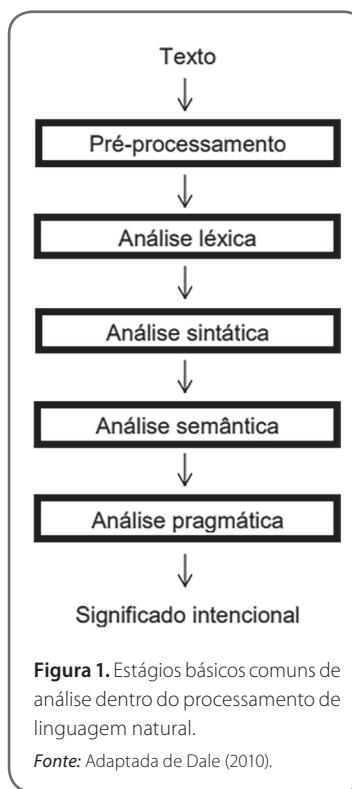
## Introdução

Você já deve ter ouvido falar sobre o processamento de linguagem natural (PLN) ou, do inglês, *natural language processing* (NPL), por exemplo, pela própria vertente da ciência da computação, em alguma aplicação específica ou em outro contexto interdisciplinar. Aqui, o principal objetivo consiste em apresentar e contextualizar o PLN como uma subárea da ciência da computação, que também integra áreas como a inteligência artificial, o aprendizado de máquina (ou, como mais conhecido em inglês, *machine learning*) e a própria linguística, em que é responsável pelo estudo da interpretação e geração automática da linguagem natural. Assim, pretende-se que você consiga compreender, em um panorama geral, em que contexto está inserida a disciplina dentro da ciência da computação, além da vasta aplicabilidade de toda a metodologia envolvida.

Neste capítulo, você aprenderá como se dá o PLN, além de tornar-se capaz de reconhecer as principais áreas de conhecimento relacionadas a esse processamento. Ao final, conseguirá realizar comparações para a compreensão do papel do PLN para diferentes problemas práticos.

## 1 Processamento de linguagem natural

Esse tipo de processamento segue um fluxo básico de análises e ações, conforme visto na Figura 1 (DALE, 2010). Assim, nesse momento você compreenderá inicialmente o que é a própria linguagem natural, analisando as ideias e a estrutura básica para a implementação, para, depois, entender cada uma das ações descritas no diagrama da Figura 1. Um texto ou o conjunto de textos, também denominado *corpus* ou *corpora*, corresponde à entrada do sistema de processamento de linguagem natural (PLN). Em seguida, tem-se o estágio de pré-processamento, que poderá envolver até mesmo análises do som e que consiste geralmente na segmentação de unidades lexicais e de sentenças, até que possam ser executadas cada uma das análises que trarão, ao final, o **significado intencional**, no qual ocorre o entendimento por parte da máquina.



Antes de compreender as motivações por trás dessa vertente, que passou a se configurar inclusive como uma área interdisciplinar de pesquisa, você deverá entender, de antemão, o que é a **linguagem natural humana**.

## Linguagem natural

Linguagem natural (ou, simplesmente, língua natural ou idioma) comprehende qualquer tipo de linguagem desenvolvida de forma natural pelo ser humano, e não premeditada, considerada o resultado da própria facilidade inata humana e basicamente subdivida entre língua falada e de sinais (MATTHEWS, 2003).

Contudo, fazendo um breve histórico, sabe-se que não há dados certos quanto ao surgimento da linguagem nos seres humanos ou até mesmo em seus ancestrais, ainda que se estime, a partir de evidências recentes, que a linguagem natural tenha sido elaborada ou evoluído no continente africano, antes mesmo da movimentação dos humanos em todo o planeta, há cerca de 50 mil anos, tornando-se importante lembrar que há, essencialmente, linguagem natural na comunicação entre indivíduos, o que inclui até mesmo os povos ancestrais (WADE, 2003).

Já do ponto de vista do processamento natural, pode-se pensar sobre as relações entre a linguagem e o cérebro humano além do próprio funcionamento da linguagem. Embora reitere-se que sabe muito pouco acerca das relações exatas entre a linguagem, a maneira como a percebemos e o próprio cérebro humano, alguns pesquisadores atribuem um crescimento já constatado do cérebro humano, comparando-o a tempos anteriores, como o próprio surgimento da linguagem (LORITZ, 2002). Com relação às características envolvidas na linguagem natural, como a própria fala, entende-se que esse processo se dá pela proferição de unidades, denominadas palavras ou, em menores escalas, letras, vogais e consoantes, embora essa divisão possa se tornar desafiadora em diversos contextos (LADEFOGED; MADDIESON, 1996). Além disso, como é esperado, a fala reflete diretamente na linguagem falada e entende-se também, por consequência, que a linguagem falada variará no tempo e no espaço e que informações relevantes sobre essa característica advêm de estudos da forma como as línguas faladas mudam, além do fato de que, independentemente da velocidade de fala de um idioma, a transmissão de informação é estimada em uma velocidade de 39 bits por segundo (MATAVIC, 2019).

Ainda no contexto do funcionamento da linguagem natural, pode-se destacar as linguagens de sinais, outra de suas variedades, com complexidades também relacionadas às estruturas gramaticais, além da própria capacidade de expressão atrelada, caso em que está diretamente associada à comunicação (SACKS, 2009).

Todavia, conforme o contexto já apresentado e o seu próprio contato com uma linguagem, seja um idioma do país de origem com o qual você se comunica, seja uma língua de sinais, há o estabelecimento de uma série de regras para a construção tanto da própria forma escrita quanto da falada, aspectos que serão explorados a seguir para que você entenda como isso se dá de forma básica e geral, além de certos exemplos e aplicações específicas. Ademais, você verá as principais motivações envolvidas no PLN, para compreender o surgimento desta vertente interdisciplinar.

## Motivações

O estudo do PLN poderá avaliar, por exemplo, problemas relacionados à geração e à compreensão automática de línguas naturais humanas, o que mostra a infinidade de sua aplicabilidade, visto as próprias premissas de estabelecimento de linguagem natural: a comunicação falada e de sinais. Além disso, um dos principais desafios, de fato, consiste na compreensão da linguagem natural, relacionada ao “fazer com que o computador comprehenda”, tornando-o capaz de extrair “sentido” da linguagem natural humana e, também, à geração de linguagem natural, outra ação dentro do PLN.

Um dos principais objetivos a destacar no PLN é simplesmente fazer com que os computadores executem tarefas úteis que envolverão a própria linguagem humana, como a habilitação da comunicação entre o humano e a máquina, atuação na melhoria da comunicação entre os próprios humanos ou até mesmo na execução de processos úteis que se utilizem de textos ou falas (JURAFSKY; MARTIN, 2008).

Assim, compreendendo as principais razões para o desenvolvimento do PLN, você deverá agora entender o histórico no qual este ocorreu, analisando toda a linha cronológica com os principais fatos e pesquisadores envolvidos.

## Histórico do processamento de linguagem natural

Sem dúvidas, o primeiro marco a ser citado para o PLN se deu em 1950, pelo matemático, cientista da computação e pesquisador Alan Turing, quando da publicação na ocasião do trabalho *Computing machinery and intelligence* (traduzido para o Brasil como “Computadores e Inteligência”), pioneiro não somente no ramo de inteligência artificial, mas também em toda a ciência da computação, tornando-se o responsável especialmente pela resposta da questão: “as máquinas podem pensar?”. Além disso, com esse trabalho Turing formulou o que hoje se conhece como **teste de Turing**, basicamente uma maneira de avaliar a capacidade de uma máquina em exibir um comportamento considerado inteligente, pela semelhança com o comportamento humano, essencialmente. Ademais, a proposta é converter a pergunta apresentada em outra menos ambígua: “Há como imaginar um computador digital que faria bem ‘o jogo da imitação’?”.



### Saiba mais

O trabalho de Alan Turing pode ser visto em *Parsing the Turing Test* (TURING, 2009). Ainda, é possível visualizar várias outras informações sobre esse importante pesquisador da área em buscas simples na internet, contextualizando com o cenário atual dentro da inteligência artificial, *machine learning* e o próprio PLN.

Sequencialmente a Turing, em 1954, destacou-se a **experiência de Georgetown** (HUTCHINS, 2004), na qual se fez a tradução automática de mais de 60 frases em russo para o inglês, fato que marcou tanto o PLN quanto a própria ciência da computação. Por sua vez, surgiram na prática sistemas estatísticos de tradução no final da década de 1980, que contradiziam a expectativa de que em até 5 anos a tradução automática estivesse resolvida, além do fato de que experimentos e pesquisas anteriores ao surgimento desses sistemas tinham se desenvolvido em meio a um cenário desmotivador, pelos resultados insuficientes inesperados e pela redução de investimentos.

Adicionalmente, na década de 1960, é possível citar sistemas bem-sucedidos de PLN denominados **SHRDLU**, programas desenvolvidos para a interação humana com termos em inglês, com base no arranjo das teclas de letras em uma máquina Linotype (ETAOIN SHRDLU), organizados em ordem decrescente de frequência de uso na língua inglesa (WINOGRAD, 1972). Esse sistema trabalhava com os denominados **mundos dos blocos**, um dos mais famosos domínios da inteligência artificial, em linguagem com vocabulário restrito e com a simulação **ELIZA** (WEIZENBAUM, 1966), o primeiro *software* para simulação de diálogos, empregando pouca informação sobre o pensamento e/ou a emoção humana para a criação do diálogo, também tido como um tipo de *chatterbot* (em português, “programas que simulam humanos na conversação interpessoal”).

Já por volta dos anos 1970 e 1980, diversos programadores iniciaram a escrita do que se chamou “ontologias conceituais”, responsáveis pela estruturação das informações do mundo real em dados que eram compreensíveis para o computador, cujos alguns exemplos de resultados são organizados em ordem crescente no tempo: MARGIE (RIESBECK *et al.*, 1975), TaleSpin (MEEHAN, 1976), QUALM (LEHNERT, 1977), SAM (CULLINGFORD, 1978), PAM (SCHANK; WILENSKY, 1978), Politics (CARBONELL, 1979) e Plot Units (LEHNERT, 1981).



### Saiba mais

Basicamente, uma ontologia, com relação à própria ciência da computação, refere-se a um conjunto de especificações formais e explícitas de uma conceitualização, que, no PLN, pode ser observada em exemplos específicos, como a facilitação da tradução de textos médicos, quando se baseia em textos especializados e até mesmo dicionários médicos.

Ademais, no período é possível citar diversos *chatterbots*, como o PARRY, de 1971 (COLBY; WEBER; HILF, 1971), o Racter (CHAMBERLAIN, 1984) e o Jabberwacky, criado pelo programador britânico Rollo Carpenter, com os primeiros resultados datados por volta de 1981.



## Saiba mais

Para mais detalhes sobre o *chatterbot* Jabberwacky existe um site oficial, com histórico e demais informações. Acesse no *link* a seguir.

<https://qrgo.page.link/weFCV>

Até meados da década de 1980 a maior parte dos algoritmos era realizada com base em conjuntos de regras manuscritas, contudo, ao final da década, houve o que foi considerada a **revolução no processamento de linguagem natural**, sobretudo pela introdução de algoritmos de aprendizagem automática (a aprendizagem de máquina, ou, como mais conhecida, *machine learning*), o que foi possível pelo aumento constante da capacidade dos computadores, dentro da expectativa da lei de Moore, e, também, por modificações específicas dentro da linguística. Além disso, processos como a marcação de partes da fala (conhecida como *part-of-speech tagging*) introduziram na época o uso de **modelos ocultos de Markov**, voltando a construção dos algoritmos ao uso de **modelos estatísticos**, que utilizam, por exemplo, pesos diferenciados aos componentes de dados de entrada, nos quais se enquadram os **modelos de linguagem de cachê**, muito usados em sistemas destinados ao reconhecimento de fala.

Do ponto de vista dos próprios modelos e algoritmos, observa-se que boa parte das implementações de tarefas dentro do PLN envolvia o processo de codificação direta de muitos conjuntos de regras. Analisando ainda de maneira geral, entende-se que boa parte dos resultados vistos até os dias atuais no campo do PLN está relacionada à tradução automática, como os trabalhos desenvolvidos pela International Business Machines Corporation (IBM) e os sistemas com modelos estatísticos aplicados ao Parlamento do Canadá e da União Europeia. Em contrapartida, atualmente observa-se que os algoritmos mais modernos se baseiam em processos de aprendizagem mecânica, principalmente na aprendizagem de máquinas estatísticas dentro da própria aprendizagem automática, sobretudo em algoritmos de árvores de decisão (HUTCHINS, 2005). Nos últimos 5 anos, estima-se ainda uma nova tendência no processamento, em razão do destaque do desenvolvimento de algoritmos de aprendizagem profunda (mais conhecidos como *deep learning*).



## Saiba mais

Caso o leitor deseje saber mais detalhes sobre os modelos mencionados, desenvolvidos pela IBM, pode obter mais informações sobre o Projeto Candide, um de seus pioneiros, em Berger *et al.* (1994).

Assim, analisando o contexto histórico atual, com o cenário de inserção tecnológica estabelecido principalmente com o uso cada vez mais frequente de computadores e o surgimento do *smartphone*, o PLN assume um papel cada vez mais protagonista, em virtude, de maneira geral, do grande fluxo de dados. Além disso, prevê-se e sabe-se que a quantidade de informações disponíveis no meio digital tende ainda a se manter em crescimento exponencial, o que traz à tona a necessidade do PLN para facilitar, por exemplo, a própria absorção de informações digitais pelo ser humano. Para compreender melhor alguns exemplos de aplicações do PLN, considere o exemplo a seguir.



## Exemplo

### Considere o contexto estabelecido pela rotina apresentada a seguir.

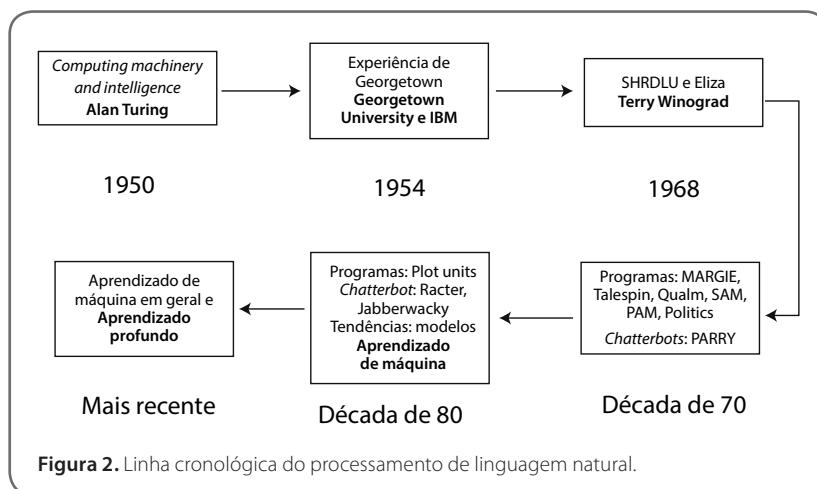
Assim, às 6h da manhã, o relógio desperta para a pessoa ir trabalhar; a esse ponto do dia, um sensor na cama detecta ativando a cafeteira e ligando o chuveiro. Em seguida, enquanto toma seu café, essa pessoa pergunta para o próprio *smartphone* qual será a previsão do tempo para o dia de hoje e solicita a ele um resumo das principais notícias até o horário. Depois, já ao sair de casa, também com o auxílio do *smartphone* essa pessoa acessa um mapa, ligando o GPS (do inglês, *global positioning system* ou “satélite de posicionamento global”) que a guiará; ao chegar ao trabalho, precisa fazer uma busca no Google; depois, tem em suas mãos um documento escaneado por alguém; e, durante sua pausa, ao marcar uma consulta médica, uma atendente automática o guiará por meio de opções até a marcação.

No final do dia, ao chegar em casa, essa pessoa nota que recebeu uma multa em seu nome, pois um radar detectou que ela avançou o sinal vermelho há 2 meses. Durante seu descanso, acessa o YouTube, porém o vídeo que ela quer ver está inaudível, levando-a a ligar a ferramenta de legenda automática para ajudá-la a entender.

Dessa forma, note que essa rotina descrita poderá ser a de boa parte das pessoas nos dias de hoje, com diversas ações e tarefas que contam com o uso da tecnologia. Entretanto, é necessário entender nesse ponto: o que grande parte dos acontecimentos descritos tem em comum? A resposta é simples: todos utilizam técnicas de PLN, seja em ações para auxiliar o usuário, seja na comunicação entre diferentes tecnologias.

Logo, como já esperado, hoje o PLN está presente em diversas situações cotidianas, por exemplo nas duas das principais redes sociais em todo o mundo: o Instagram e o Facebook. Dessa forma, por meio de técnicas de aprendizado de máquina, existem algoritmos nessas redes para encontrar padrões em grandes conjuntos de dados, por exemplo, capazes de reconhecer a linguagem natural para a análise de sentimentos, em que esses algoritmos poderão procurar por padrões em postagens, até mesmo para compreender como os indivíduos se sentem em relação a empresas e/ou a determinados produtos em específico.

A Figura 2 resume a linha cronológica do PLN com alguns dos fatos, sistemas e algoritmos mais marcantes vistos neste tópico.



**Figura 2.** Linha cronológica do processamento de linguagem natural.

A seguir, você verá uma introdução ao PLN e suas principais ações relacionadas, com o fluxo básico de ideias de um algoritmo.

## Informações gerais sobre o processamento de linguagem natural

A partir de agora, você observará a ideia básica por trás do processamento de linguagem natural, a fim de compreender também as principais ações proporcionadas com algoritmos da área.

Assim, do ponto de vista da própria linguagem, como já foi possível compreender, a linguagem natural é a principal e mais comum ferramenta de comunicação da relação humana. E o computador também dispõe de mecanismos análogos de comunicação, representados pelas linguagens formais para o processamento das informações, como Java, Python, etc. Entretanto, na linguagem natural, tanto na forma escrita quanto na falada, há uma infinidade de regras complexas e que resultam em diversos casos de ambiguidade, sobretudo quando consideramos os diversos idiomas existentes. Assim, em geral um dos principais desafios no PLN consiste no desenvolvimento de sistemas capazes de entender a linguagem natural dentro de todos os parâmetros comuns envolvidos na construção da linguagem formal como conhecida, independentemente do idioma e de acordo com as especificidades de cada idioma e das aplicações, além de outros desafios e mais detalhes mencionados adiante.

Para entender o fluxo básico do PLN, considere os dois grandes exemplos apresentados a seguir: o **agente de conversação (sistema de diálogo)** e a **máquina de tradução**. O primeiro é um exemplo de tarefa bastante útil e de grande aplicabilidade dentro do PLN: um dos mais conhecidos e citados de agente de conversação é o próprio computador HAL 9000, do filme “2001: uma odisseia no espaço”. Na ocasião, demonstrou-se que o HAL é um agente artificial capaz de realizar comportamentos linguísticos avançados, como falar e entender o inglês e, conforme demonstrado no enredo do próprio filme, até mesmo realizar a leitura de lábios em certos momentos. Além disso, um agente de conversação moderno incluirá um sistema que proporcionará o reconhecimento automático da fala e a compreensão da linguagem, que será, em termos resumidos, o estágio de entrada da linguagem e a saída desta, em que parte do sistema ou um sistema adicional é responsável pelo diálogo, pelo planejamento da resposta e pela síntese de fala. Continuando, outra tarefa útil que se relaciona diretamente ao idioma refere-se à tradução automática, proporcionada pela introdução de algoritmos e ferramentas matemáticas, que, por sua vez, está ainda distante de soluções definitivas pelas dificuldades já mencionadas, como a complexidade do estabelecimento da linguagem natural, que se dá por meio de regras e mecanismos específicos e que podem provocar ambiguidades (JURAFSKY; MARTIN, 2008).

No Quadro 1, tem-se uma visão geral do PLN, com os principais objetivos, conhecimentos necessários sobre artes e humanidades e ciência e engenharia.

**Quadro 1.** Habilidades e conhecimentos envolvidos no processamento de linguagem natural

<b>Objetivos principais</b>	<b>Conhecimentos de artes e humanidades</b>	<b>Conhecimentos de ciência e engenharia</b>
Análise da linguagem	Manipular grandes <i>corporas</i> (conjuntos de textos escritos ou registros orais), explorando modelos linguísticos e testando afirmações empíricas	Uso de técnicas dentro da modelagem de dados, mineração de dados e da descoberta de conhecimento para analisar a linguagem natural
Linguagem tecnológica	Construir sistemas robustos para performar tarefas linguísticas com aplicações tecnológicas	Uso de algoritmos linguísticos e estrutura de dados em programas robustos de processamento de linguagem

*Fonte:* Adaptado de Bird, Klein e Loper (2009).

Assim, é possível notar que, de maneira geral, o que distinguirá sistemas cujas aplicações se destinam ao PLN dos demais sistemas de processamento será o próprio uso do conhecimento da linguagem, em níveis mais complexos ou não conforme a aplicação em si. A seguir, você verá as principais áreas nas quais o PLN atua, envolvendo os conhecimentos linguísticos.

## 2 Atuação do processamento de linguagem natural quanto às áreas do conhecimento linguístico e exemplos práticos

O conhecimento linguístico principal para o uso do PLN nas mais diversas aplicações modernas e clássicas envolverá em grande parte dos casos, com exceção de aplicações específicas e restritas usos determinados, as competências linguísticas (JURAFSKY; MARTIN, 2008) listadas a seguir.

- **Fonética e fonologia:** o conhecimento relacionado ao som da linguagem.
- **Morfologia:** tratará sobre os componentes significativos das palavras.
- **Sintaxe:** retrata a relação estrutural existente entre palavras.
- **Semântica:** conhecimento dos significados.

- **Pragmática:** retrata a relação entre o significado e as intenções e os objetivos do locutor.
- **Discurso:** responsável pelo conhecimento das unidades maiores da linguística.

Essas competências e seu respectivo tratamento serão vistos de modo geral, dentro de suas funções e exemplos no PLN, com a introdução das tarefas executadas nesse processamento.

## Principais tarefas no processamento de linguagem natural e exemplos práticos

Algumas das principais tarefas a realizar dentro do PLN poderão estar ligadas diretamente à sintática, como a **segmentação**, o **part of speech tagging** e a **análise de dependência**. Na segmentação, tem-se a divisão de um texto em frases ou em palavras, por exemplo, ou mesmo a divisão de palavras que formam esse texto nos denominados morfemas, unidades significativas dentro da palavra. No *part-of-speech tagging* (“marcação de partes da fala”, há uma análise baseada no “rótulo” de cada palavra dentro de uma frase dada. Para entendê-lo, considere a frase: “Eu me interesso por inteligência artificial” — “eu” é o sujeito, por exemplo, e nesses rótulos apresenta-se a função sintática de cada palavra. Por último, existe a análise de dependência, que pode ser feita por meio da construção de uma árvore de dependência como uma estratégia para compreensão da relação entre as partes da frase. Assim, considerando a frase exemplificada, o verbo poderá ter outras partes relacionadas a ele, como um objeto ou até mesmo um pronome. Uma das principais aplicações da análise de dependência dentro do PLN consiste na análise de sentimento, que, considerando novamente a dependência entre os itens da frase, poderá revelar à máquina, por exemplo, se determinada pessoa se interessa por inteligência artificial.

No exemplo a seguir, você verá a execução da segmentação de palavras, para entender essa ação no contexto do PLN.



## Exemplo

Considere a seguinte frase:

As eleições dos EUA resultaram na vitória do candidato A.

Realizando a segmentação por palavras, deverão ser retornadas as seguintes divisões:

As; eleições; dos; EUA; resultaram; na; vitória; do; candidato; A

Assim, o computador deverá conseguir não somente compreender que os espaços em branco significam, nesse caso, o começo e o fim das palavras, mas também entender corretamente que EUA também devem ser considerados um elemento único.

Outra questão bastante relevante com relação à segmentação e que deve ser considerada reside no fato de que essa tarefa não é algo tão trivial, especialmente quando consideramos a multiplicidade de idiomas existentes e suas estruturas diversas. Assim, considerando como exemplo a língua inglesa, pode-se notar que as palavras nela são separadas por espaços em branco, o que compreende um padrão para muitas outras línguas, como o português, embora haja exceções a essa regra (p. ex., o chinês).

Agora tenha em conta o próximo caso, em que se utiliza a análise de sentimento, tomando como exemplo o Twitter para estudar o cenário estabelecido na Eleição Geral de 2015 no Reino Unido.



## Exemplo

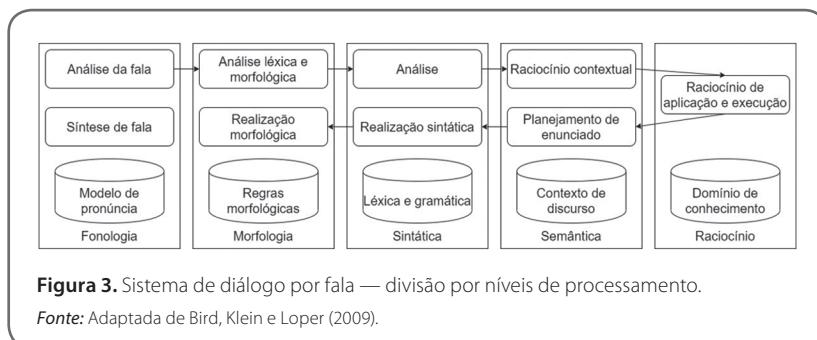
Considere o seguinte exemplo de aplicação do PLN, destinado às análises sobre a Eleição Geral de 2015 no Reino Unido (BURNAP *et al.*, 2015): pela análise dos 140 caracteres de cada mensagem elaborada no ambiente do Twitter, os denominados *tweets*, obtiveram-se diversos resultados a analisar e que forneceram previsões concretas sobre a mudança de compartilhamento de voto em 2015 a partir de 2010, com base na própria rede social e em vitórias projetadas.

Contudo, embora a análise de sentimentos configure-se como uma ferramenta para geração de diversos resultados e grande aplicabilidade, deve-se tomar alguns cuidados, sobretudo do ponto de vista do uso de dados disponibilizados pela própria internet. Assim, entende-se que, ao mesmo tempo que há uma vasta quantidade de informações disponíveis, existe uma quantidade expressiva nesse montante que poderá ser irrelevante ao contexto de análise. Ainda, as fontes e os formatos dos dados são diversos, em muitos casos não há uma estrutura coesa dos dados, além do fato de que, especialmente no caso de dados digitais, o uso de informalidades na fala e na escrita é expressivamente mais comum, o que também pode dificultar as análises. Também é preciso considerar a possibilidade de manipulação dos dados, a existência de boatos e a variável temporal, pois em diversas análises e para vários tipos de dados o tempo poderá atuar diretamente na usabilidade das informações fornecidas pelos dados.

Com relação à semântica, é possível citar a tradução automática, o *named entity recognition* (NER), determinadas **representações semânticas, a análise de tópicos e a própria análise de sentimento**. A tradução automática, já introduzida, é, sem dúvidas, uma das principais funcionalidades dentro do PLN. Já o NER, o reconhecimento de entidades nomeadas, proporciona a distinção em textos ou frases do que é, dentro dessas estruturas, uma pessoa, um lugar ou mesmo uma data, a partir emprego apenas das estruturas sintáticas e semânticas, sem o auxílio externo de outros bancos de dados. Na representação semântica, analisa-se como é possível representar uma palavra ou uma frase semanticamente, para extrair relações para comparação, analisando palavras plurais, sinônimos, em contextos similares, etc. Por análise de tópicos, entende-se a capacidade de extração de características gerais de textos, como os assuntos tratados, as similaridades semânticas entre diferentes documentos, etc. Já a análise de sentimento, conforme mencionado, poderá revelar posicionamentos positivos e negativos, mas também ser utilizada em aplicações diversas do PLN, executada do ponto de vista semântico.

Adicionalmente à competência da semântica, é necessário definir a tarefa de análise semântica, que, conforme Goddard e Schalley (2010), tem como objetivo final, tanto para pessoas quanto para sistemas de PLN, entender o enunciado: não apenas ler o que está escrito, mas também compreender a declaração. Além disso, do ponto de vista pragmático, há aplicações como extração de informação, criação automática de resumos, *data mining*, tradução automática, tratamento de parâmetros de pesquisa de usuários, sistemas de representação do conhecimento, etc.

Por último, há ainda tarefas diretamente relacionadas ao discurso e à fala, como a sumarização automática, em que são utilizados algoritmos capazes de coletar todas as informações de um livro e fornecer resumos, e o próprio diálogo, cujos principais exemplos são os *chatterbots*, importantes algoritmos que proporcionam o diálogo direto entre a máquina e o indivíduo, além de assistentes dentro de sistemas, como o Google Assistant, a Siri, etc. Nesses exemplos, destaca-se a capacidade de “pensar” adquirida. Na Figura 3, é observada a divisão por níveis de um sistema dessa natureza, de diálogo por fala.



Note que o sistema de diálogo por fala é bastante completo do ponto de vista do PLN, com boa parte dos elementos linguísticos e complexidades vistos até agora. Assim, para compreender como de fato esse sistema funciona, considere o exemplo a seguir.



### Exemplo

Suponha que você resolva perguntar a um sistema de diálogo como o Google Assistant se você deve levar um guarda-chuva com você amanhã, quando sair de casa. Em um primeiro momento, esse áudio será gravado e se fará a análise da fala, quebrando o áudio em palavras e frases, traduzindo-a em forma de texto e realizando uma análise morfológica, sintática e léxica, transformando o que foi falado de fato em palavras isoladas. Na etapa de análise sintática (mais conhecida do inglês *parsing*), entendem-se quais os elementos são verbos, sujeitos, etc., e, assim, se realiza a análise semântica, para entender de fato o que o interlocutor disse. No processamento semântico, tem-se o

“raciocínio” do diálogo, caso em que o Google, por exemplo, analisará pela localização via GPS a previsão do tempo no local para fornecer a resposta correta; então, no processamento de raciocínio, tem-se a mudança de sentido do fluxo de execuções, para que o assistente consiga responder corretamente ao indivíduo. Durante a “montagem” da resposta, são levados em consideração processos similares, para a correta formulação da resposta e análise morfológica da resposta a caminho para que sejam estabelecidos, primeiro, os morfemas e, por último, a estrutura da própria palavra.

Além disso, o processo de conversão da linguagem natural em uma representação útil ao computador, por meio de ferramentas da linguística, é considerado comumente o primeiro componente geral do PLN: o NLU (do inglês, *natural language understanding*, ou “entendimento da linguagem natural”). Para tal finalidade, é possível citar, por exemplo, a análise de sentimentos, a análise de palavras-chave dentro de um texto, etc. O segundo componente visto dentro da PLN, por sua vez, compreende o processo de geração da linguagem natural, a partir da saída de uma máquina, abreviado como NLG (do inglês, *natural language generation*, ou “geração de linguagem natural”). Nesse caso, ainda é possível citar assistentes virtuais e até mesmo o processo de geração de legendas a partir de um vídeo.

A seguir, tem-se uma introdução aos principais desafios, que também está diretamente relacionado à linguística: a ambiguidade.

## Ambiguidade no processamento de linguagem natural

A ambiguidade pode representar um desafio persistente dentro do PLN como um todo, como você verá a seguir. Assim, Jurafsky e Martin (2008) destacaram que a ambiguidade envolve todas as complexidades e análises linguísticas vistas anteriormente, pois a maioria das tarefas ao longo do processamento não só da fala, mas também da linguagem pode ser vista como uma ambiguidade resolutiva, em uma ambiguidade a cada um desses níveis.

Diz-se, então, que uma entrada é ambígua para o PLN se várias estruturas linguísticas alternativas ambíguas puderem ser construídas para esta, como no exemplo a seguir.



## Exemplo

Considere a seguinte frase:

Andréia pediu a Fabiano que pegasse sua mochila na sala.

Note que, para essa simples frase, existem ainda outras formas de construir uma estrutura diferente, mas que produz o mesmo significado, pois, afinal, a mochila de quem deveria ser pegada?

A ambiguidade poderá ocorrer também em outras formas, dependendo da construção do texto e das frases, o que gera, basicamente, incoerências no PLN. Para que isso se torne ainda mais claro, basta você pensar o seguinte: o seu entendimento sobre a frase do exemplo foi dificultado? A instrução estava completamente clara? O que se deve analisar é que, similarmente ao tratamento de ambiguidades na comunicação humana, dentro das construções da linguagem natural, o computador estará ainda mais propenso a erros.

A seguir, você entenderá de maneira mais direta a atuação de algoritmos e sistemas de PLN em diferentes problemas e aplicações, evidenciando, sobretudo, a interdisciplinaridade do uso das técnicas.

### 3 Atuação do processamento de linguagem natural quanto aos diferentes problemas

A partir de agora, serão enfatizadas as aplicações e resoluções de problemas a partir de técnicas do PLN, que envolvem diversas áreas e disciplinas.

Iniciando pela própria linguística, em problemas destinados à melhoria de comunicação a partir da realização de tradução há diversos exemplos, inclusive disponibilizados gratuitamente, como a transcrição de fala, em que o locutor fala e o computador é responsável pela transcrição da fala, ou o contrário. O próprio tradutor do Google é um exemplo desse tipo de aplicação, em que se pode citar a seguinte situação de transcrição de fala a partir do PLN: o ato de digitar uma palavra, selecionar o idioma para aprender a pronúncia é um recurso bastante utilizado para ouvi-la a partir da resposta do tradutor. Ainda sobre tradutores, tem-se o *neural machine translation* (NMT), a máquina de tradução automática também fornecida pelo Google, que se configura como

uma modificação que promoveu diversos ganhos à tradução obtida ao final, já que a tradução nessas ferramentas era feita, no início, frase à frase, e hoje o é a partir de implementações de técnicas de *deep learning*. Assim, atualmente o uso desse tipo de técnica permite a memorização das palavras, impactando diretamente no sentido final do texto ou no conjunto de textos traduzidos, melhorando o sentido final construído.

Os *chatbots* configuraram-se também como grandes exemplos do uso atual do PLN, como alguns dos principais e mais usados de sistemas de diálogo. Há, ainda, sistemas de perguntas de respostas, criados a partir do uso do PLN, em que geralmente o usuário passa por um sistema de PLN que, então, será capaz de fornecer as respostas necessárias. Além disso, uma das principais possibilidades de aplicação desse tipo de sistema é sua inserção em fóruns e até mesmo em plataformas destinadas ao ensino a distância, possibilitando aos alunos respostas sobre dúvidas de determinados conteúdos.



### Saiba mais

No ramo das linguagens de programação, existem vários exemplos de sistemas de perguntas e respostas, capazes de fornecer auxílio e exemplos práticos, a fim de sanar dúvidas sobre o tema de forma *on-line*. Um deles é o Stack Over Flow, um site de perguntas e respostas destinado a programadores, curiosos e entusiastas da área, que exige apenas a inscrição do usuário para que possa usufruir dos serviços. Além disso, no sistema qualquer um desses usuários pode realizar as perguntas e qualquer outro respondê-las, além do fato de as melhores respostas receberem votações positivas até atingirem a classificação de 1º lugar.

<https://qrgo.page.link/2SeFt>

Os sistemas de perguntas e respostas funcionam na prática da seguinte maneira: um robô é treinado para aprender as respostas, para que estar apto a responder às perguntas de um aluno ou outro participante de um fórum sem auxílio humano.

Contudo, destaca-se outro problema, que pode ser auxiliado por técnicas e programas do PLN, relacionado à absorção de um grande volume de conteúdos e informações: a sumarização de textos. Um exemplo prático de algoritmos que usam essa tarefa pode ser visto em Cabral (2015), em que se propõe uma plataforma destinada à tarefa.

O *Image Captioning*, a tarefa de ter uma imagem dada e descrevê-la, é proporcionado pelas técnicas do PLN e constitui-se em uma aplicação também bastante multidisciplinar, funcionando basicamente da seguinte forma: suponha que há uma fotografia de uma flor em um jardim. Para a tarefa de reconhecimento, o computador deverá ser capaz de armazenar todas as dependências e elementos, com técnicas de redes neurais e em vídeo. Um exemplo prático dessa tarefa para imagens é apresentado a seguir.



### Exemplo

O reconhecimento de imagem pode ser utilizado em aplicações visando à resolução e ao atendimento de diversos problemas. Assim, neste exemplo você verá um trabalho realizado para auxiliar na monitoração de alagamentos.

Em um dos trabalhos do pesquisador Roger Wang, um especialista na aplicação de dados na investigação de problemas ambientais, da Universidade de Dundee, na Escócia, foram usadas duas técnicas dentro da inteligência artificial — o processamento de linguagem natural e a visão computacional — para monitorar a ocorrência de alagamentos, utilizando como dados os tweets no Twitter e recursos gráficos como fotos disponibilizadas no MyCoast (um aplicativo criado para supervisionar o litoral norte-americano com imagens fornecidas pelos usuários).

A própria análise de sentimentos, vista ao longo deste texto, representa uma das principais aplicações do PLN, especialmente quando se observam a aplicabilidade e a tendência atual do uso desse tipo de análise no universo crescente do uso de redes sociais e de grandes volumes de dados digitais diversos. Além dos exemplos destacados, sabe-se que grande parte do próprio funcionamento das redes sociais se dá pela análise de sentimentos, com uma tendência cada vez maior de uso em estratégias comerciais.



### Saiba mais

Existem trabalhos diretamente relacionados às redes sociais e ao uso do PLN, capazes de fornecer informações interessantes para diversas áreas. Nesse sentido, sugerimos ao leitor o livro *Análise de redes sociais: uma visão computacional*, de Gabardo (2015).

Por último, as pesquisas mais recentes indicam ao estado da arte do PLN que o uso de técnicas como a aprendizagem profunda (*deep learning*) aponta um novo horizonte de possibilidades, tanto no PLN quanto na própria resolução de problemas ainda em aberto, referentes à inteligência artificial, mas com interface para áreas multidisciplinares.



## Referências

- BERGER, A. L. et al. The Candide system for machine translation. In: WORKSHOP ON HUMAN LANGUAGE TECHNOLOGY, 1994, Stroudsburg. *Proceedings* [...]. Stroudsburg: Association for Computational Linguistics, 1994. p. 157–162.
- BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python*: analyzing text with the natural language toolkit. [S. l.]: O'Reilly Media, 2009.
- BURNAP, P. et al. 140 characters to victory? Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, [s. l.], v. 41, p. 230–233, 2016.
- CABRAL, L. S. *Uma plataforma para sumarização automática de textos independente de idioma*. 2015. Tese (Doutorado em Engenharia Elétrica) — Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Pernambuco, Recife, 2015.
- CARBONELL, J. G. *Subjective understanding*: computer models of belief systems. New Haven: Yale University, 1979.
- CHAMBERLAIN, W. *The policeman's beard is half constructed*. [S. l.]: Warner Books, 1984.
- COLBY, K. M.; WEBER, S.; HILF, F. D. Artificial paranoia. *Artificial Intelligence*, [s. l.], v. 2, n. 1, p. 1–25, 1971.
- CULLINGFORD, R. E. *Script application*: computer understanding of newspaper stories. 1978. Disponível em: <https://apps.dtic.mil/docs/citations/ADA056080>. Acesso em: 07 mar. 2020.
- DALE, R. Classical approaches to natural language processing. In: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Handbook of natural language processing*. 2nd ed. Boca Raton: Chapman & Hall, 2010.
- GABARDO, A. C. *Análise de redes sociais*: uma visão computacional. [S. l.]: Novatec, 2015.
- GODDARD, C.; SCHALLEY, A. C. Semantic analysis. In: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Handbook of natural language processing*. 2nd ed. Boca Raton: Chapman & Hall, 2010.
- HUTCHINS, J. *The history of machine translation in a nutshell*. [S. l.: s. n.], 2005.
- HUTCHINS, J. The Georgetown-IBM experiment demonstrated in January 1954. In: CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS, 6., 2004, Washington. *Proceedings* [...]. Springer: Heidelberg, 2004. p. 102–114.

- JURAFSKY, D.; MARTIN, J. H. *Speech and language processing*. 2nd ed. [S. I.]: Pearson Prentice Hall, 2008.
- LADEFOGED, P.; MADDIESON, I. *The sounds of the world's languages*. Oxford: Blackwell, 1996.
- LEHNERT, W. G. Plot units and narrative summarization. *Cognitive Science*, [s. I.], v. 5, n. 4, p. 293–331, 1981.
- LEHNERT, W. G. A conceptual theory of question answering. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 5., 1977, Cambridge. *Proceedings [...]*. Cambridge: MIT, 1977. v. 1, p. 158–164.
- LORITZ, D. *How the brain evolved language*. Oxford: Oxford University Press, 2002.
- MATACIC, C. Human speech may have a universal transmission rate: 39 bits per second. *Science*, [s. I.], 4 Sept. 2019. Disponível em: <https://www.sciencemag.org/news/2019/09/human-speech-may-have-universal-transmission-rate-39-bits-second>. Acesso em: 07 mar. 2020.
- MATTHEWS, P. H. *Linguistics: a very short introduction*. Oxford: Oxford University Press, 2003.
- MEEHAN, J. R. *The metanovel: writing stories by computer*. 1976. Disponível em: <https://apps.dtic.mil/docs/citations/ADA031625>. Acesso em: 07 mar. 2020.
- RIESBECK, C. K. et al. Inference and paraphrase by computer. *Journal of the ACM*, [s. I.], v. 22, n. 3, p. 309–328, 1975.
- SACKS, O. *Seeing voices: a journey into the world of the deaf*. [S. I.]: Pan Macmillan, 2009.
- SCHANK, R. C.; WILENSKY, R. A goal-directed production system for story understanding. In: WATERMAN, D. A.; HAYES-ROTH, F. (ed.). *Pattern-directed inference systems*. Cambridge: Academic Press, 1978. p. 415–430.
- TURING, A. M. Computing machinery and intelligence. In: TURING, A. M. *Parsing the turing test*. Dordrecht: Springer, 2009. p. 23–65.
- WADE, N. Early voices: the leap to language. *The New York Times*, New York, 15 July 2003. Disponível em: <https://www.nytimes.com/2003/07/15/science/early-voices-the-leap-to-language.html>. Acesso em: 07 mar. 2020.
- WEIZENBAUM, J. ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, [s. I.], v. 9, n. 1, p. 36–45, 1966.
- WINOGRAD, T. Understanding natural language. *Cognitive Psychology*, [s. I.], v. 3, n. 1, p. 1–191, 1972.

## Leituras recomendadas

JABBERWACKY.COM. *About the Jabberwacky AI*. 2011. Disponível em: <http://www.jabberwacky.com/j2about>. Acesso em: 07 mar. 2020.

STACK OVERFLOW. *Explore nossas perguntas*. 2020. Disponível em: <https://pt.stackoverflow.com/>. Acesso em: 07 mar. 2020.

WINOGRAD, T. *Procedures as a representation for data in a computer program for understanding natural language*. 1971. Disponível em: <https://dspace.mit.edu/handle/1721.1/7095>. Acesso em: 07 mar. 2020.



### Fique atento

Os *links* para sites da Web fornecidos neste capítulo foram todos testados, e seu funcionamento foi comprovado no momento da publicação do material. No entanto, a rede é extremamente dinâmica; suas páginas estão constantemente mudando de local e conteúdo. Assim, os editores declaram não ter qualquer responsabilidade sobre qualidade, precisão ou integralidade das informações referidas em tais *links*.

Encerra aqui o trecho do livro disponibilizado para esta Unidade de Aprendizagem. Na Biblioteca Virtual da Instituição, você encontra a obra na íntegra.

Conteúdo:



SOLUÇÕES  
EDUCACIONAIS  
INTEGRADAS