

ONTOLP: CONSTRUÇÃO SEMI-AUTOMÁTICA DE ONTOLOGIAS A PARTIR DE TEXTOS DA LÍNGUA PORTUGUESA

Mestrando: Luiz Carlos

Orientadora: Renata Vieira



SUMÁRIO

- Introdução
- Objetivo
- Metodologia
- OntoLP
- Avaliação
- Recursos
- Considerações Finais



INTRODUÇÃO (MOTIVAÇÃO)

○ Web Semântica

- *“Para o funcionamento da Web Semântica, computadores devem ter acesso a coleções estruturadas de informação e conjuntos de regras que eles possam usar para conduzir raciocínio automático”* (Berners-Lee, 2001)

○ Construção Manual de Ontologias

[Brewster et al.,2003]

- Complexo
- Tedioso
- De alto custo



INTRODUÇÃO (OBJETIVO)

- Propor e avaliar técnicas para a construção automática de ontologias a partir de textos da língua portuguesa com base em técnicas já desenvolvidas para outras línguas



INTRODUÇÃO (CONSTRUÇÃO AUTOMÁTICA DE ONTOLOGIAS)

- Pontos Chave:
 - Etapas Executadas
 - Extração de Termos
 - Organização Hierárquica dos Termos
 - Fonte de Conhecimento
 - Textos



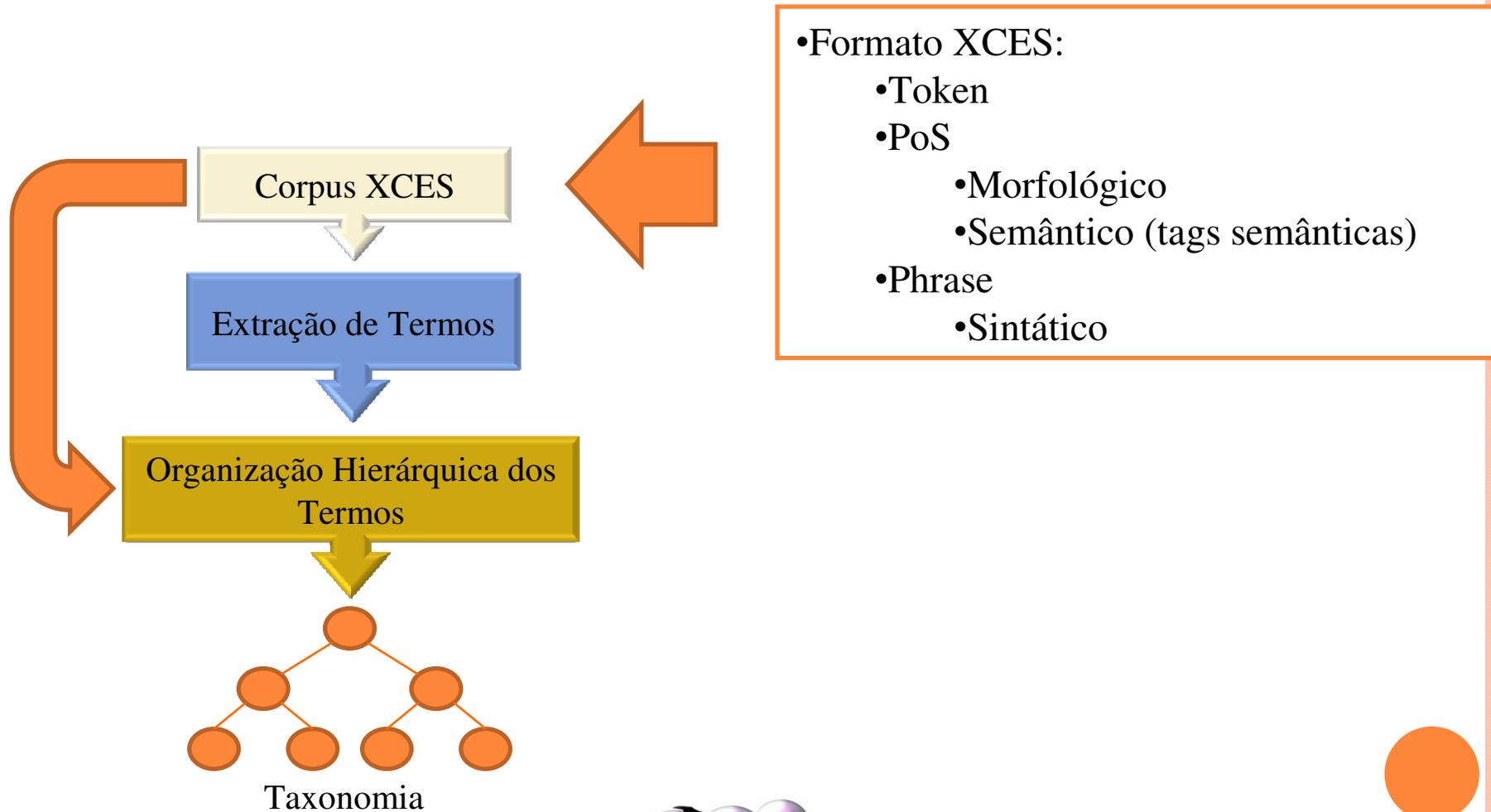
INTRODUÇÃO

(CONSTRUÇÃO A PARTIR TEXTOS)

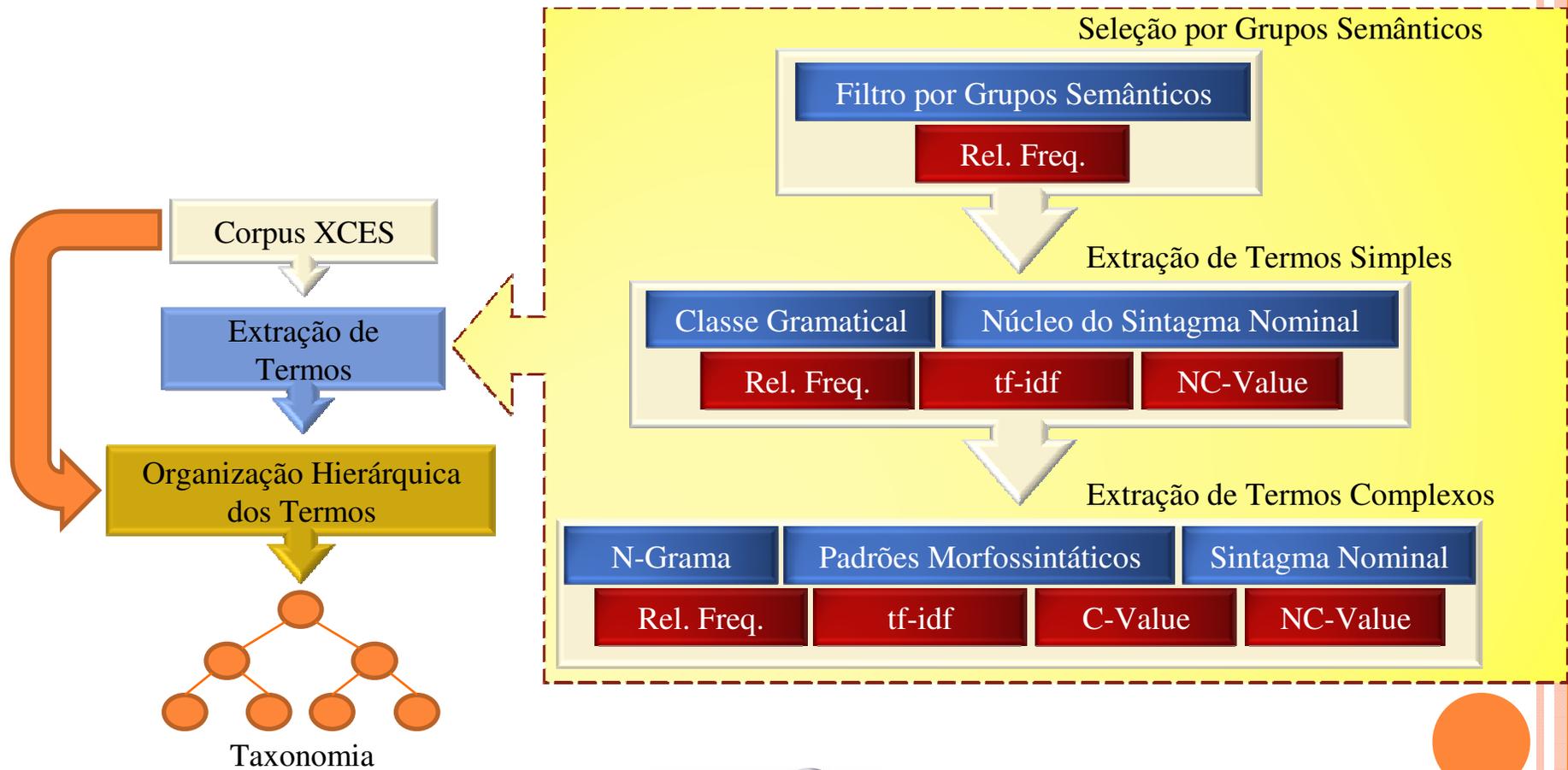
- Principais Métodos (Extração de Termos):
 - Estatísticos
 - Lingüísticos
 - Híbridos
- Principais Métodos (Organização Hierárquica):
 - Baseado em Termos Complexos
 - Baseado em Padrões
 - Baseado nas medidas de Especificidade e Similaridade



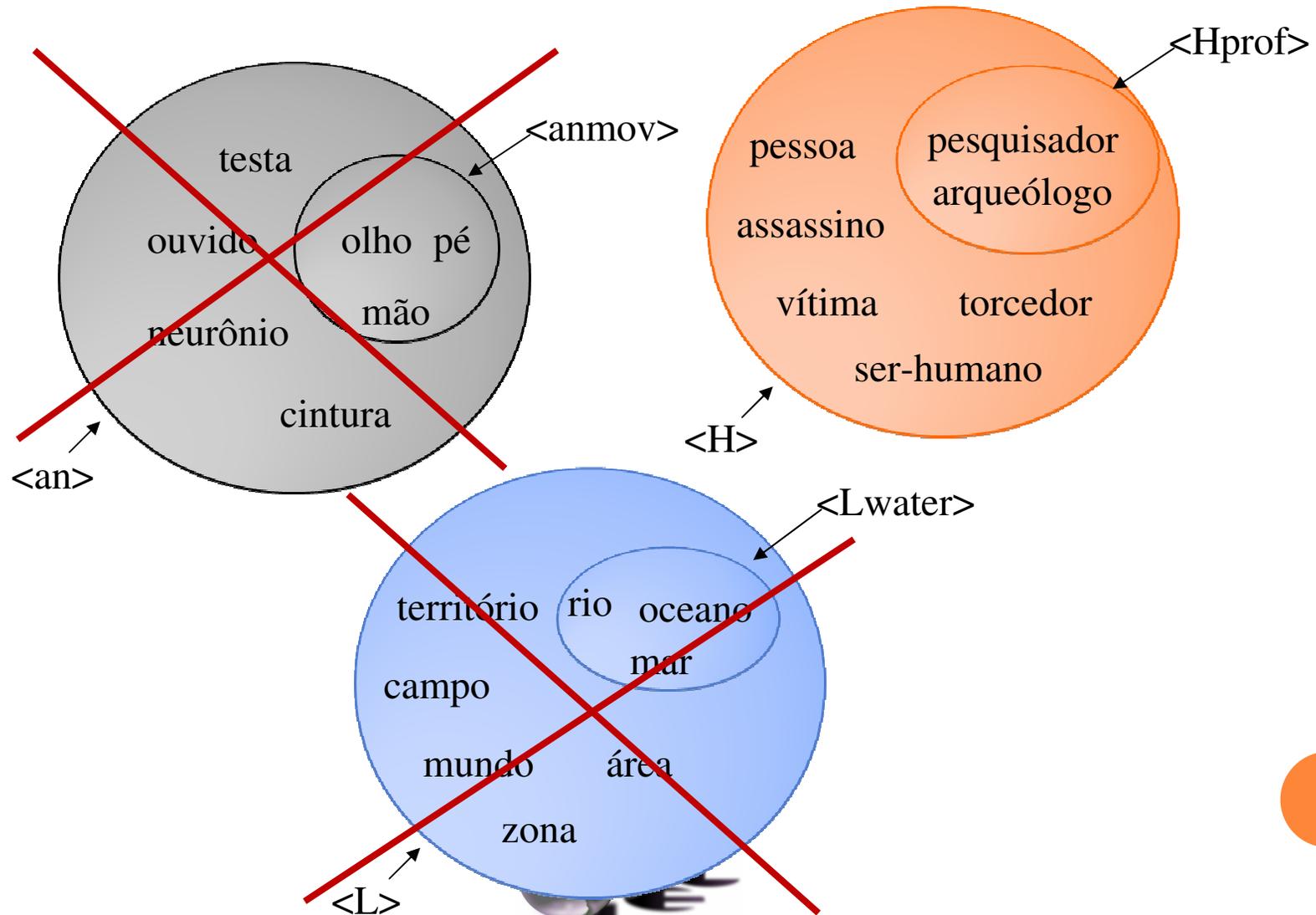
METODOLOGIA PROPOSTA



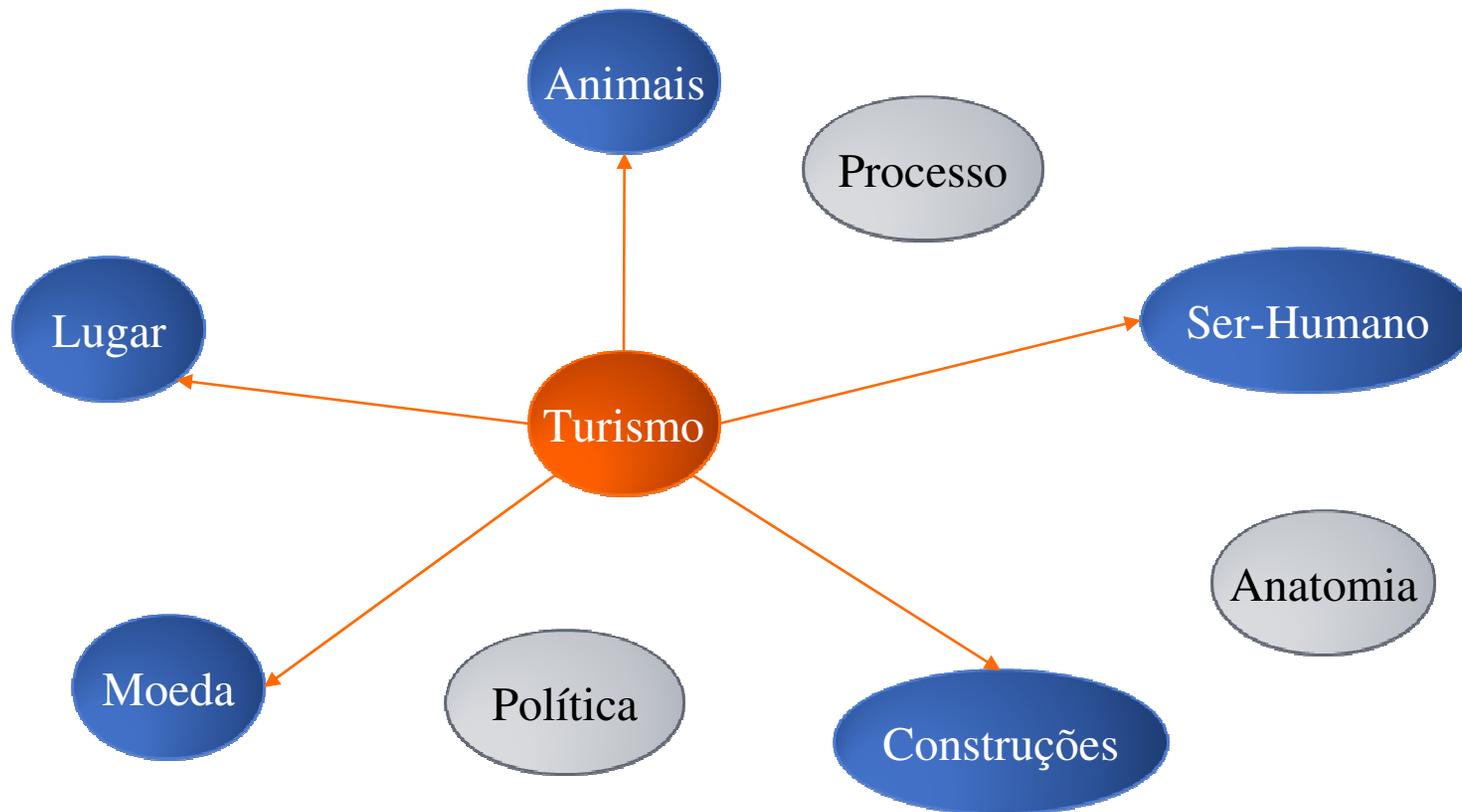
METODOLOGIA (EXTRAÇÃO DE TERMOS)



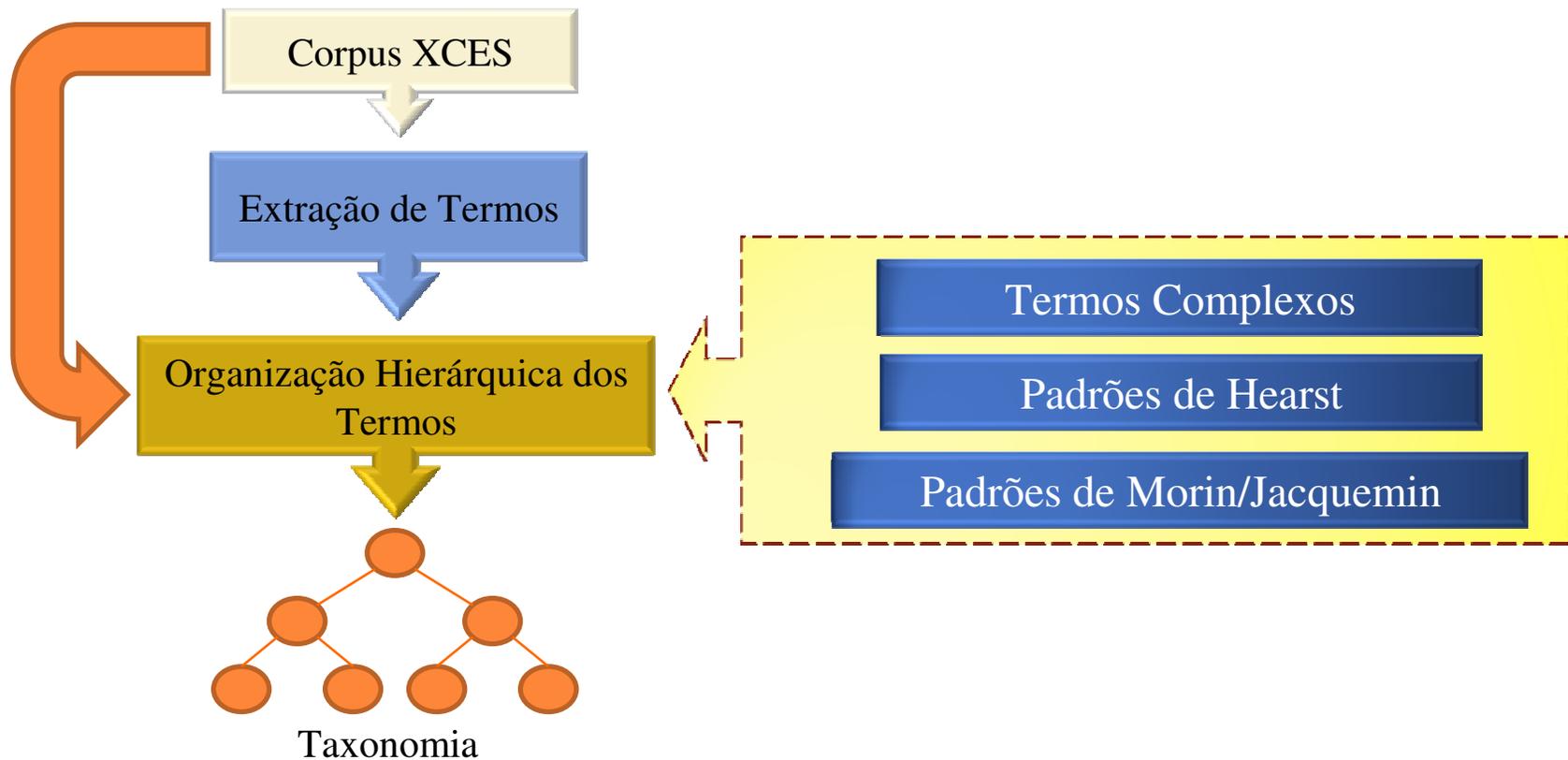
METODOLOGIA (GRUPOS SEMÂNTICOS)



METODOLOGIA (GRUPOS SEMÂNTICOS)



METODOLOGIA (ORGANIZAÇÃO HIERÁRQUICA DOS TERMOS)



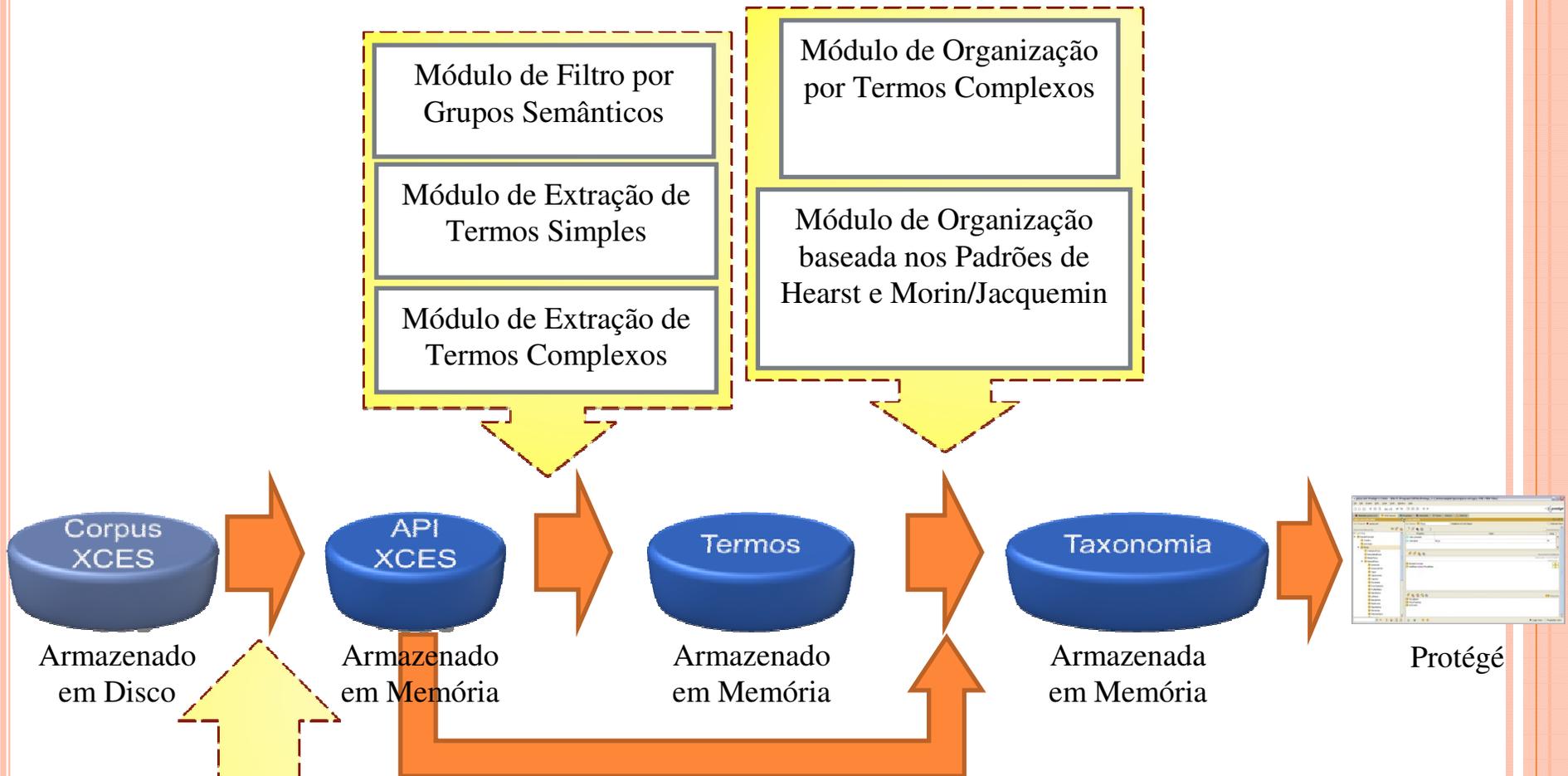
PADRÕES DE HEARST E MORIN/JACQUEMIN

Hearst/Baségio	Baségio/Termos Complexos
SUB como {(SUB ,) * (oule)} SUB	T como {(T ,) * (oule)} T
SUB tal(is) como {(SUB ,) * (oule)} SUB	T tal(is) como {(T ,) * (oule)} T
tal(is) SUB como {(SUB ,) * (oule)} SUB	tal(is) T como {(T ,) * (oule)} T

Morin/Baségio	Baségio/Termos Complexos
SUB1 (LIST_ SUB2)	T1 (LIST_ T2)
SUB1 : LIST_ SUB2	T1 : LIST_ T2
SUB1 e (notadamente - em particular) SUB2	T1 e (notadamente - em particular) T2



ONTOLP



Módulo de Importação do Corpus (API XCES)

Módulo de Filtragem por Grupos Semânticos
Módulo de Extração de Termos Simples
Módulo de Extração de Termos Complexos

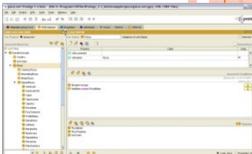
Módulo de Organização por Termos Complexos
Módulo de Organização baseada nos Padrões de Hearst e Morin/Jacquemin

Corpus XCES

API XCES

Termos

Taxonomia



Protégé



ONTOLP (INTERFACE DE EXTRAÇÃO DE TERMOS)

<new> Protégé 3.3.1

File Edit Project Window Help

Classes Slots Forms Instances OntoLP (lastVersion)

Corpus Termos Taxonomia

Exatção de Termos:

Informações Semânticas

Filtro Semântico (10:43:33)

Termos Simples:

Classe Gramatical (10:43:59)
Núcleo do Sintagma Nominal (10:44:31)

Termos Compostos:

N-Grama (10:44:20)
Padrões Morfossintáticos (10:44:20)
Sintagma Nominal (10:44:20)

Lista de Termos: Total de tags: 131

Tag Semântica	Descrição	Freq. Rel.
[ac]	Abstrato e Contável	0,0685789
[act]	Ação	0,0534550
[Hprof]	Profissão Humana	0,0435463
[HH]	Grupo de Humanos	0,0346806
[am]	Massa Abstrata/Não Contável	0,0333768
[H]	Humano	0,0320730
[event]	Evento não organizado	0,0320730
[sem-c]	Produto de Cognição	0,0315515
[Azo]	Animal Terrestre	0,0312907
[inst]	Instituição	0,0271186
[sem-r]	Trabalho de Leitura	0,0252934
[Labs]	Lugar Abstrato	0,0208605
[sick]	Doença	0,0190352
[act-d]	Realizar uma Ação	0,0169492
[occ]	Evento Social Humano	0,0166884
[amount]	Substantivo Quantitativo	0,0166884
[dur]	Duração	0,0164276
[cc]	Objeto Concreto e Contável	0,0156454
[per]	Período	0,0148631
[Ltop]	Lugar Natural, Geográfico	0,0148631
[activity]	Atividade	0,0148631
[domain]	Domínio de Conhecimento	0,0140808
[percep-f]	Sentimento	0,0138201
[tool]	Ferramenta	0,0135593
[Lstar]	Objetos Espaciais	0,0125163
[cm-chem]	Substância Química e Biológica	0,0125163
[pp]		0,0122555
[cm]	Massa Concreta/Não-Contável, substância	0,0122555
[temp]	Objeto Temporal, Ponto no Tempo	0,0119948
[Acell]	Células Animais	0,0119948
[mat]	Material, Substância	0,0106910
[Hnat]	Nacionalidade Humana	0,0104302

Ordenar termos por:

Freq. Rel. tf-idf C-Value NC-Value

Tag Cloud:

astrofísico astronauta astrônomo defesa fazendeiro escritor
estudante cronista senador dramaturgo ministro caçador-coletor
auxiliar autor médico assessor moderador gerente-executivo
médica modelo prefeito agrônomo presidente geocientista criado

secretária geneticista caçador pirata **cientista** co-autor
arqueólogo ministra secretário gerente técnico vice-diretor filósofo
paleoantropólogo animador secreta diretor-presidente
ambientalista mergulhador

pesquisador

químico ator físico autora geólogo biólogo paleontólogo
trabalhadora coordenador psicólogo especialista

ONTOLP (INTERFACE DE ORGANIZAÇÃO HIERÁRQUICA)

The screenshot displays the Protégé 3.3.1 software interface. The main window title is "<new> Protégé 3.3.1". The menu bar includes File, Edit, Project, Window, and Help. The toolbar contains various icons for file operations and editing. The interface is divided into several panes:

- Classes:** A tabbed interface with "Classes", "Slots", "Forms", and "Instances" selected. The current project is "OntoLP(Beta)".
- Corpus:** A section with tabs for "Termos" and "Taxonomia".
- Construção de Taxonomia:** A panel on the left with three sections:
 - Termo Composto:** Includes a "Termos Compostos (1:45:34)" list and a "Padrões de Hearst (1:53:36)" list.
 - Padrões de Hearst:** A section for Hearst patterns.
 - Padrões de Morin:** Includes a "Padrões de Morin (1:53:52)" list.
- CLASS BROWSER (Left):** Displays a "Class Hierarchy" for "For Project: Taxonomia.Compostos (1:45:34)". The hierarchy starts with ":SYSTEM-CLASS" and includes classes like [cabana], [sistema], [estrada], [pousada], [vôo], and [vôo]_regular, with various sub-classes.
- CLASS BROWSER (Right):** Displays a "Class Hierarchy" for "For Project: ". The hierarchy starts with ":THING" and includes classes like [armadilha], [cardume], [equipamento]_obrigatório, [passeio], [pensão]_completo, and [espécie]_de_mamífero, with various sub-classes.

AVALIAÇÃO

- Avaliação dos Métodos:
 - Precisão, Abrangência e *F-measure*
 - Recursos Necessários:
 - Corpus
 - Ontologia de Referência
 - **Construída Manualmente**
- Avaliação feita pelo usuário:
 - Avaliação das Funcionalidades do plug-in
 - Avaliação da etapa de Extração de Termos
 - **Seleção por Grupos Semânticos**



AVALIAÇÃO/RECURSOS

- Avaliação dos Métodos:
 - *CórpusEco* (Ecologia)
 - Lista de Termos uni, bi e trigrama
 - Ontologia do subdomínio “Ecologia das Comunidades”
- Avaliação feita pelo usuário:
 - *NanoTerm* (Nanotecnologia & Nanociência) (USP/SCar)
 - *JPED* (Pediatria) (UFRGS)



CONSIDERAÇÕES FINAIS

- Principal Dificuldade:
 - Escassez de recursos de avaliação
- Principais Contribuições:
 - Avaliação do uso de informações semânticas na construção de ontologias para o Português
 - Criação de uma ferramenta de uso geral para auxílio ao processo de construção de ontologias
 - Desenvolvimento dos módulos de avaliação automática das etapas executadas



CONSIDERAÇÕES FINAIS

- EDITAL CT-INFO 2007 - “Grandes Desafios da Computação”
 - Objetivo Geral:
 - Organizar recursos, métodos de construção e avaliação, e publicações da área de Ontologias para a Língua Portuguesa;
 - Objetivos Específicos:
 - Impulsionar as pesquisas na área;
 - Promover a interação entre grupos de pesquisa;
 - Disponibilizar subsídios para comparação entre métodos (*Benchmark*)

