# LINGUÍSTICA COMPUTACIONAL E SUAS SUBÁREAS

Wilton Ribeiro Cruz*

**Resumo:**
*O objetivo desta resenha é apresentar as teorias que compõe o campo de pesquisa e desenvolvimento em Linguistica Computacional e mostrar a relação existente entre Linguística de Corpus e Processamento de Língua Natural, pois são complementares.*

**Palavras-chave:** linguística computacional; PLN; processamento de língua natural.

**Abstract:**
*The objective of this review is to present the theories that compose the field of research and development in Computational Linguistics and to show the relation between Corpus Linguistics and Natural Language Processing, as they are complementary.*

**Keywords:** computational linguistics; PLN; natural language processing.

## Introdução

Um linguista americano chamado *Ray Jackendoff* fez uma observação irônica entre os linguistas computacionais que dizia assim: Quando os informatas resolvem pedir auxílio aos linguistas teóricos em algum de seus projetos de Processamento de Linguagem Natural (*PLN*), seus programas acabam se tornando menos eficientes. Complementando essa linha de pensamento com um linguista brasileiro da PUC-RS, Othero (2001), afirma que esses dois tipos de pesquisadores nem sempre têm os mesmos objetivos no estudo da linguagem humana.

* Graduado em Letras Port/Inglês UEMG e com especialização em Língua Portuguesa.Pesquisador nas áreas de Linguística textual,Linguística Aplicada,abordando conceitos interdisciplinares em Tecnologia e Linguagem Humana. Lattes: http://lattes.cnpq.br/3984604274622971. Contato: piumhi10@gmail.com.

Linguistas dessa área fazem uma pequena comparação em dizer que o linguista tem um compromisso com a verdade, e o informata, com a eficácia.

Aristóteles mesmo já havia afirmado que, "o objetivo do conhecimento teórico é a verdade, enquanto o do conhecimento prático é a eficácia".

Com toda essa informação podemos entender que ainda há uma ampla relação entre *PLN* e Linguística de Corpus onde cresce uma área totalmente nova e interdisciplinar.

## 2. Evolução dos sistemas de Processamento de Língua Natural (PLN)

No caráter histórico existe uma síntese de evolução nos estudos de *PLN*, e um grau de sofisticação linguística que ela trouxe para o ser humano é inegável e basilar.

Historicamente podemos resumir a fio como tudo começou no decorrer de décadas, leremos os tópicos abaixo:

*Década de 50*: A Tradução automática foi a que mais motivou uma sistematização computacional das classes de palavras e da gramática tradicional logo depois veio uma possível identificação computacional de poucos tipos de constituintes oracionais.

*Década de 60*: Depois de novas aplicações e criações de formalismos gramaticais foram os primeiros tratamentos computacionais das gramáticas livres de contexto, ou melhor, Gramática de Cláusulas Definidas, sigla em inglês, (*DCG's*) depois bem mais tarde veio a criação dos primeiros analisadores sintáticos, temos a exemplos do *Parser* sintático usado como uma linguagem de programação muito conhecida do *Prolog*[1]. As primeiras formalizações do significado em termos de redes semânticas foram baseadas em diagramas arbóreos e hierarquia de conceitos.

*Década de 70*: Consolidou-se os estudos de *PLN*, e mais adiante propôs uma implementação de parcelas das primeiras gramáticas e analisadores sintáticos em busca de formalização de fatores pragmáticos e discursivos.

Temos um histórico minucioso de sua evolução e concluímos em resumo onde tudo começou como motivação a tradução automática de determinada língua estrangeira para a materna.

### 2.1. Campos de estudos em PLN

[1] Prolog é uma linguagem de programação que se enquadra no paradigma de Programação em Lógica Matemática. É uma linguagem de uso geral que é especialmente associada com a inteligência artificial e lingüística computacional.

- Processamento de Língua Natural (PLN) ou (*Natural Language Processing*).

Para termos de análise, *PLN* é uma subárea da *Inteligência Artificial*[2] e da *linguística aplicada*[3] que estuda os problemas da geração e compreensão automática de línguas humanas naturais. Ela se preocupa diretamente com o estudos de linguagem voltado para a criação de softwares e sistemas multiagentes de auto padrão e robusteza, exemplo de Tradutores automáticos, sintetizadores de voz e reconhecimento, *chatterbots*, *parsers*, etc.

São programas inteiramente capazes de processar a linguagem natural.

S. L. Pereira afirma que a pesquisa em *PLN* está voltada essencialmente a três aspectos da comunicação em língua natural:

- Som: Fonologia
- Estrutura: morfologia e sintaxe
- Significado: semântica e pragmática

A fonologia a priori, está a relacionada ao reconhecimento dos sons que compõem as palavras de uma língua, já a morfologia reconhece as palavras em termos das unidades primitivas que a compõem e por fim a sintaxe que define a estrutura de uma frase, e sempre com base na forma de como essas palavras se relacionam na frase.

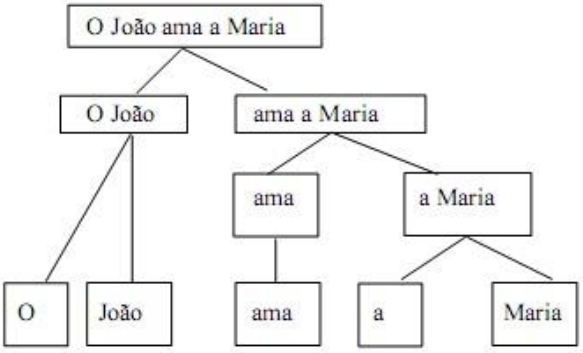
A semântica se associa significado a uma estrutura sintática, em termos dos significados das palavras que a compõem podemos associar o significado; diferentemente da pragmática que verifica se o significado associado a uma estrutura sintática é realmente o significado mais apropriado no contexto considerado, teoricamente um termo apropriado para isso seria, “agramatical”.

Logo abaixo, analisaremos para fins didáticos um diagrama arbóreo que melhor definirá a explicação de S.L Pereira em relação ao significado apropriado:

[2] A Inteligência Artificial (IA) é uma área de pesquisa da ciência da computação e Engenharia da Computação, dedicada a buscar métodos ou dispositivos computacionais que possuam ou simulem a capacidade racional de resolver problemas, pensar ou, de forma ampla, ser inteligente.

[3] Linguística aplicada é um campo interdisciplinar de estudo que identifica, investiga e oferece soluções para problemas relacionados com a linguagem da vida real. Alguns dos campos acadêmicos relacionados à lingüística aplicada são educação, linguística, psicologia, antropologia e sociologia, informática, história, geografia.

O João ama a Maria.




Na gramática em que permite analisar a sentença "O João ama a Maria", a diferença entre regras sintagmáticas e lexicais pode ser claramente identificada pelo fato de que as últimas definem conjuntos, que são os conjuntos dos itens lexicais relativos a cada uma das categorias lexicais, formando assim o diagrama contistuinte sintagmática.

## 2.2. O Formalismo Gramatical

Não iremos encarar a sentença como um mero aglomerado de palavras, unidas uma a outra de qualquer forma pois há entre o nível da palavra e o da frase uma outra forma de organização que é o sintagma (ou constituinte), Radford (1981: 69).

A princípio não entraremos em detalhes sobre alguma formalização de estrutura frasal do *PB*, porque estaríamos aprofundando o tema a respeito de *PLN*.

O *PLN* pode ser definido como " a utilização de conhecimentos sobre a língua e a comunicação humana, tanto para a comunicação com sistemas computacionais como para melhorar a comunicação entre os seres humanos" (Santos, 2002).

## 3. A Linguística de Corpus

Linguística de corpus (quer dizer corpo em latim) é uma área da Linguística que se ocupa da coleta  e análise de corpus, onde um determinado conjunto de dados linguísticos serão coletados.O léxico é o mais importante para a Linguistica de corpus

segundo Tony Berber 2004, podemos perceber os dicionários de inglês atuais que são produzidos com base de corpus, temos também o COBUILD[4] que é um dos maiores bancos de dados da língua inglesa já criados para produzir dicionários, gramáticas e livros didáticos para o ensino do inglês.

Até aqui percebemos a importância do corpus para se pesquisar uma determinada língua, pois é uma área de extensa participação de pesquisa, baseada em corpora[5].

### 3.1. Um pouco de História

Segundo estudiosos afirmam que já existia um corpora antes do computador, pois na Grécia Antiga o grande imperador Alexandre, o Grande, definiu o Corpus *Helenístico*, já na antiguidade produziam-se corpora de citações da Bíblia. Depois em meados do século XX houve muitos pesquisadores que se dedicaram à descrição da linguagem por meio de corpora, pesquisadores como *Thordike* e vários linguístas empenhado no papel de coletar o corpora mas nessa época não eram ainda eletrônicos, e ainda eram mantidos e manuseados manualmente. Linguistas acreditam que a funçao primordial naquela época de se usar um corpus era no ensino de línguas, e logo foi um corpus não-computadorizado que definiu um conceito de corpora atual.

*SEU* (Survey of Engish Usage), compilado por Raldolf Quirk em Londres, a partir de 1959, também foi planejado para um milhão de palavras, serviu para outras referências de corpora, tudo isso tornou-se grandes ferramentas a serviço da Linguística de Corpus.

### 3.2. A importância do Corpus

A linguística de corpus surgiu com a necessidade de que estudiosos da língua precisavam se apoiar em usos reais, para fazerem generalizações ou esboçarem teorias a respeito do funcionamento linguístico. Atualmente, a linguística de corpus está intimamente ligada ao uso do computador, visto que os corpora/córpora (plural

[4] Cobuild, um acrônimo para Collins Birmingham University International Language Database, é um centro de pesquisa britânico criado naUniversidade de Birmingham em 1980 e financiado pela Collins Publishers.

[5] Corpora/córpora (plural de corpus) são eletrônicos. Assim, a linguística de corpus contemporânea caracteriza-se pela coleta e análise de corpora eletrônicos com o auxílio de ferramentas eletrônicas.

de corpus) são eletrônicos. Assim, a linguística de corpus contemporânea caracteriza-se pela coleta e análise de corpora eletrônicos com o auxílio de ferramentas eletrônicas.

O corpus deve ser constituído de dados autênticos (não inventados), legíveis por computador e representativos de uma língua ou variedade da língua da qual se deseja estudar.

O computador é a ferramenta primordial da Linguística de Corpus.

As ferramentas computacionais são geralmente utilizadas para reorganização e extração de informações no corpus, e que serve para uma observação e interpretação de dados, fornecendo novas perspectivas para uma análise linguística.

As ferramentas computacionais mais comuns são: Programas para listar palavras - fazem a contagem das palavras em um corpus;

Concordanciadores - programas que permitem que o usuário procure por palavras específicas em um corpus, fornecendo exaustivas listas para as ocorrências da palavra em contexto;

Etiquetadores - fazem análises automáticas do corpus e inserem etiquetas (códigos) de ordem morfossintática, sintática, semântica ou discursiva.

### 3.3. A definição de Corpus

Berber Sardinha, (2000), afirma que a definição mais completa de um conjunto de dados linguístico se baseia na origem, e os dados devem ser autênticos, pois o propósito do corpus deve ser um objeto de estudo linguístico, depois vem a composição do corpus que deve ser criteriosamente escolhido, a formatação também deve ser legível por computador, a representatividade de uma língua ou variedade e finalmente a extensão do corpus que é vasto para ser representativo.

## 4. Considerações finais acerca do conhecimento em Linguística Computacional.

Historicamente podemos entender que a Linguística Computacional, (*LC*) somente começou a ser divulgada depois da década de 60 nos Estados Unidos, e com o único intuito de se fazer computadores traduzirem textos automaticamente de uma determinada língua estrangeira para a materna.

Depois quando surgiu a Inteligência Artificial a *LC* se converteu em um ramo da *IA*, tratando com o nível de entendimento humano e o *PLN*. Para uma melhor definição do que é *LC* uma citação de *Vieira* et al:

> "De acordo com Vieira & Lima (2001, p. 1), a Linguística Computacional pode ser entendida como "a área de conhecimento que explora as relações entre lingüística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural";

Mesmo que a *PLN* seja uma disciplina aplicada à Ciência da Computação ela compartilha em vários temas com a Linguística de Corpus, porém hoje essas duas áreas se mantém independentes, e entendendo essa relação poderíamos olhar para os dois lados, pois como seria o processamento de língua natural sem um corpora autêntico e confiável?

*Thorndike* ousou coletar 18 milhões de palavras por meios manuais, o que não era confiável, porque o ser humano não é talhado para tarefas desse tipo. O trabalho era realizado com grande contigentes de assistentes. A pesquisa de *Käding* ,por exemplo, sobre a ortografia do alemão, consumiu a mão-de-obra de 5.000 analistas! As possibilidades de erro eram tão grande que para se obter de um léxico em uma gramática que processasse determinada língua caberia a Linguística de Corpus dar essa confiabilidade à *PLN* tratando-se em um termo mais restrito e definido dessa relação, Berber Sardinha afirma que " o *PLN* é uma disciplina com laços fortes com a Ciência da Computação, embora compartilhe vários temas com a Linguística de Corpus, as duas ainda mantêm-se independentes.

No brasil, a Linguística Computacional está em estágio inicial. A pesquisa em corpus se dá em universidades voltadas ao *PLN*, *Lexicografia* e a Linguística de Corpus, ambos as subáreas ganham campo tanto para fins comerciais como acadêmicos.

Ainda falta muito para aprimorar nossos conceitos em Linguística Computacional e pesquisadores que gostam de inovar poderiam ser mais práticos e menos teóricos.

Terminamos esta resenha com uma breve inferência acerca de trabalhos baseados na língua, com uma maravilhosa citação de Perini:

> "Para quem gosta de certezas e seguranças, tenho más notícias: a gramática não está pronta. Para quem gosta de desafios, tenho boas notícias: a gramática não está pronta. Um mundo de questões e

problemas continua sem solução, à espera de novas idéias, novas teorias, novas análises, novas cabeças. Perini (2003: 85)".

## Referências

BERBER SARDINHA, T. **Linguística de Corpus** - (8520416764).

JACKENDOFF, R. **Foundations of language**: brain, meaning, grammar, evolution. Oxford: Oxford University Press, 2002.

PERINI, M. A. **Sofrendo a gramática**. 3ª edição. São Paulo: Ática, 2003.

VIEIRA , R. e LIMA, V. L. S. (2001) **Lingüística computacional**: princípios e aplicações. In: IX Escola de Informática da SBC-Sul. Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul.

OTHERO, G. de A. **Lingüística Computacional**: uma nova área de pesquisa para os estudantes de Letras. Entrelinhas, ano 2, n. 5. São Leopoldo: UNISINOS, 2002.

OTHERO, G. de A. & MENUZZI, S. de M. **Lingüística computacional**: teoria e prática. São Paulo, Parábola Editorial, 2005, 128 p. ISBN: 85-88456-39-X.

OTHERO, G. de A. **A gramática da frase em português** - Algumas reflexões para a formalização da estrutura frasal em português.

OTHERO, G. de A. **Teoria X-Barra** - Gabriel de Ávila Othero.

OTHERO, G. de A. **Lingüística Computacional** - princípios e aplicações.

PEREIRA. S. L. **Processamento de Língua Natural**, artigo.