N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

Introdução ao Processamento de Línguas Naturais

PLN

Alexsandro dos Santos
Jorge Marques Pelizzoni
(organizadores)

N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

O que é PLN?

O PLN é uma área de trabalho situada entre a **linguística** e a **ciência da computação** que lida com os **aspectos computacionais da capacidade humana de linguagem**.

Objetivo: construir sistemas que, de alguma forma, tratam de problemas intrínsecos ao processo comunicativo verbal.

N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

Evolução dos sistemas de PLN

Década de 50: A Tradução automática

- sistematização computacional das classes de palavras da gramática tradicional
- identificação computacional de poucos tipos de constituintes oracionais

Década de 60: Novas aplicações e criação de formalismos

- primeiros tratamentos computacionais das gramáticas livres de contexto
- criação dos primeiros analisadores sintáticos
- primeiras formalizações do significado em termos de redes semânticas

Década de 70: Consolidação dos estudos do PLN

- implementação de parcelas das primeiras gramáticas e analisadores sintáticos
- busca de formalização de fatores pragmáticos e discursivos

Década de 80: Sofisticação dos sistemas

- desenvolvimento de teorias lingüísticas motivadas pelos estudos do PLN

Década de 90: Sistemas baseados em “representações do conhecimento”

- desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos lingüísticos e extralingüísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos lingüísticos

N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

Aglutinação de esforços

N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

A essência lingüística e tecnológica do PLN

- **Sistema de processamento automático de conhecimentos:**

"Assumimos que um computador não poderá simular uma língua natural satisfatoriamente se não compreender o assunto que está em discussão. Logo, é preciso fornecer ao programa um modelo detalhado do domínio específico do discurso. Além disso, o sistema possui um modelo simples de sua própria mentalidade. Ele pode se lembrar de seus planos e ações, discuti-los e executá-los. Ele participa de um diálogo, respondendo, com ações e frases, às frases digitadas em inglês pelo usuário; solicita esclarecimentos quando seus programas heurísticos não conseguem compreender uma frase com a ajuda das informações sintáticas, semânticas, contextuais e do conhecimento de mundo físico representadas dentro do sistema." (Winograd, 1972: 1)

- **Metas:**
  - projetar e implementar sistemas computacionais em que a comunicação homem-máquina possa se realizar, em última instância, por meio de línguas naturais;
  - projetar e implementar sistemas computacionais voltados, de um lado, para a própria investigação de teorias e modelos lingüísticos e, de outro, para a realização de tarefas que envolvem conhecimentos de natureza lingüística como, por exemplo, fazer revisão ortográfica e gramatical, traduzir, interagir por meio de perguntas e respostas e elaborar resumos.

N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

Estratégia de pesquisa integrada

DOMÍNIOS	PROBLEMAS	RECURSOS
Lingüístico	Como explicitar o conhecimento e o uso lingüístico?	Teorias da competência e do desempenho
↓↑	↓↑	↓↑
Representacional	Como representá-los?	Linguagens formais de representação
↓↑	↓↑	↓↑
Implementacional	Como coficar as representações?	Linguagens de programação e sistemas computacionais

N

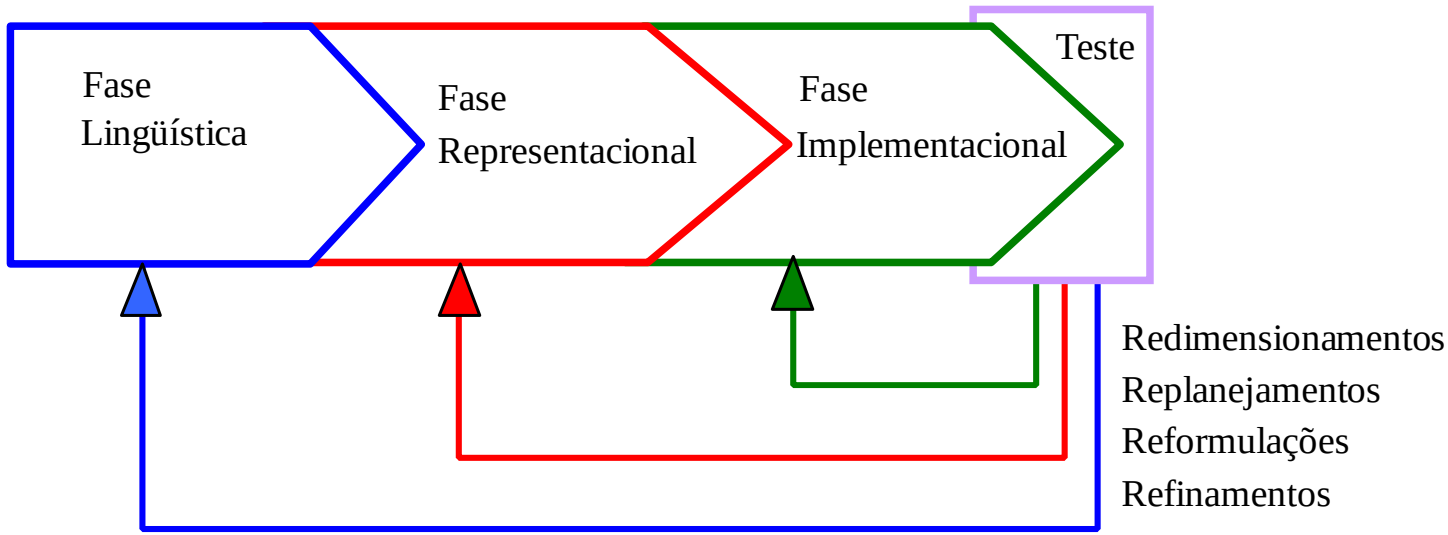
I

L

C

USP
UNESP
UFSCAR
UNICAMP

Fases de construção de um Sistema de PLN (SPLN)



N

I

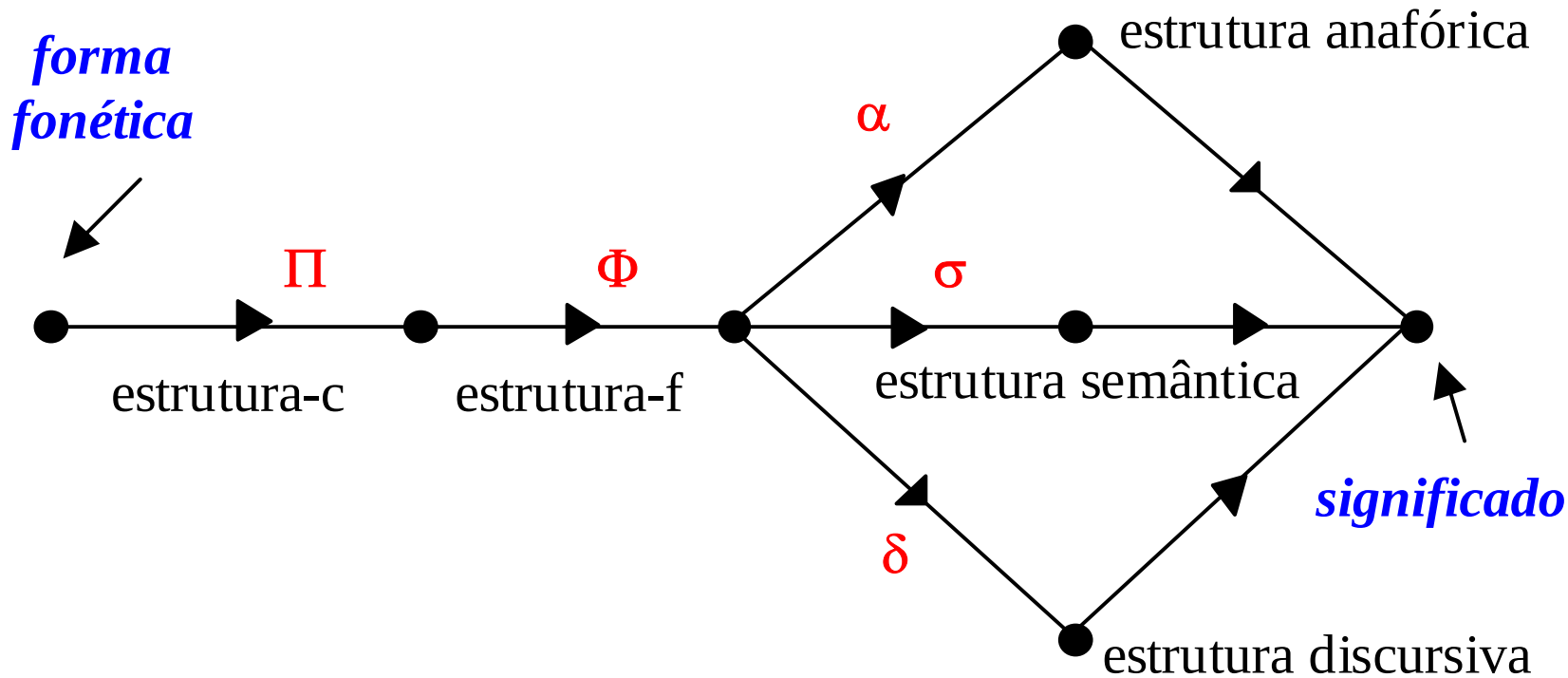
L

C

USP
UNESP
UFSCAR
UNICAMP

Fases de construção de um SPLN

Fase lingüística



N

I

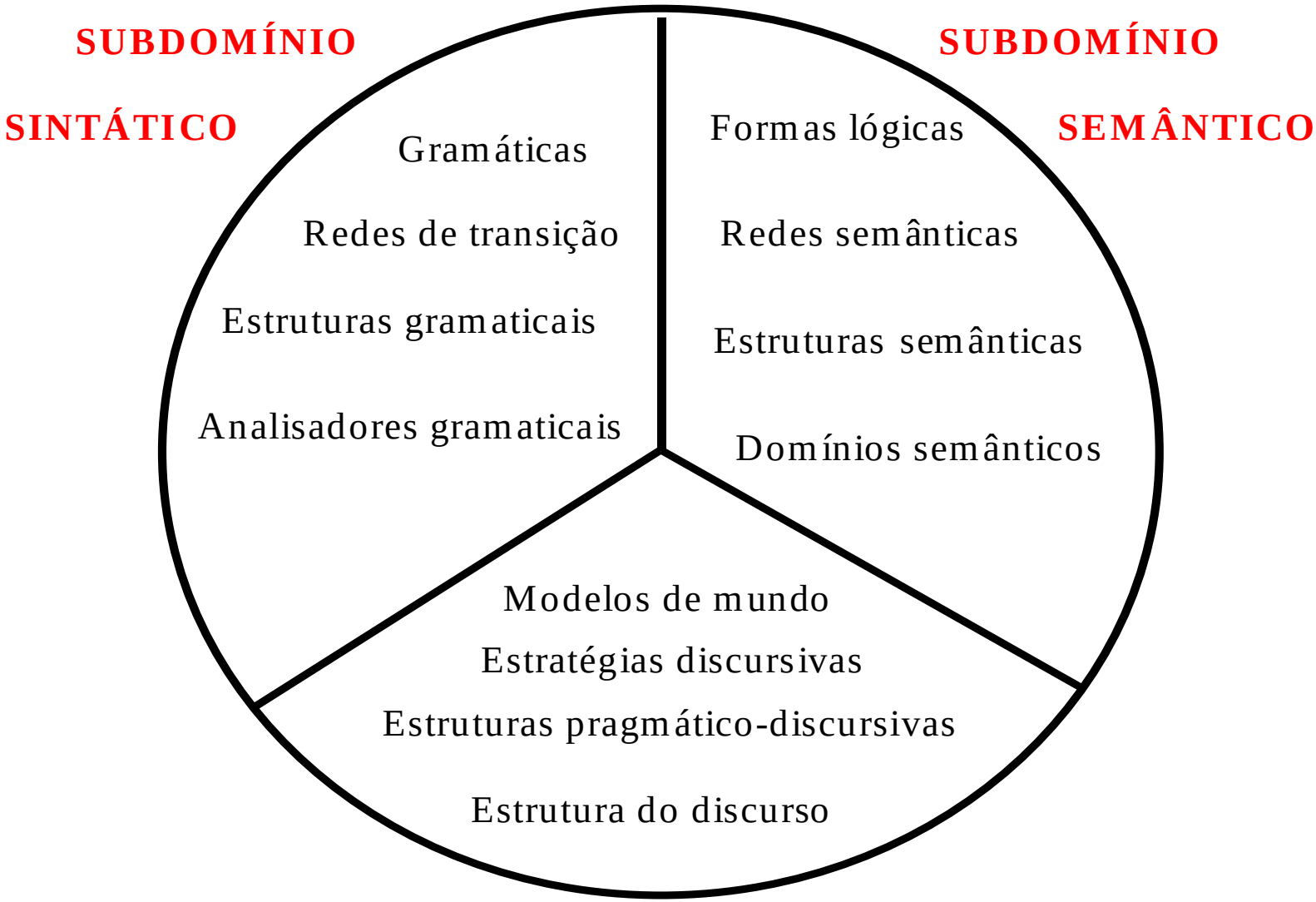
L

C

USP
UNESP
UFSCAR
UNICAMP

Fases de construção de um SPLN

Fase representacional (recursos)



N

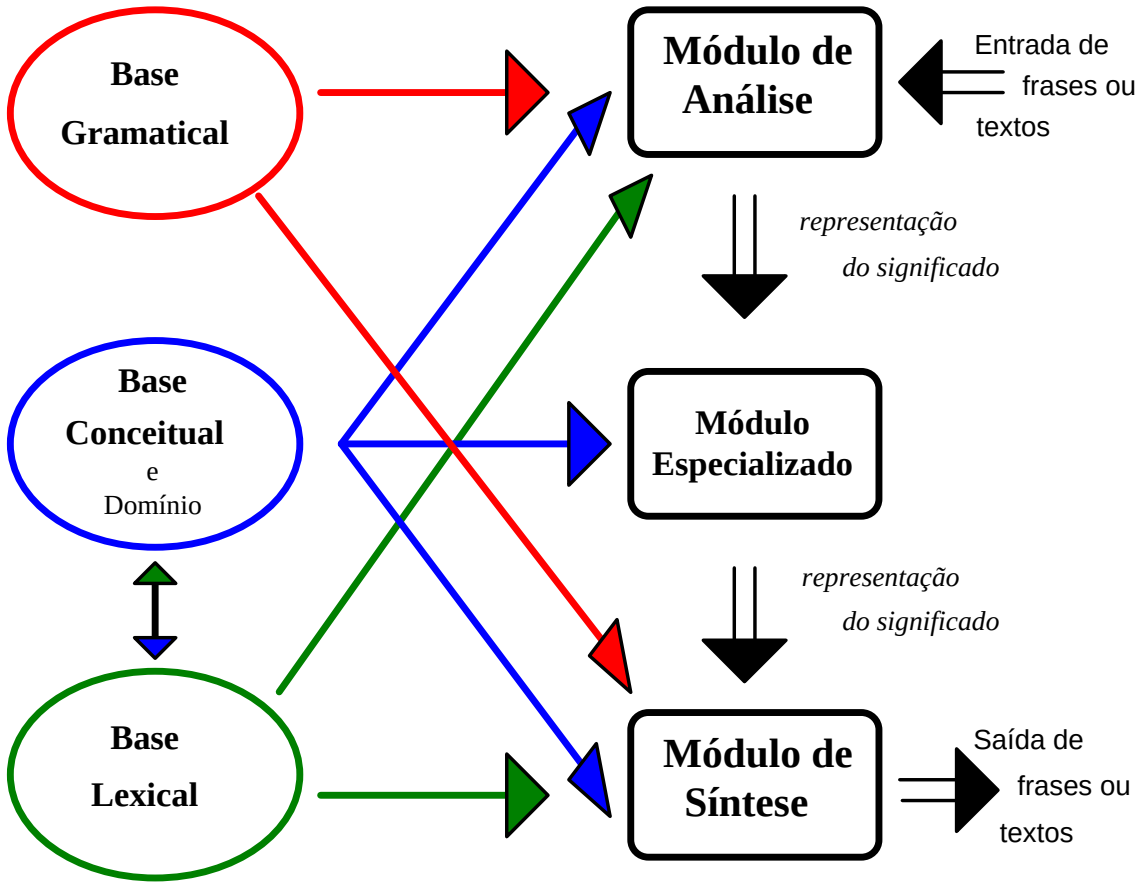
I

L

C

USP
UNESP
UFSCAR
UNICAMP

Fase de implementação



N

I

L

C

USP
UNESP
UFSCAR
UNICAMP

Aplicações

- Manipulação de bases de dados
- Sistemas tutores
- Automação de tarefas administrativas e gerenciais
- Programação automática
- Processamento automático de textos científicos
- Sistemas especialistas
- Tradução automática
- Sistemas “acadêmicos”

N

I

L

C

São Carlos
SP

Por que a Lingüística?

- É a ciência da comunicação verbal.
- A Lingüística.....⬇ preocupa-se com as formas lingüísticas

e seu conteúdo

⬇ descreve a língua

apresenta: - as formas

- os *processos*

- as *possibilidades* de formação

Cenário PLN

- Necessidade de *sistematizar* e *formalizar*

⬇ representação de formas e de processos

⇓

visualizar as possibilidades de boa formação lingüística

N

I

L

C

São Carlos
SP

Os segmentos lingüísticos

I. Unidades mínimas da *comunicação*

- Texto
- Sentença
  - ⬇ “Eu quero ir ao banheiro”
  - ⬇ “Atenção!”
  - ⬇ “Você já encontrou?”
- Palavra
  - ⬇ “Perigo!”
  - ⬇ “coelho”

II. Unidades mínimas *distintivas de formação*

- do som (*fonemas*) ⬇ lata / bata / cata
- da forma (*morfemas*) ⬇ relacionar / relacionamento

N

I

L

C

São Carlos SP

Sentenças

* Unidade escolhida para o processamento lingüístico

◆ Lidam com *relações*

◆ Podem ser segmentadas em sintagmas

Sintagmas

◆ Grupo de palavras organizado em torno de um *núcleo sintático*

- ⬇ NP (SN) ⇒ nome ⇒ ex.: “A mãe” / “presentes”
- ⬇ VP (SV) ⇒ verbo ⇒ ex.: “tinha comprado”
- ⬇ PP (SP) ⇒ preposição ⇒ “para os filhos”
- ⬇ AP (SAdv) ⇒ advérbio ⇒ “na galeria”

N

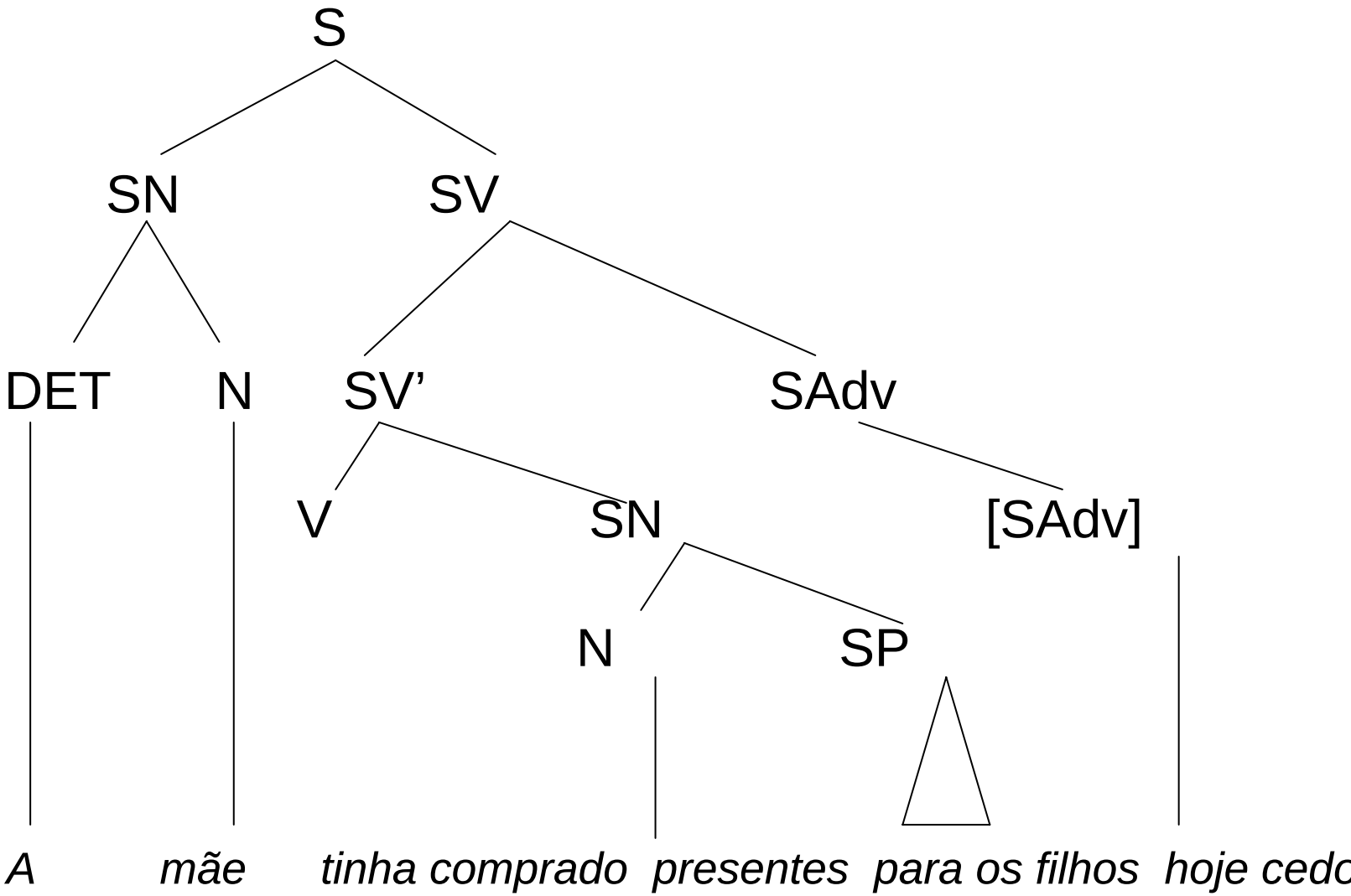
I

L

C

São Carlos
SP

Análise componencial ⇒ *estruturas*



N

I

L

C

São Carlos
SP

Níveis de processamento

- Modelo de processamento modular 🡻 módulos
- *Módulos lingüístico*: cada etapa do processamento da língua
- Léxico: um módulo lingüístico?

Os módulos lingüísticos

- *Fonético-fonológico*: quando se trata de depreender a identidade sonora dos elementos que constituem a palavra.
- *Morfológico*: quando os morfemas são isolados para a compreensão do processo de *formação* e *flexão* das palavras
- *Sintático*: quando a distribuição das palavras resulta em determinadas funções que elas desempenham na sentença
- *Semântico*: quando o conteúdo significativo da palavra implica relações de natureza ontológica e referencial para a identificação dos objetos no mundo
- *Pragmático-discursivo*: quando a significação implica relações com contexto de enunciação e intenção dos interlocutores.

N

I

L

C

São Carlos
SP

Língua

◆ “Sistema simbólico em que um conteúdo mental - o *significado* - se integra numa dada expressão oral, que podemos chamar o *significante*” (Saussure, 1922)

Informações lingüísticas

I. Informações fonético-fonológicas

◆ Fonemas: entidade representacional das formas distintivas do som

Ex.: - *pato* ➧ /p/ : consoante labial surda

- *bato* ➧ /b/ : consoante labial sonora

- *morte* ➧ /o/ : vogal central média

- *morto* ➧ /o/ : vogal posterior média-alta

* Fonema ≠ Letra ➧ grafema (ex.: mexer / fechar / assim / açaí

N

I

L

C

São Carlos
SP

- Alguns problemas para o PLN: - variação de timbre
  - posição silábica dos sons
  - palavras homófonas

II. Informações morfológicas

- *Morfemas*: unidades mínimas dotadas de significado

Exs.: - sonh- ⬇ raiz léxica
- -a- ⬇ vogal temática
- -va- ⬇ desinência modo-temporal
- -m ⬇ desinência número-pessoal

⬇ “SONHAVAM”

N

I

L

C

São Carlos
SP

◆ Tipos de morfemas:

1. Gramaticais: ex.: desinências, vogais temáticas
2. Lexicais: ex.: afixos (prefixo, infixo, sufixo)

Ex.:

1. *florid<u>as</u>* ⬇ -*a*-: morfema gramatical indicativo de gênero
   -s: morfema gramatical indicativo de número
2. *flor<u>ista</u>* ⬇ -*ista*: morfema lexical indicativo de ocupação/emprego

◆ Processos morfológicos: - flexão
- derivação
- composição

◆ Alguns problemas para o PNL: - palavras homônimas
- derivação/neologismos

N

I

L

C

São Carlos
SP

II. Informações sintáticas

◆ *Sentença*: uma unidade mínima da comunicação.

Exs.: 1. Fernando caiu da bicicleta.

2. Pare!

3. Que férias!

* Sentença = Oração = Frase

* Sentença ≠ reunião de vocábulos / ≠ sentido completo

◆ Sintaxe: Posição/ordem/distribuição dos vocábulos

Exs.: - [*saia menina de] / menina de saia

- [*queria O rapaz emprego de mudar]

⬇ Categorias gramaticais (nome, adjetivo, etc.)

⬇ Funções sintáticas (exs.: sujeito, complemento nominal)

⬇ *Regras* de boa formação sintática

N

I

L

C

São Carlos
SP

◆ Outras noções importantes:

1. Estrutura argumental

Exs.: - *falar* 🡻 2 argumentos — sobre alguma coisa
com alguém
em algum lugar

- *falar* 🡻 1 argumento

2. Papel temático

Ex.: *A menina <u>destruiu</u> o brinquedo com o martelo.*

- *a menina*: papel temático 🡻 agente
- *o brinquedo*: papel temático 🡻 paciente
- *com o martelo*: papel temático 🡻 instrumento

N

I

L

C

São Carlos SP

◆ Alguns problemas para o PLN:

- ambigüidade sintática (Ele atirou *a pedra da ponte*)

- determinação dos papéis temáticos

IV. Informações semânticas

◆ *Significado*: entidade abstrata responsável pela articulação de idéias a expressões lingüísticas.

* Propriedades semânticas

⬇ Significado ≠ Sentido ≠ Acepção de...

◆ *Primitivos semânticos*:

1. Traços semânticos (ex.: animado, concreto, humano)

2. Categorias ontológicas (ex.: evento, propriedade, entidade)

N

I

L

C

São Carlos
SP

◆ Especificação semântica: Exs.:

A

1. *homem* {entidade, animado, concreto, humano, macho} ⬇ “homem”

2. *mulher* {entidade, animado, concreto, humano, fêmea} ⬇ “mulher”

B

1. *banco* {coisa, concreto [feito para sentar]} ⬇ “assento”

2. *banco* {lugar, concreto,[envolve dinheiro]} ⬇ “instituição financeira”

◆ Alguns problemas para o PLN:

- ambigüidade lexical

Ex.: *cabo* / *canto* / *manga*)

- contradições

Ex.: *Eu sei que ela saiu, mas não está em casa*.

N

I

L

C

São Carlos
SP

V. Informações pragmático-discursivas

- Texto: unidade lingüística maior na atividade comunicativa
- Discurso: atividade lingüística [língua] *atualizada*:

  - em determinado *tempo* ⎫ contexto
  - em um *lugar* preciso ⎭
  - prevê um *locutor* determinado ⎫ interlocutores
  - prevê um *ouvinte* suposto ⎭

- *Pragmática* ⬇ leva em conta: - o *contexto* da enunciação e
  - a *intenção* dos interlocutores
- *Análise do discurso* ⬇ ocupa-se: - da *formação discursiva*
  - das *condições de produção*

N

I

L

C

São Carlos
SP

◆ Noções trabalhadas:

1. *Pressuposição*
2. *Inferência*

Coerência

3. *Implicação*

Coesão

⇩

conectivos (elementos de coesão)

⬇ Coerência: fenômeno que reflete a harmonia entre as partes do texto

* solidariedade, lógica

⬇ Coesão: conexão entre os partes do texto evidenciando as relações de sentido nele existentes.

* articulação, concatenação

⬇ Conectivos: preposições, conjunções, pronomes, advérbios

Problemas típicos

Interpretação

Extração de uma representação conceitual da *mensagem de um texto*

João dormiu no cinema.

existe(humano:**x**,valor(**x**,'João'),
evento(**e**,dormir(**x**,local(**x**,cinema))),
tempo(**e**,passado))

Geração

Produção de uma *representação coerente do discurso*, pelo inter-relacionamento de suas *proposições elementares*;

Expressão lingüística da *estrutura coerente do discurso*

Problemas lingüísticos

Anáfora

*Maria comprou um ingresso para o filme e deu-**o** a seu irmão.*

Ambigüidade

João viu um homem no parque com um binóculo.

Problemas lingüísticos

Representações múltiplas

Paráfrases

João aconselhou Maria a beber vinho.
João disse a Maria que beber vinho lhe faria bem.

Metáforas

Ele está mais para tartaruga do que para lebre.

Extraposição (referências encadeadas)

O cabelo da menina que sentou no banco que foi pintado pelo servente ficou manchado pela tinta que estava fresca.

Problemas de representação

Conhecimento lingüístico

Morfológico, lexical, sintático, semântico

Conhecimento extralingüístico

Interação com o "mundo" e a linguagem (pragmática)

Modelos de representação do conhecimento

Redes semânticas, ontologias diversas, dicionários

Armazenamento

- Inter-relacionamento entre os diversos tipos de informação
- Limitações computacionais

Problemas de comunicação

Intenções e inferências
Recepção e produção de mensagens

Composição da mensagem
Escolha do conteúdo informacional

Estruturação da mensagem
Escolha da organização do conteúdo

Verbalização
Escolha de construções da língua em uso

Recepção da mensagem
Avaliação do comportamento do receptor

N

I

L

C

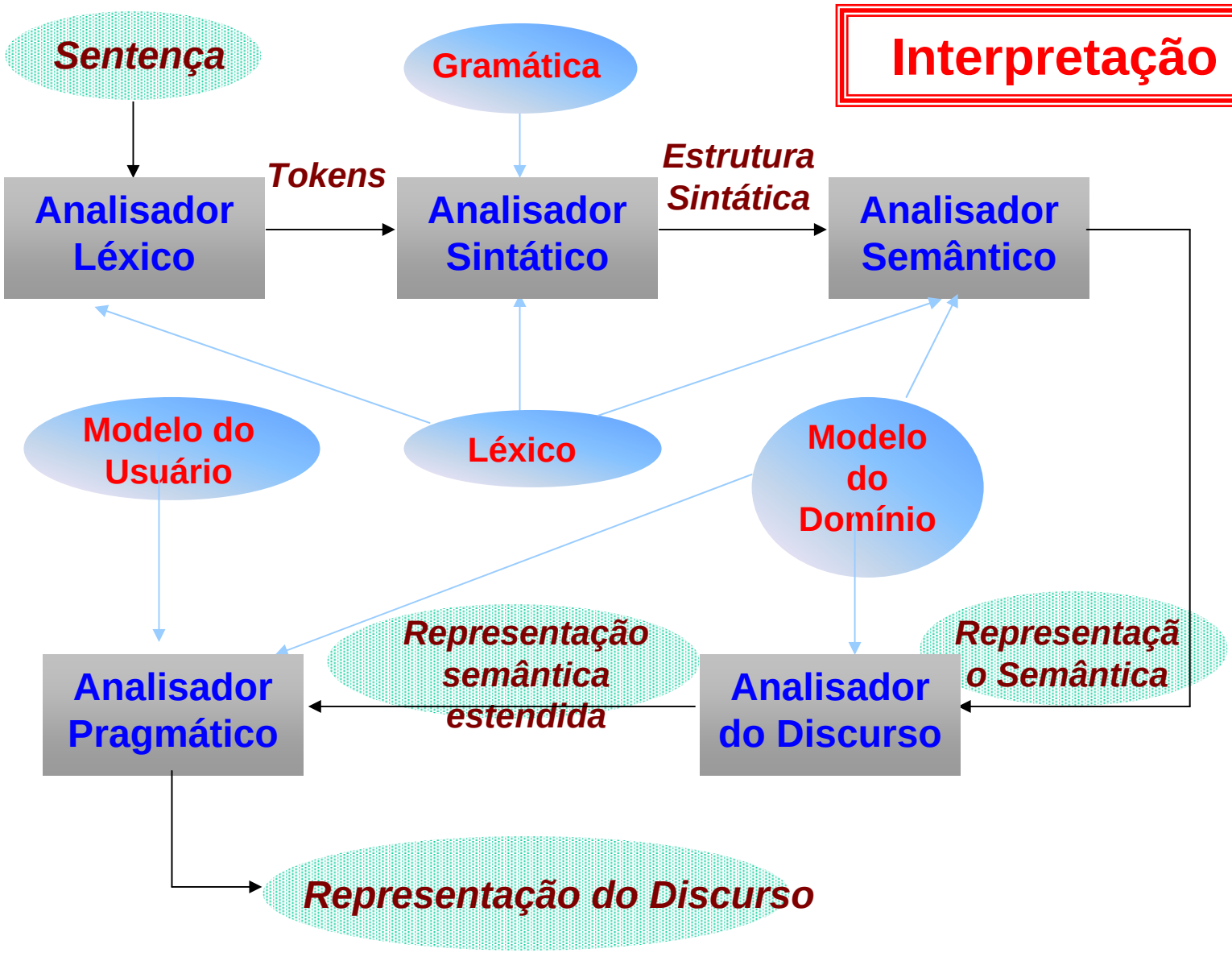
São Carlos
SP

Lingüística Computacional

Modelagem e formalização
de
aspectos lingüísticos e comunicativos

⇕

Engenharia do conhecimento lingüístico


Interpretação
Sentença
Gramática
Tokens
Estrutura Sintática
Analisador Léxico
Analisador Sintático
Analisador Semântico
Modelo do Usuário
Léxico
Modelo do Domínio
Representação semântica estendida
Representaçã o Semântica
Analisador Pragmático
Analisador do Discurso
Representação do Discurso

Exemplo 1

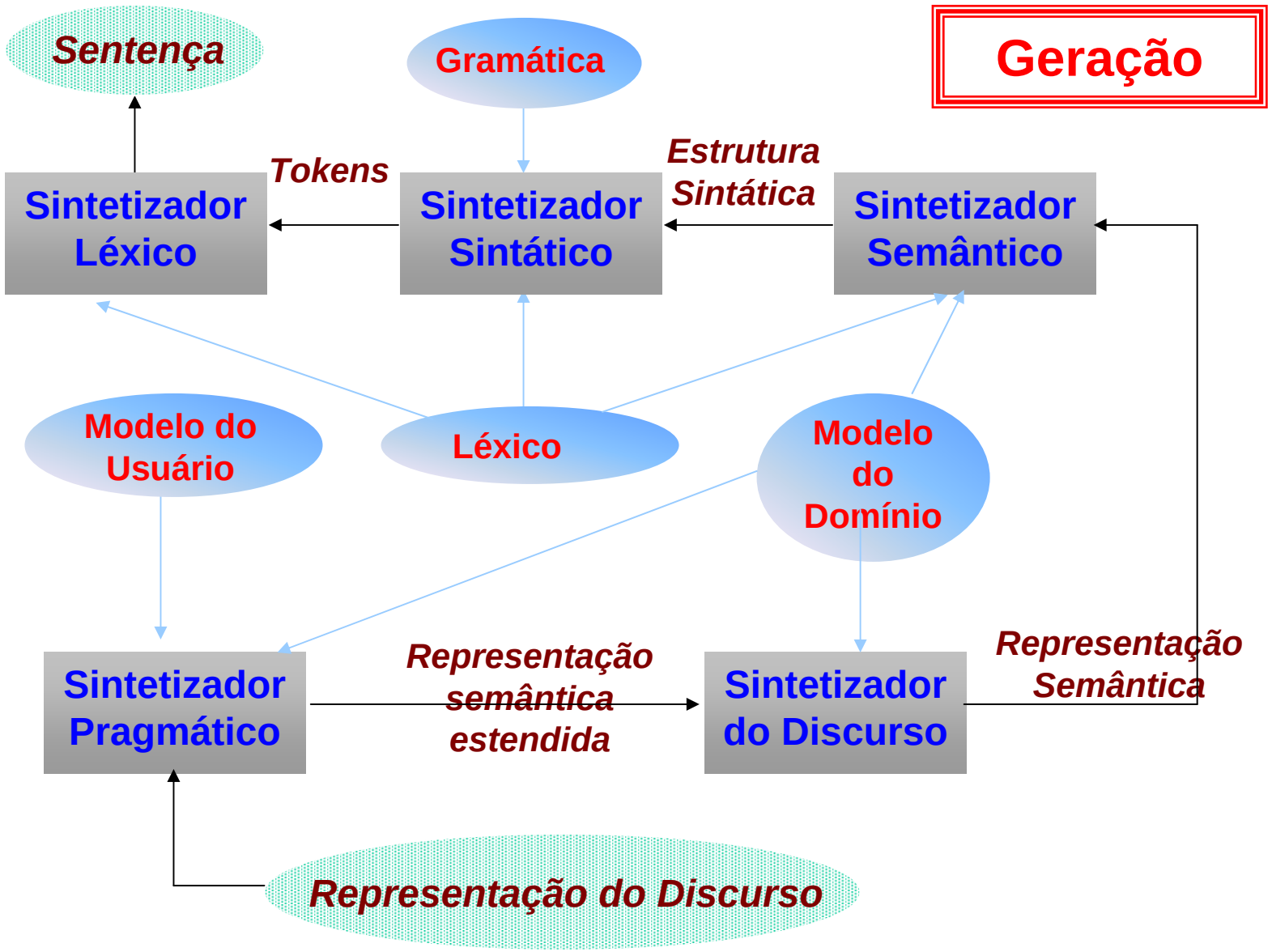
Interpretação

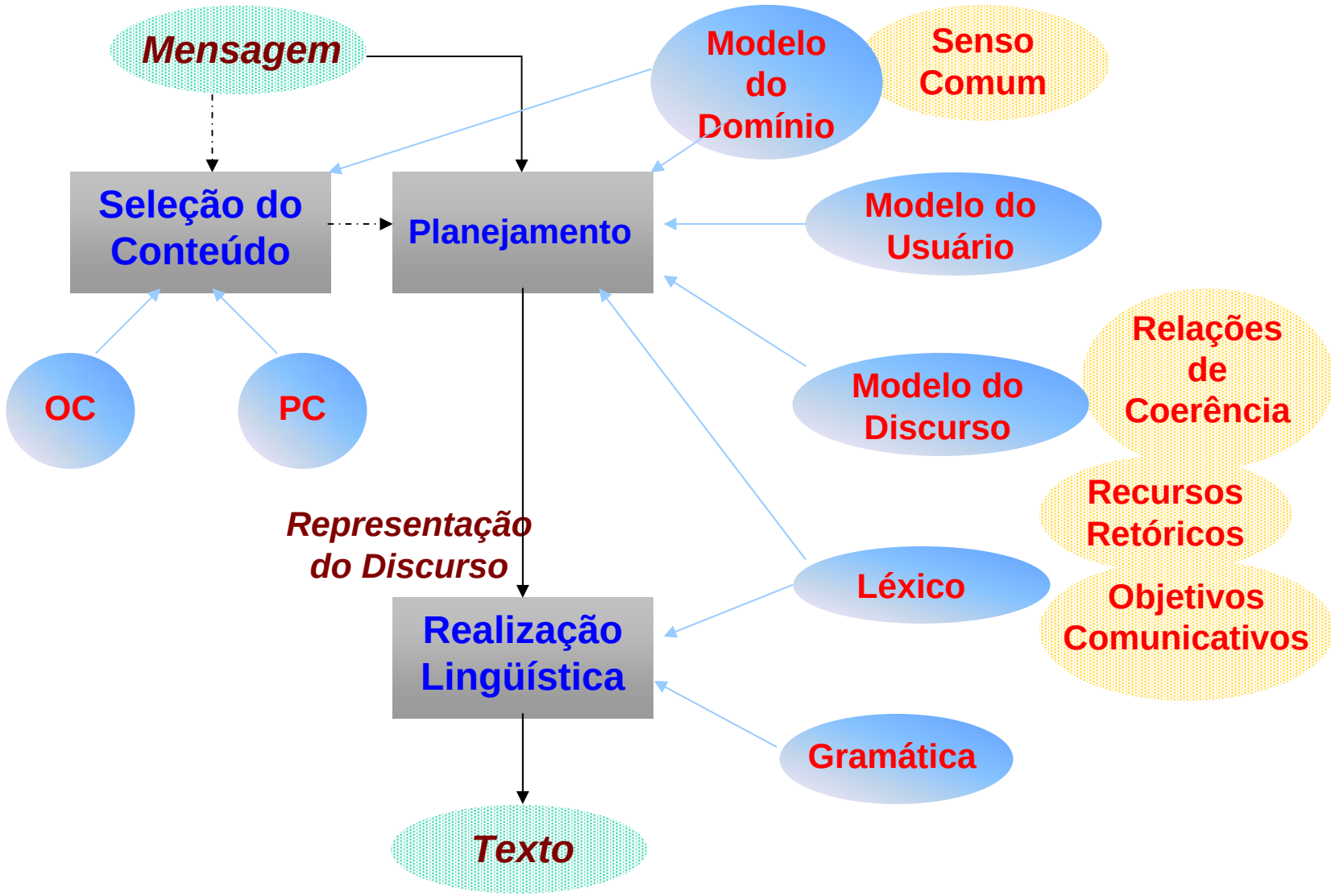
João *viu* ***um homem*** *no parque* ***com um binóculo****.*

João *viu um homem no parque* ***com um binóculo****.*

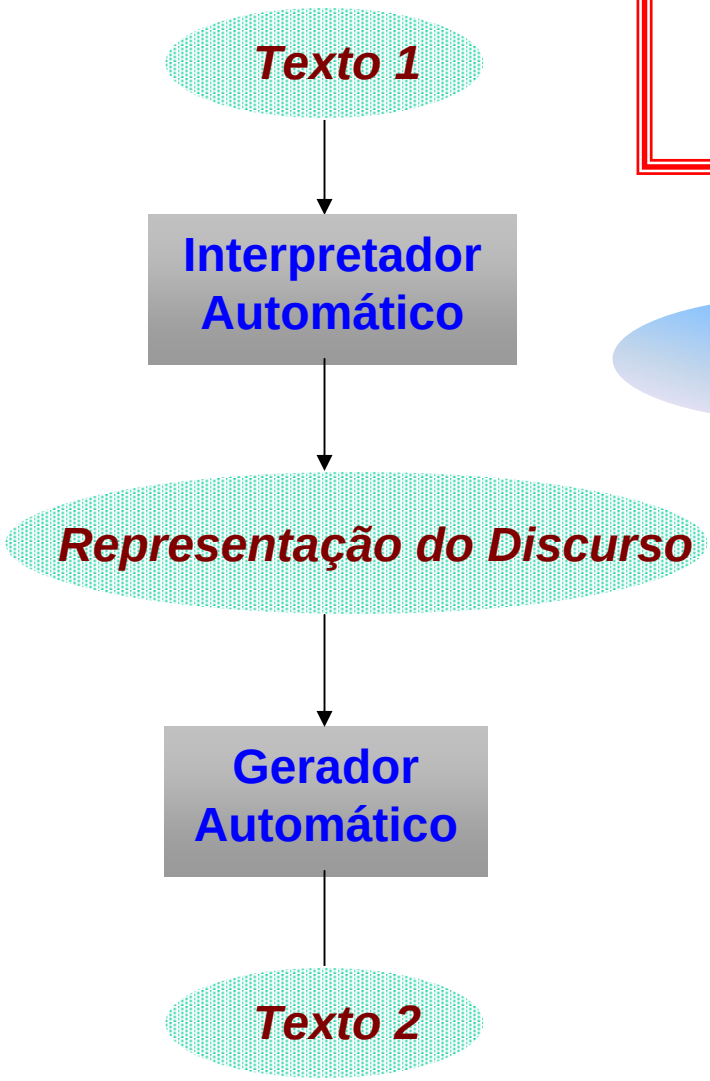
João viu ***um homem*** *no parque* ***com um binóculo****.*

João viu um homem no parque com um binóculo.
O homem, após enterrar algo no terreno, saiu correndo e escondeu-se atrás de uma árvore.
Mais tarde, ao ouvir o noticiário, João relembrou a cena e concluiu que seu binóculo o fez uma testemunha ocular de um crime.


Geração
Sentença
Gramática
Tokens
Estrutura Sintática
Sintetizador Léxico
Sintetizador Sintático
Sintetizador Semântico
Modelo do Usuário
Léxico
Modelo do Domínio
Sintetizador Pragmático
Representação semântica estendida
Sintetizador do Discurso
Representação Semântica
Representação do Discurso


Mensagem
Modelo do Domínio
Senso Comum
Seleção do Conteúdo
Planejamento
Modelo do Usuário
OC
PC
Modelo do Discurso
Relações de Coerência
Recursos Retóricos
Representação do Discurso
Léxico
Objetivos Comunicativos
Realização Lingüística
Gramática
Texto

Interpretação e Geração



Modelo do Usuário

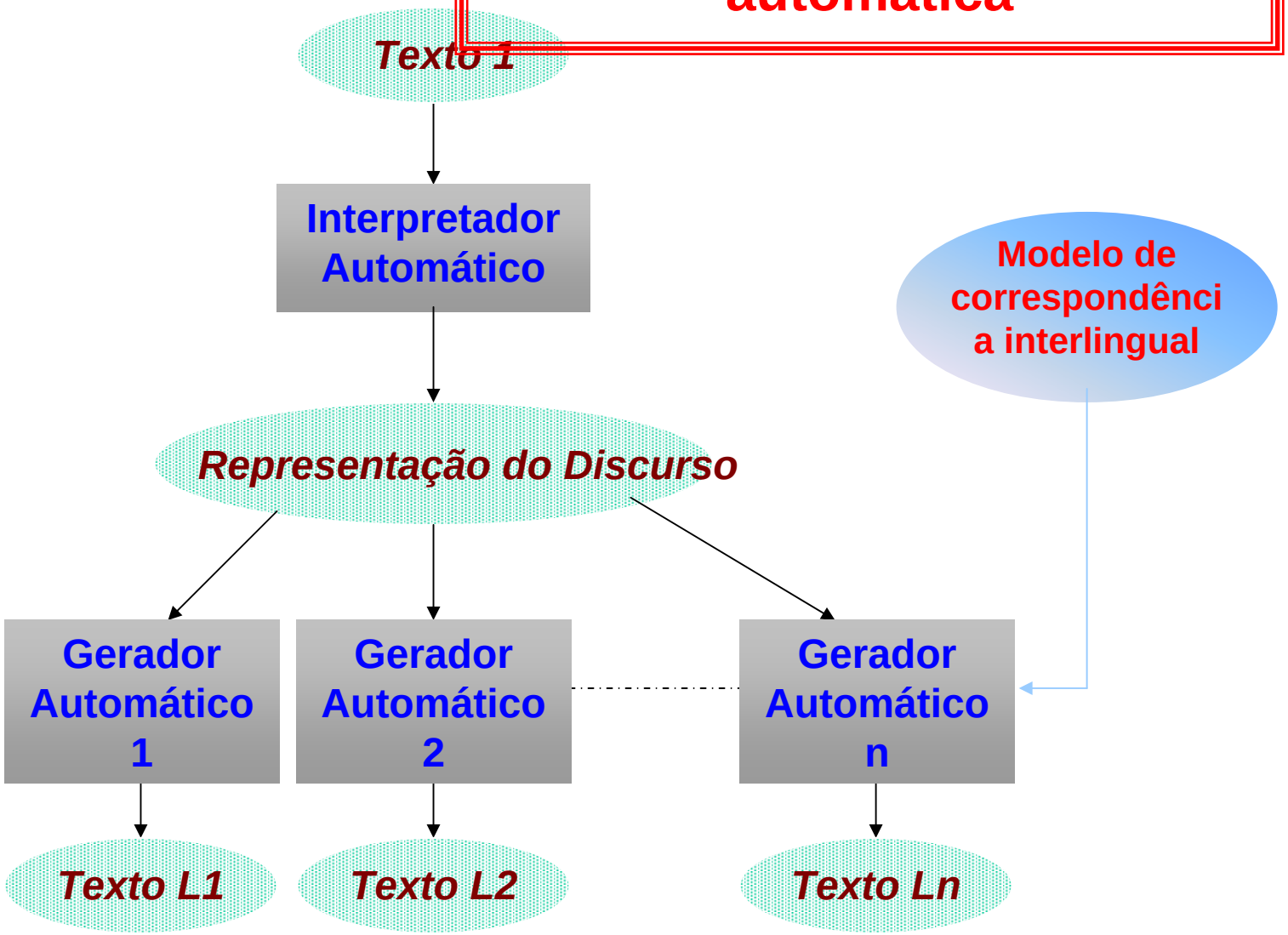
Modelo do Discurso

Gramática

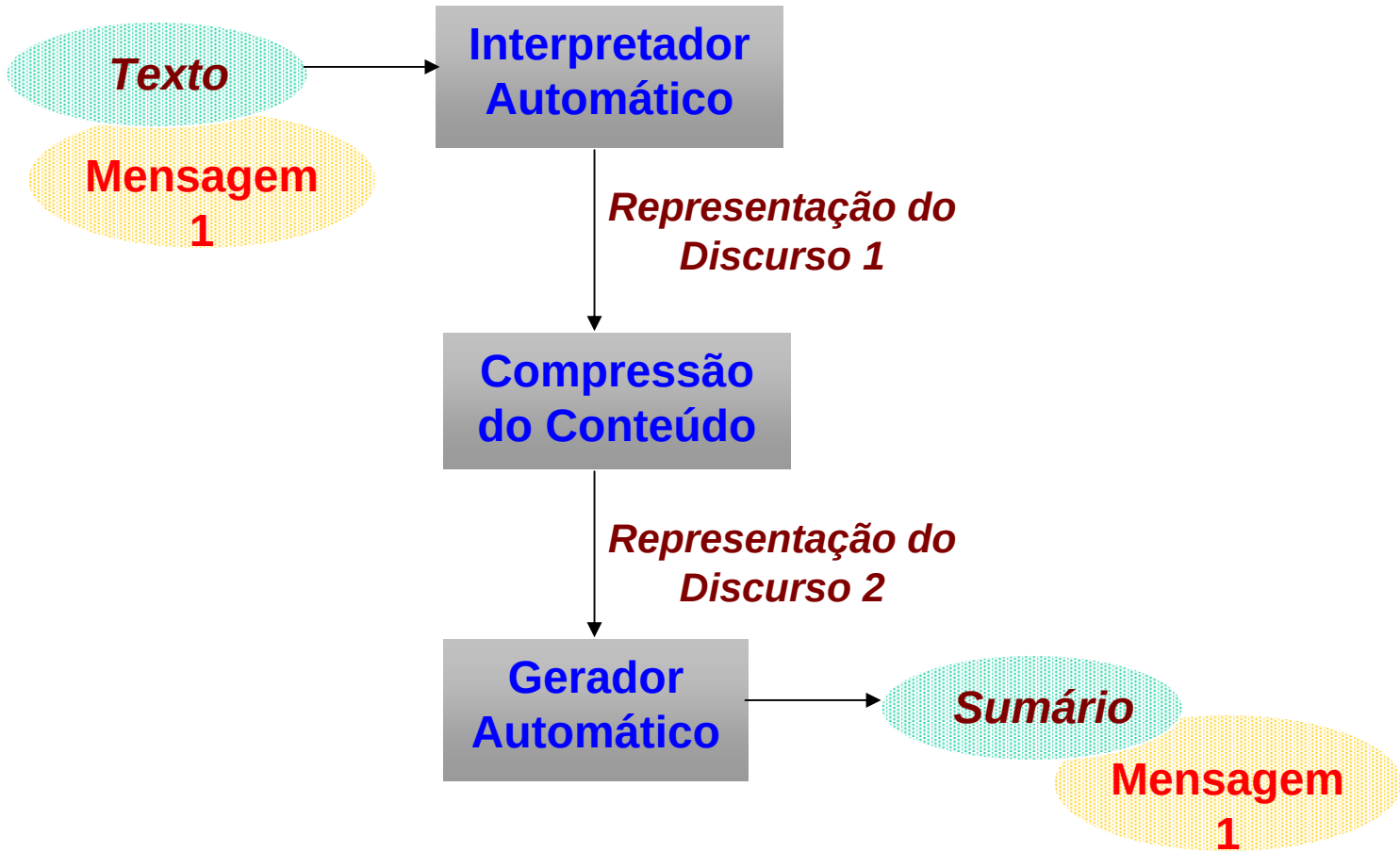
Léxico

Modelo do Domínio

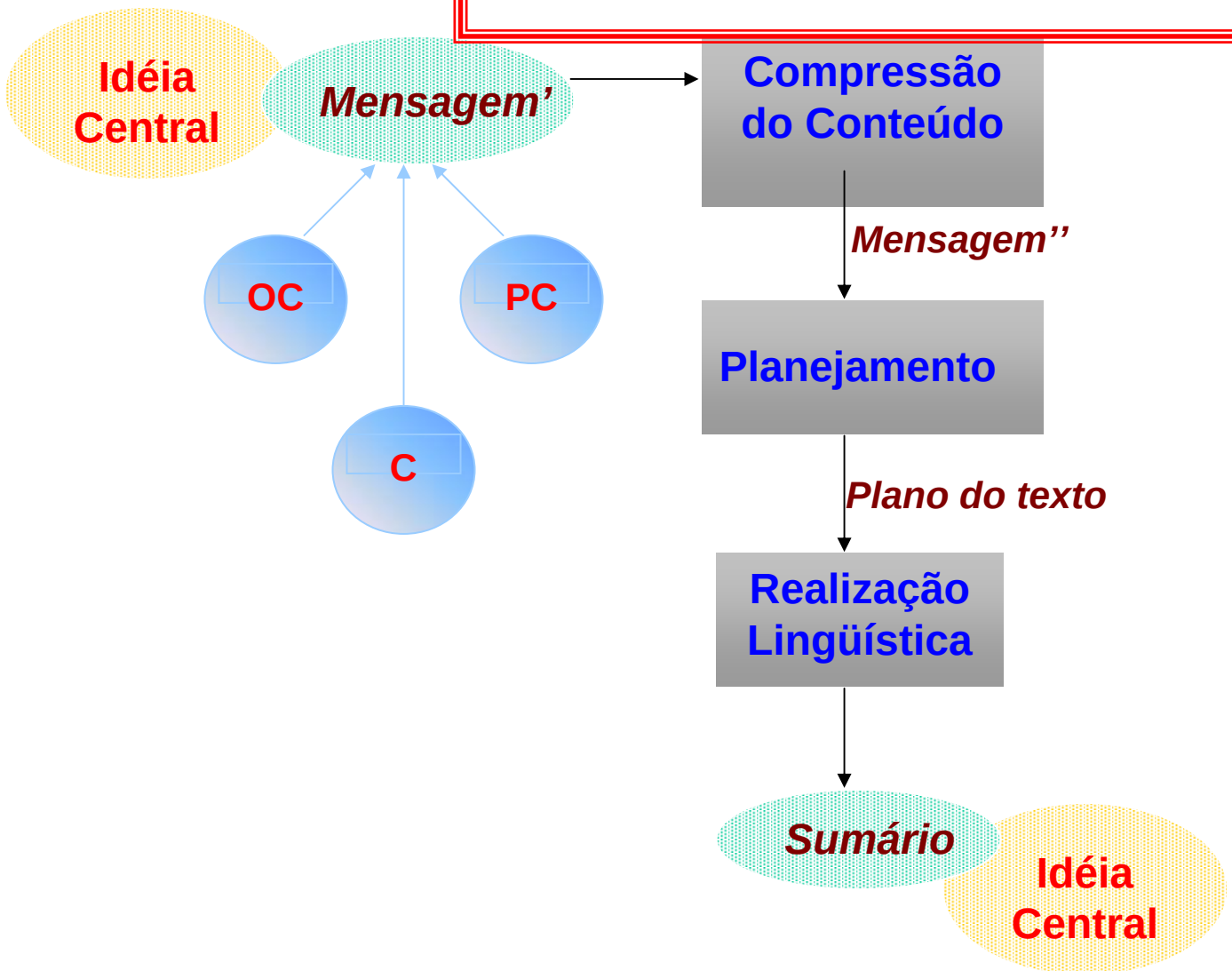
Cenário 1: Tradução automática



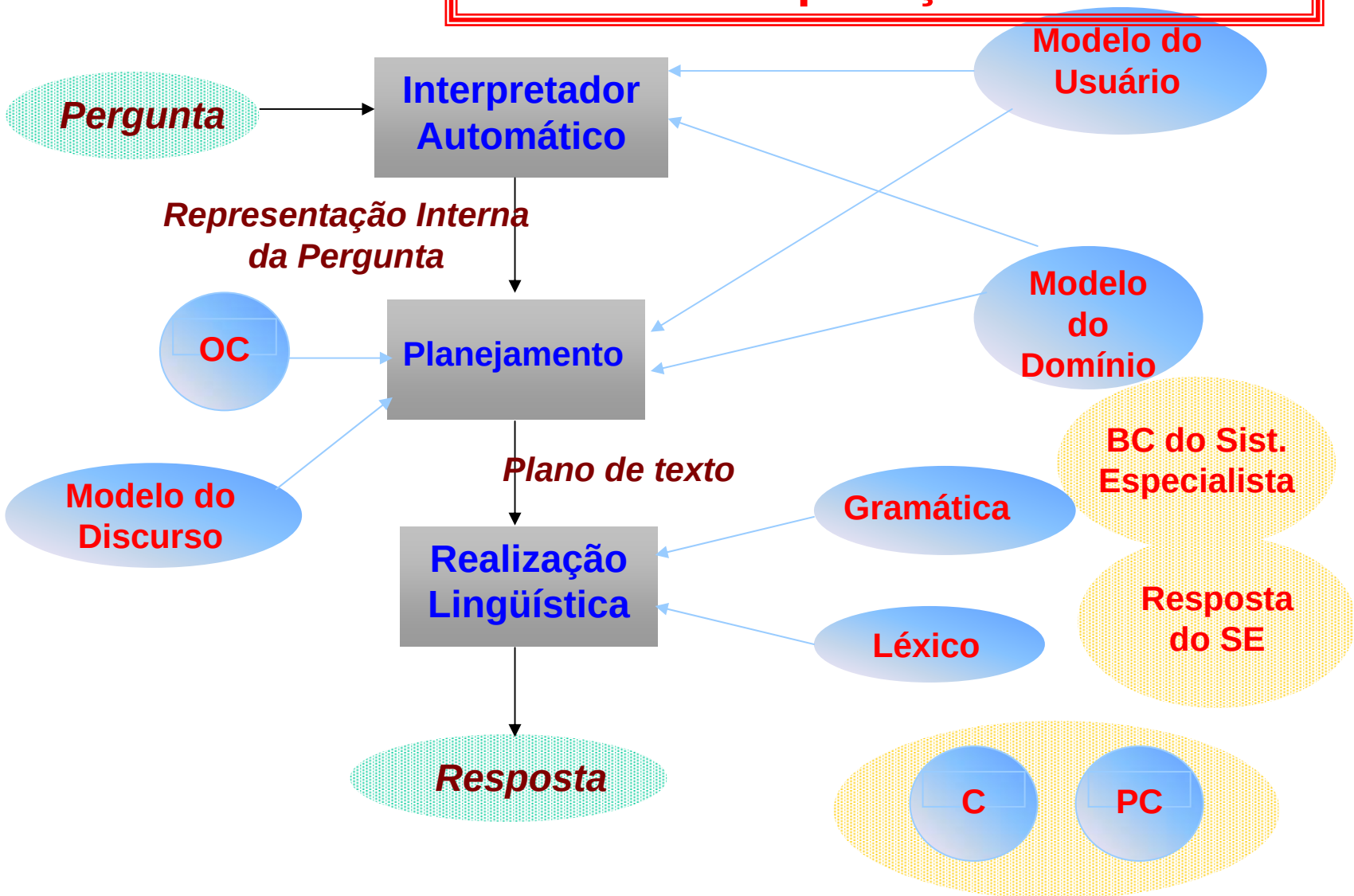
Cenário 2': Sumarização automática



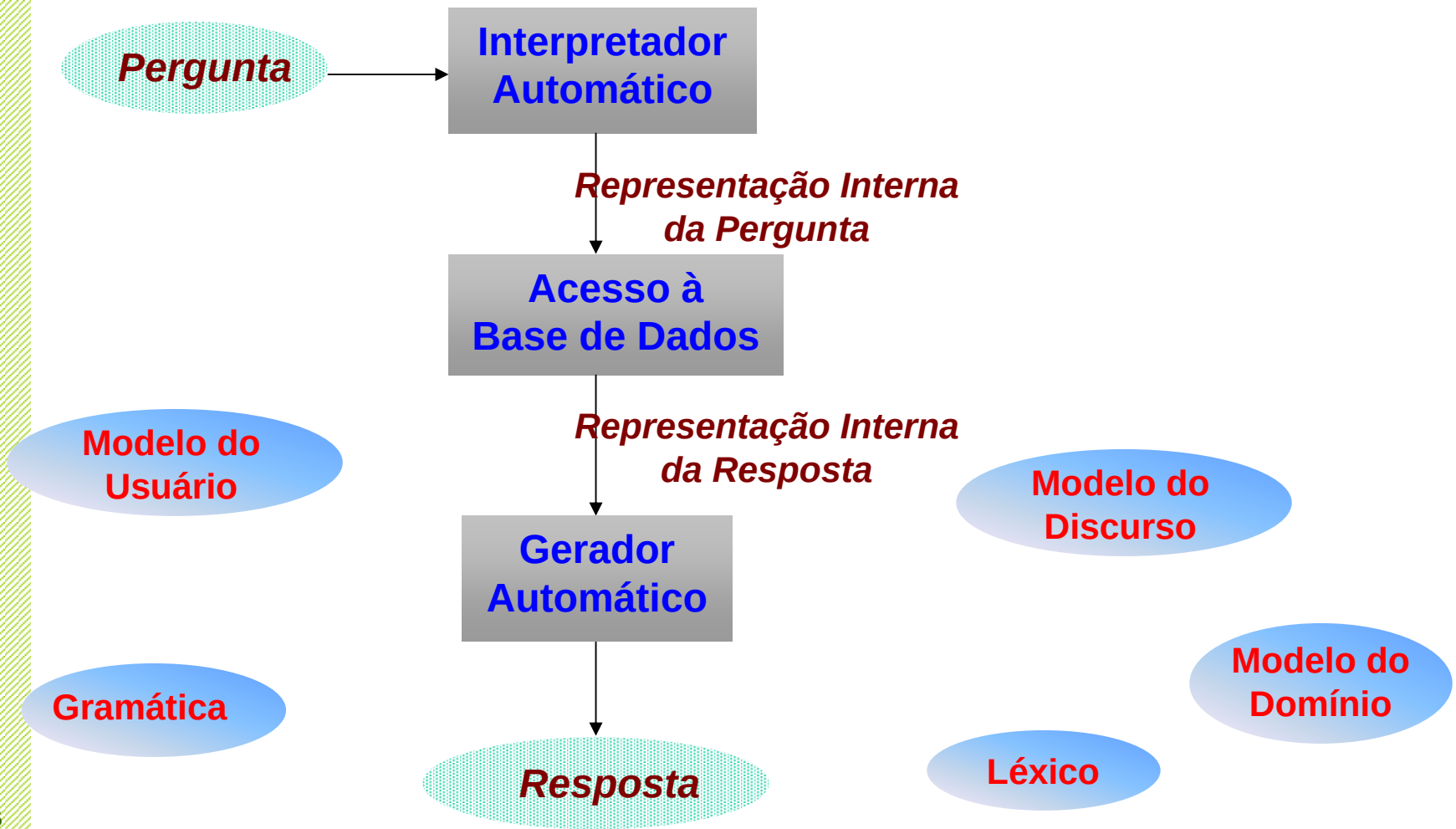
Cenário 2'': Sumarização automática



Cenário 3: Gerador de explicações



Cenário 4: Consulta a BD



Interpretação

- Distinção entre informações essenciais, complementares e supérfluas, para expressão da mensagem original

- Preservação da idéia central
  - Proposição central do discurso
  - Objetivo comunicativo

- Compreensão
  - Necessária?
  - Quando?
  - Por que?
  - Para que?

- Outros processos automáticos
  - Estatística
  - Redes neurais
  - Etc.

Problemas

- Distinção entre informações essenciais, complementares e supérfluas
- Seleção de informações complementares visando
  - Clareza
  - Informatividade
  - Expressividade
  - Legibilidade
- Exclusão de informações supérfluas
- Como reconhecer/Quando utilizar informações complementares
- Quando considerar o modelo do usuário?
- Quando considerar o gênero do discurso?
- Geração multi-sentencial, sentença por sentença
  - Impossibilidade de se recuperar a “trama do discurso”
  - Construções estilicamente pobres
  - Referências provavelmente incorretas

Engenharia do conhecimento lingüístico

⇓

Métodos computacionais que reflitam o potencial lingüístico requerido

Conhecimento necessário especificado nos moldes requeridos pelos métodos e pela aplicação

Sistema eficiente e competente, mediante

- objetivos principais de uso
- escolha do ambiente computacional adequado

N

I

L

C

São Carlos
SP

Agradecimentos

Alexsandro dos Santos

Jorge Marques Pelizzoni
(organizadores)