Estrutura Ontológica e Unidades Lexicais: uma aplicação computacional no domínio da Ecologia

Claudia Zavaglia – UNESP/IBILCE
Leandro Henrique Mendonça de Oliveira – USP/ICMC-NILC
Maria das Graças Volpe Nunes – USP/ICMC-NILC
Sandra Maria Aluísio – USP/ICMC-NILC

unesp
UNIVERSIDADE ESTADUAL PAULISTA

Objetivo

- Avaliação de métodos de extração automática de termos (EAT) a partir de córpus para a construção de ontologias.

Objetivo

- Avaliação de métodos de extração automática de termos (EAT) a partir de córpus para a construção de ontologias.

Extração de Termos

- A extração de termos é o reconhecimento dos candidatos a termos em um córpus especializado;

- O extrator de termos é um conjunto de programas ou ferramentas computacionais que reconhece e extrai as unidades terminológicas (termos) que aparecem nos córpus especializados.

Trabalho Suporte

Projeto ExPorTer:

Teline, M.F. Avaliação de Métodos de Extração Automática de Terminologia para textos em Português. ICMC-USP, São Carlos, SP, Fevereiro 2004. Dissertação de Mestrado.

Domínio: Revestimento Cerâmico;

córpus 448.352 ocorrências

15 métodos simples das 3 abordagens: lingüística, estatística e híbrida (lingüística + estatística + lingüística)

Abordagens EAT

- Métodos estatísticos
  - dependentes do tamanho do córpus 😟
    - se o córpus de aplicação é pequeno, gera-se muito silêncio,
    - mesmo quando o córpus apresenta milhões de ocorrências,
      - há sempre uma porcentagem de palavras que não podem ser recuperadas em razão de sua baixa freqüência de uso no córpus.
  - geram bastante ruído 😟
    - muitas das palavras da língua geral aparecem nos textos com uma alta freqüência
  - são independentes da língua 🙂

- Métodos lingüísticos
  - a grande quantidade de ruído gerada (entre 55% e 75%) é um dos problemas principais dos sistemas 😟
    - que trabalham apenas dados morfológicos, morfossintáticos,sintáticos e/ou léxicos.
- são dependentes da língua e até de variante 😟

- Métodos híbridos
  - aqueles que aplicam o conhecimento estatístico primeiro e depois o lingüístico,
    - mesmos problemas de silêncio dos sistemas puramente estatísticos. 😟
  - aqueles que utilizam a estatística apenas como um complemento da lingüística.
    - os resultados finais melhores
    - estatística auxiliar no momento do processo de detecção, reafirmando ou recusando a condição de termo de uma unidade lingüística. 🙂

Gargalo da área EAT

- Avaliação
  - Exige a opinião de especialistas
  - É um trabalho caro e demorado
  - Uso de listas de referências (avaliação tradicional) com as medidas tradicionais de precisão e revocação (recall) traz riscos
    - Listas são incompletas, devido a constante produção de novos termos

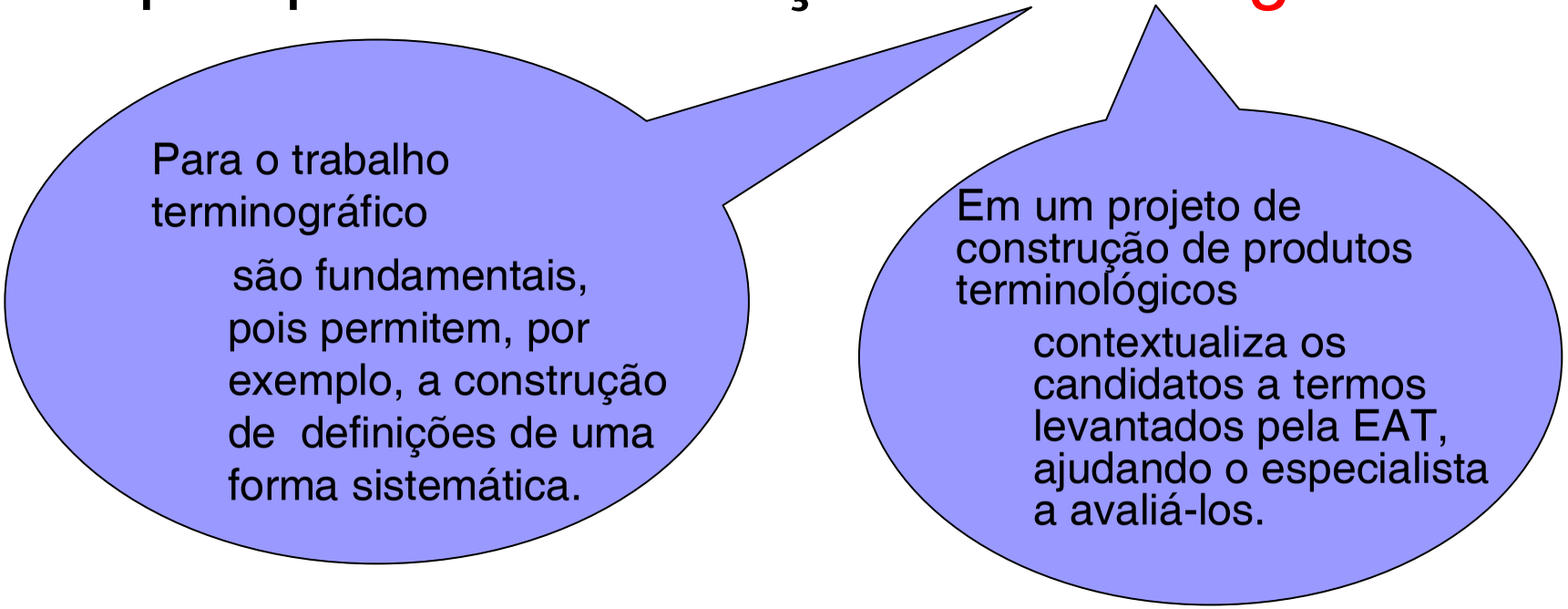
- Nossa avaliação foi tradicional
  - Feita com uma lista de referência com termos do domínio da Ecologia, usando as métricas de precisão e revocação.

Cenário no Brasil

- Um dos primeiros sistemas de EAT (TERMINO)
  - 1989, na Universidade de Quebec, Montreal.
- Muitos projetos de construção de repertórios terminológicos
  - ainda utilizam o critério semântico para a extração de termos, em uma abordagem manual a partir de córpus.
- Critério semântico
  - adequado
    - a extração manual é lenta,
    - sujeita à subjetividade e
    - à omissão de termos importantes.

Objetivo

- Avaliação de métodos de extração automática de termos (EAT) a partir de córpus para a construção de ontologias.



Conceituando “ontologia”

“Uma ontologia é o vocabulário usado para representar um certo domínio do conhecimento e a conceituação que estes termos pretendem capturar.” (Chandrasekaran, et al. 1999).

“ Ontologias fornecem um vocabulário comum de uma área e define, com níveis distintos de formalismos, o significado dos termos e dos relacionamentos entre eles.” (Gómez-Pérez, 1999)

“Ontologias são termos e relações que compreendem o vocabulário de uma área, como também as regras para combinar estes termos e relações para definir extensões deste vocabulário.” (Novello, 2002)

Ontologias

- O domínio -

- Por que um domínio?
  - Ontologias gerais x Ontologias específicas

📂 Ontologias gerais:

- 🗀 Representam grandes e substanciosos conjuntos de elementos.
- 🗀 Representam o senso comum de uma comunidade sociolingüística.
- 🗀 Elaboração lenta dada a infinidade de informações contidas no Universo.

Ontologias

- O domínio -

- Por que um domínio?
  - Ontologias gerais x Ontologias específicas

📂 Ontologias específicas:

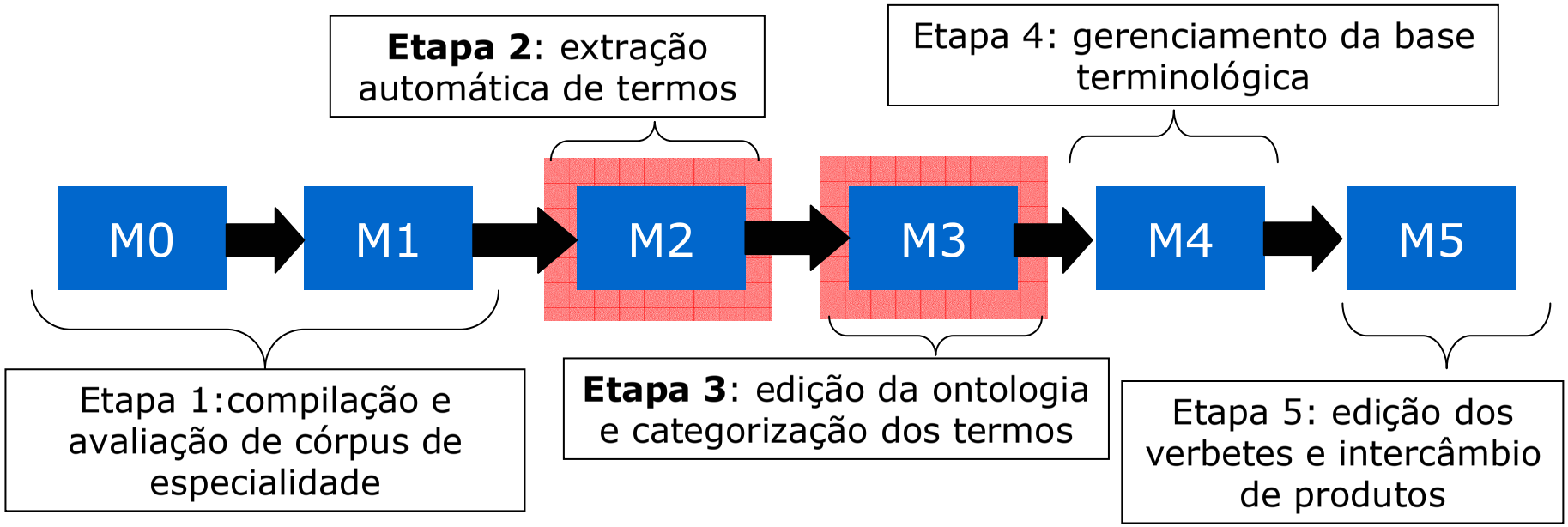
- 🗀 Representam conjuntos de objetos e suas relações de um reduzido e específico domínio.
- 🗀 Representam o consenso de um grupo de especialistas de uma área restrita e especial.
- 🗀 Elaboração "rápida" e "simples", uma vez que o número de informações é restrito e limitado.

- Este trabalho foi um exercício para refinar o ambiente e-termos, sendo desenvolvido no NILC
  - Oliveira, L. H. M. e-Termos Ambiente Web colaborativo para criação de produtos terminológicos.

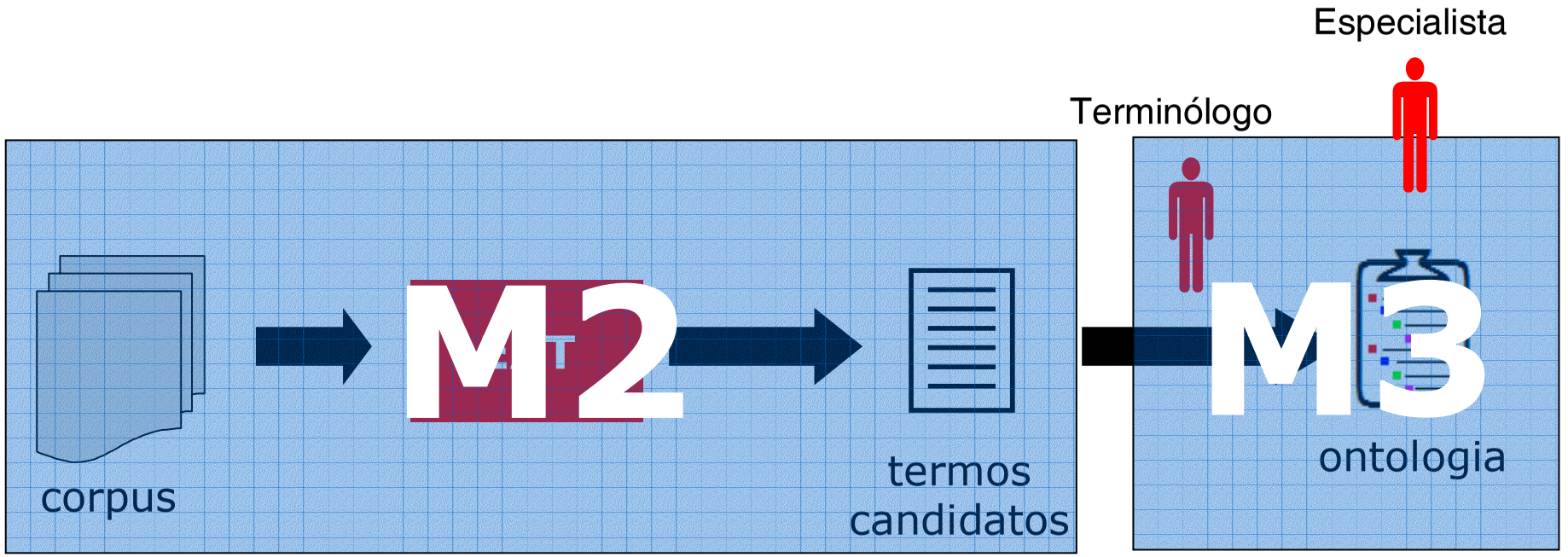
    ICMC-USP, São Carlos, SP. 2006. Qualificação de Doutorado.

O e-Termos

O ***e-Termos*** é um Ambiente Colaborativo Web (*Computer-Supported Collaborative Work - CSCW*) composto por seis módulos de trabalho independentes para a criação de produtos terminológicos.



Usando Ontologias a partir da Extração de Termos



Metodologia da Pesquisa

- Fase 1 – Preparação: O CórpusEco

- A Economia da Natureza
- Ecologia

Ed. Guanabra Koogan

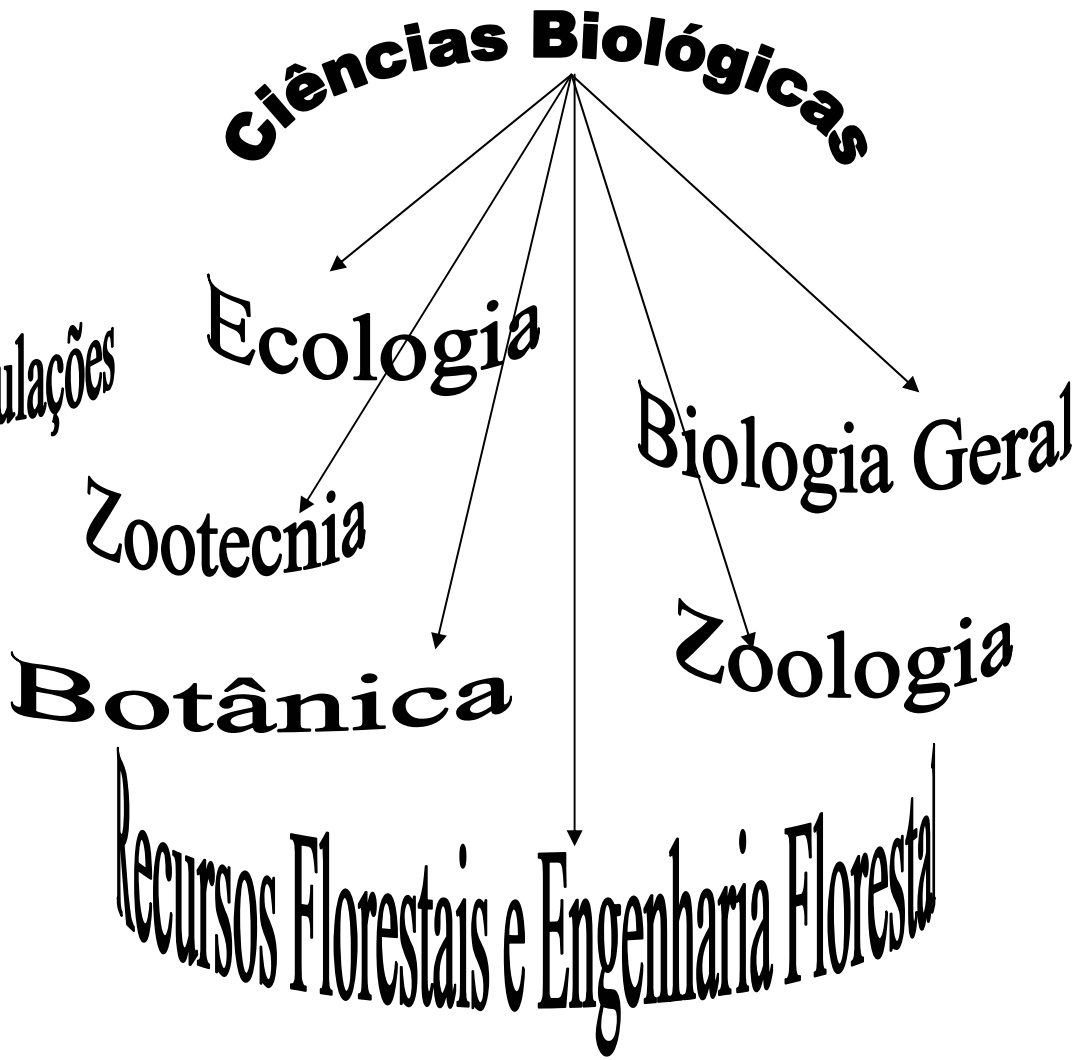
CórpusEco

260.921 ocorrências

- Córpus de Referência

Lácio-Web

Textos didáticos



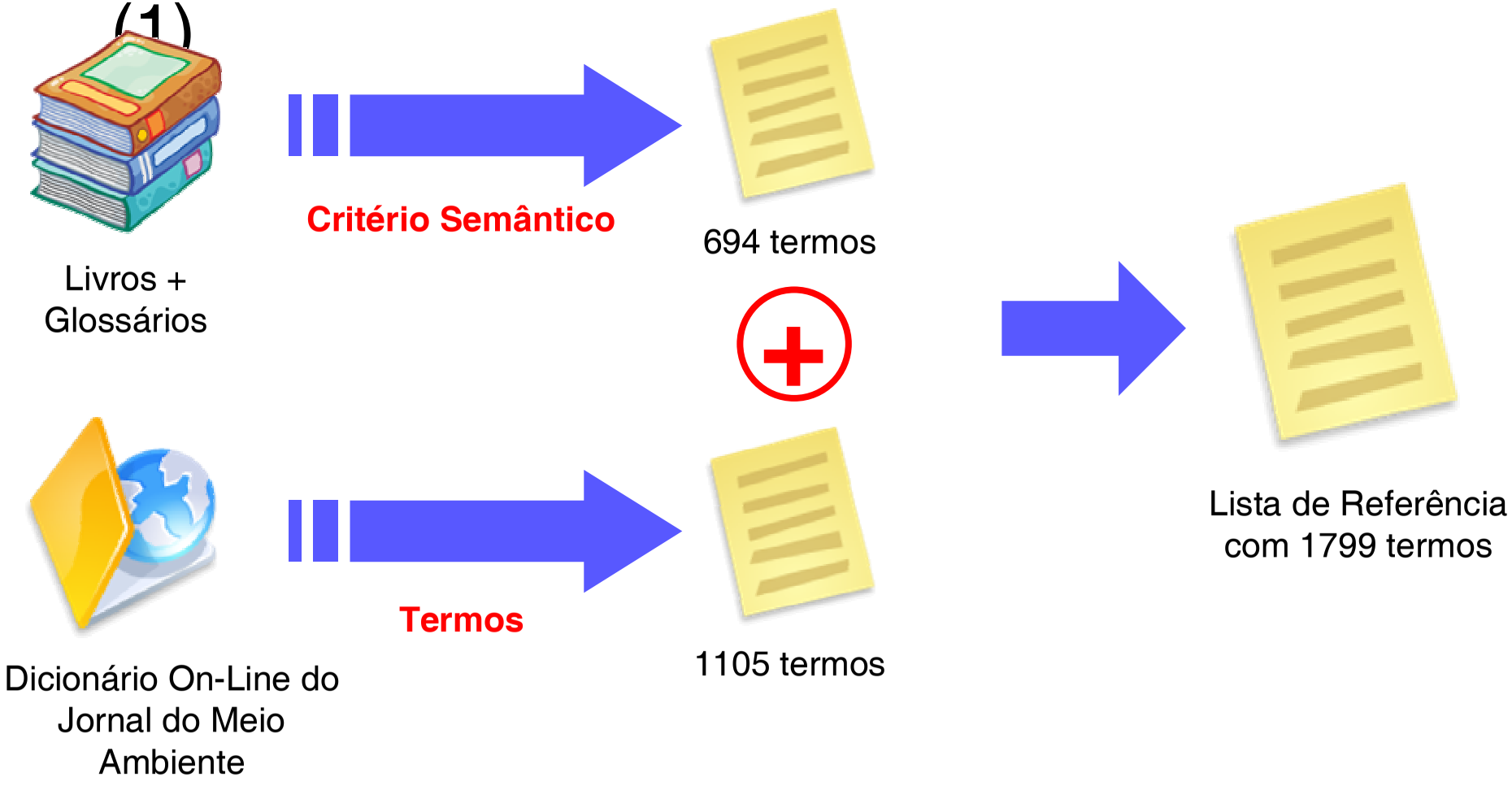
CórpusEco

Gêneros

Textos técnico-científicos do Lácio-Web

Metodologia

- Fase 1 – Preparação: Lista de Referência (1)



Metodologia

- Fase 1 – Preparação: Lista de Referência (2)

Metodologia

- Fase 2 – Aplicação dos Métodos de Extração

  Três Abordagens:

Estatística
- Oito métodos

Lingüística
- Três métodos

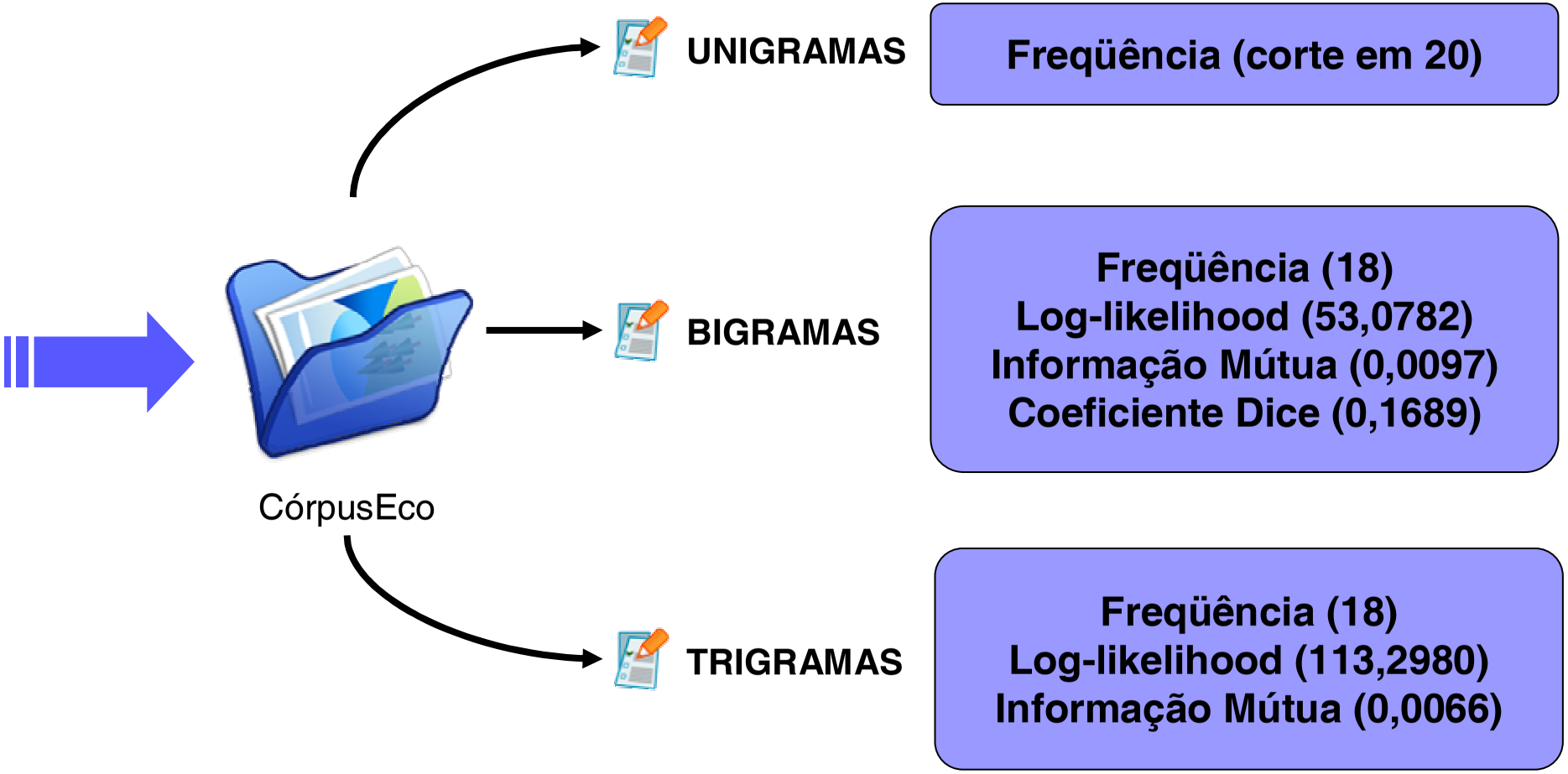
Híbrida (Lingüístico + Estatístico)
- Quatro métodos

Total de 15 métodos.

Metodologia

Abordagem Estatística

Quatro Medidas Estatísticas do pacote NSP



Metodologia

Abordagem Lingüística (Heid et al (1996); Klavans e Muresan (2000))

CórpusEco

POS tagger

CórpusEco Etiquetado

Estruturas Lingüísticas
Marcadores Estruturais
Padrões Morfossintáticos
da Ecologia
StopList

UNIGRAMAS ← **n / np / adj / verb**

BIGRAMAS ← **n_adj / n_n / adj_n / adj_adj / n_adv**

TRIGRAMAS ← **n_prep_n / n_prep_np / n_n_adj / n_adj_adj / n_prep_adj**

Classes de expressões e indicadores estruturais (expressões)

- Uso Geral (UG): as expressões apresentadas nessa classe podem ser utilizadas em qualquer domínio de especialidade.

- Conceitual do domínio de Revestimentos Cerâmicos (CD): as expressões encontradas nessa classe podem ser aplicadas preferencialmente para córpus do domínio de Revestimentos Cerâmicos.

- Sinais gráficos (SG): nessa classe considera-se “( )”, “:” e “-”

Aluísio (1995)	
apresenta	(UG)
atua	
caracterizado	(UG)
classe de	(UG)
compreendendo	
compreendido	(UG)
conhecido como	(UG)
consiste	(UG)
contém, contêm	(UG)
em outras palavras	
implica	(UG)
isto é	(UG)
ou seja	(UG)
por exemplo	(UG)
tal como	(UG)

Sager (1993)	
é	(UG)
são	(UG)
utilizado	(UG)

Almeida (2000)	
característica do	(CD)
composição do	(CD)
composto	(CD)
estado de	
matéria-prima	(CD)
método	
parte de	
processo	(CD)
propriedade de	(CD)
tipo de	(CD)

CórpusEco	
adição de	
chamamos	(UG)
constitui	(UG)
constituído	(UG)
depende	(UG)
desenvolvido	(UG)
determinado	(UG)
empregado	(UG)
expresso	(UG)
formado	(UG)
obtido	(UG)
palavra	(UG)
relacionado	(UG)

Klavans e Muresan (2000;2001a;2001b)	
chamado	(UG)
definido como	(UG)
expressão	(UG)
(se) entende	
significa	
termo	(UG)
()	(SG)
-	(SG)
:	(SG)

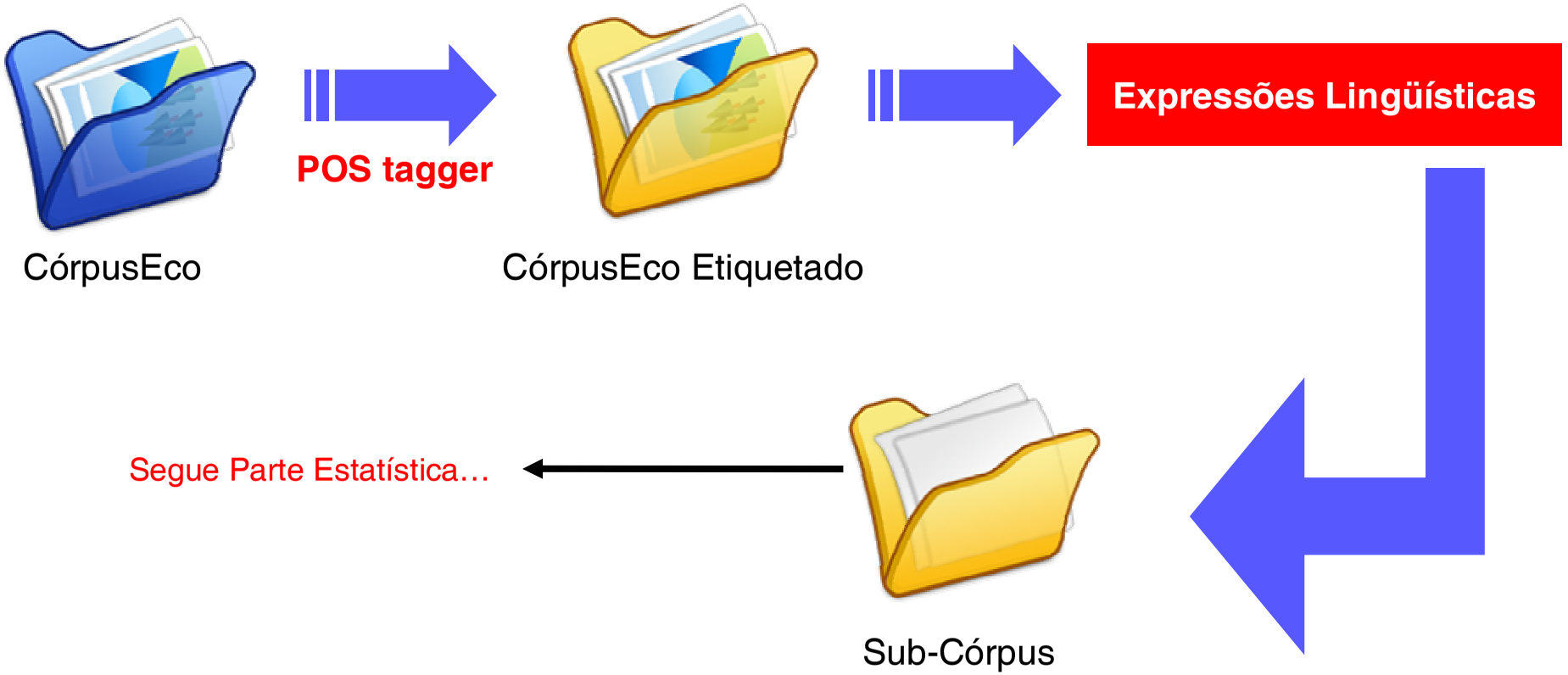
ISO/TC 37/SC 1	
conceito	(UG)
corresponde	(UG)
define	
denominado	(UG)
feito de	
usado	(UG)

56 Expressões lingüísticas e marcadores estruturais levantados; 45 implementados (Teline, 2004)

Metodologia

Abordagem Híbrida (1)

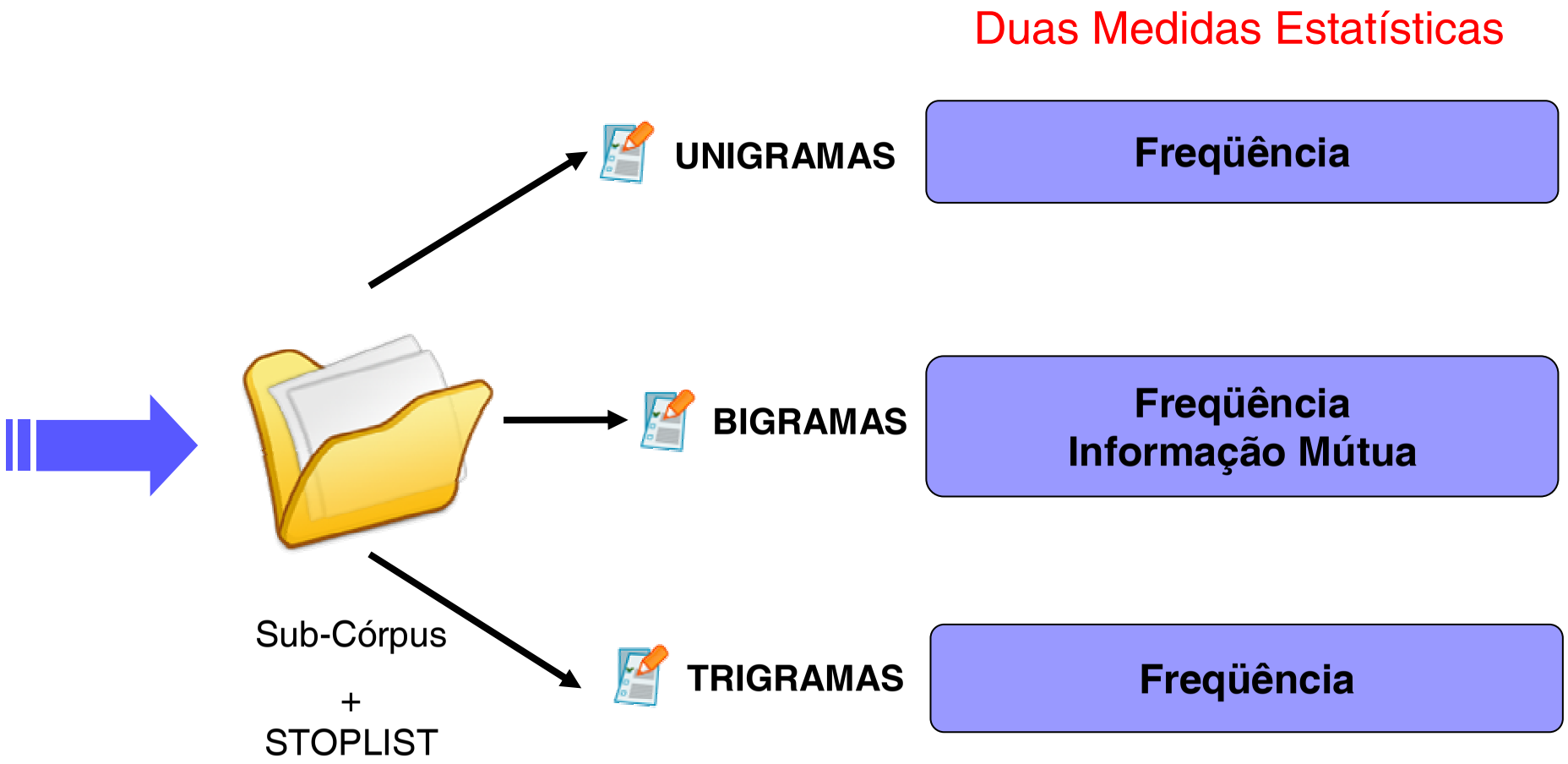
Parte Lingüística:



Metodologia

Abordagem Híbrida (2)

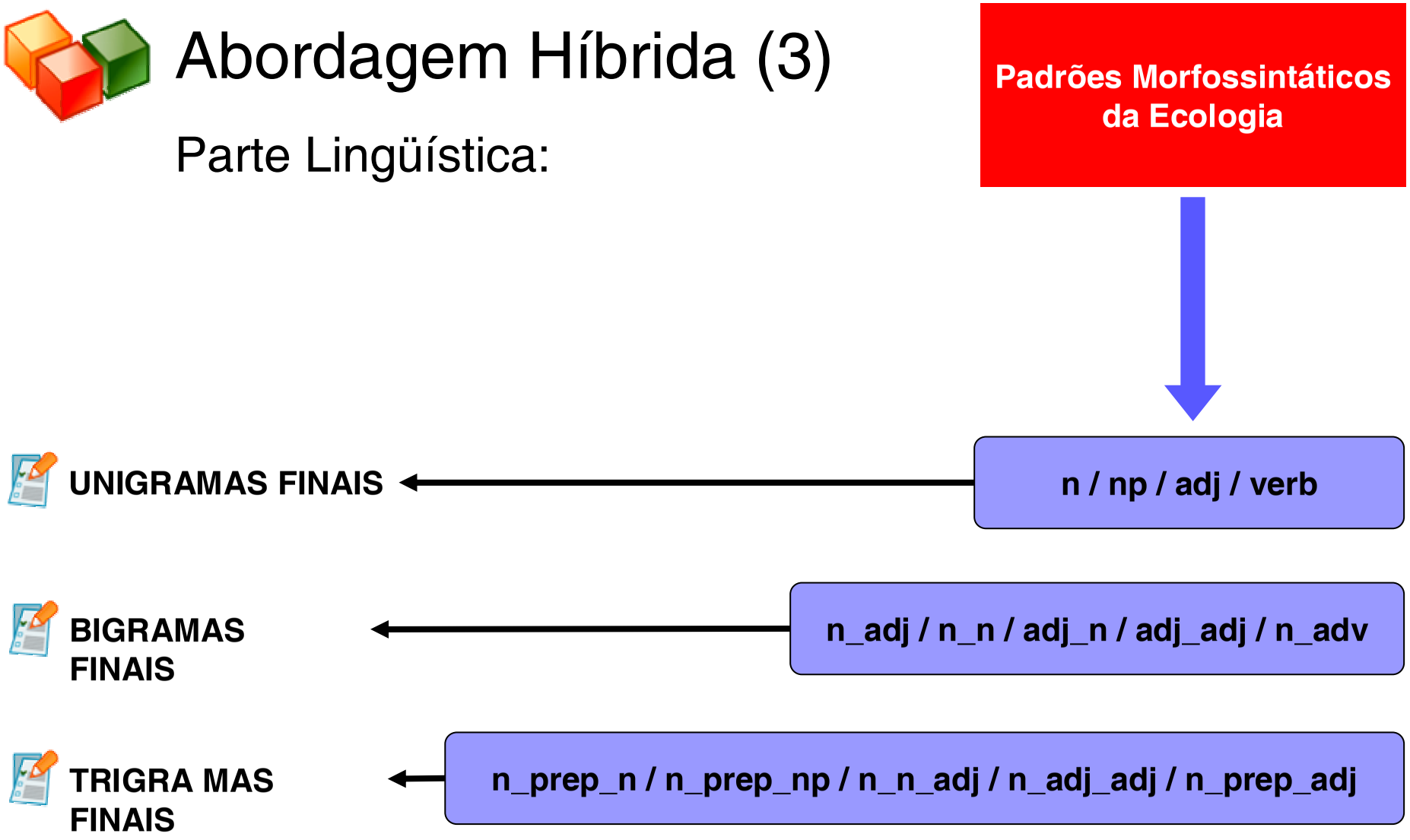
Parte Estatística:



Metodologia

Abordagem Híbrida (3)

Parte Lingüística:



Resultados - Geral

O desempenho nestas tarefas raramente excede F = 0.60

Abordagem	Métodos	Precisão	Revocação
Estatística	Freqüência – unigramas	9,48	34,27
	Freqüência – bigramas	20,31	14,44
	Log-Likelihood – bigramas	20,31	14,44
	Informação mútua – bigramas	20,31	14,44
	Dice – bigramas	20,31	14,44
	Freqüência – trigramas	2,41	10,23
	Informação mútua – trigramas	2,41	10,23
	Log-Likelihood – trigramas	2,41	10,23
Lingüística	ExPorTer_lingüístico – unigramas	2,74	89,18
	ExPorTer_lingüístico – bigramas	1,31	62,22
	ExPorTer_lingüístico – trigramas	0,89	82,95
Híbrida	ExPorTer_híbrido c/ Freqüência – unigramas	12,76	23,25
	ExPorTer_híbrido c/ Freqüência – bigramas	41,18	7,78
	ExPorTer_híbrido c/ Freqüência – trigramas	18,75	3,41
	ExPorTer_híbrido c/ Informação mútua – bigramas	1,68[1]	65,0

[1] Calculada sem corte.

Resultados - 150 primeiros termos dos métodos com melhor precisão

150 candidatos a termos analisados	Abordagem estatística	Abordagem Lingüística	Abordagem híbrida
unigramas	42 (28%) (>n)	21(14%) (n)	45 (30%) (>n)
bigramas	51(34%) (n+adj)	4 (n+adj)	19 (n+adj)
trigramas	10 (6,6%) (n+prep+n)	1 (n+prep+n)	3 (n+prep+n)
TOTAL	**103**	**26**	**67**

Considerações Finais

- A EAT alimentou o delineamento da estrutura arbórea da ontologia
  - forneceu unidades terminológicas que se caracterizaram como classes ou subclasses, tais como: *população, comunidade, energia, área, ecologia, tabela de vida*, entre outros.

- Já a abordagem lingüística carece de maior detalhamento ainda para que possamos obter resultados melhores em uma próxima tentativa de extração.

- Embora os valores dessa nossa avaliação não tenham sido altos,
  - Métodos de EAT refinados e especializados para os gêneros tratados, e tamanho de córpus em uso,
    - poderá auxiliar de maneira eficaz o trabalho de extração do terminólogo, pois são rápidos.
    - Comparados com extração de termos com base no critério semântico, essa captura de unidades lexicais demandaria tempo e seria certamente muito mais lenta e subjetiva.

Trecho do CórpusEco

Segundo o Novo Dicionário Aurélio, uma definição de “estabilidade” é: “Propriedade geral dos sistemas mecânicos, elétricos e aerodinâmicos, pela qual o sistema retorna ao estado de equilíbrio após sofrer uma perturbação." Poderíamos generalizar esta definição para incluir todos os sistemas, inclusive os ecológicos. (N. do T.)

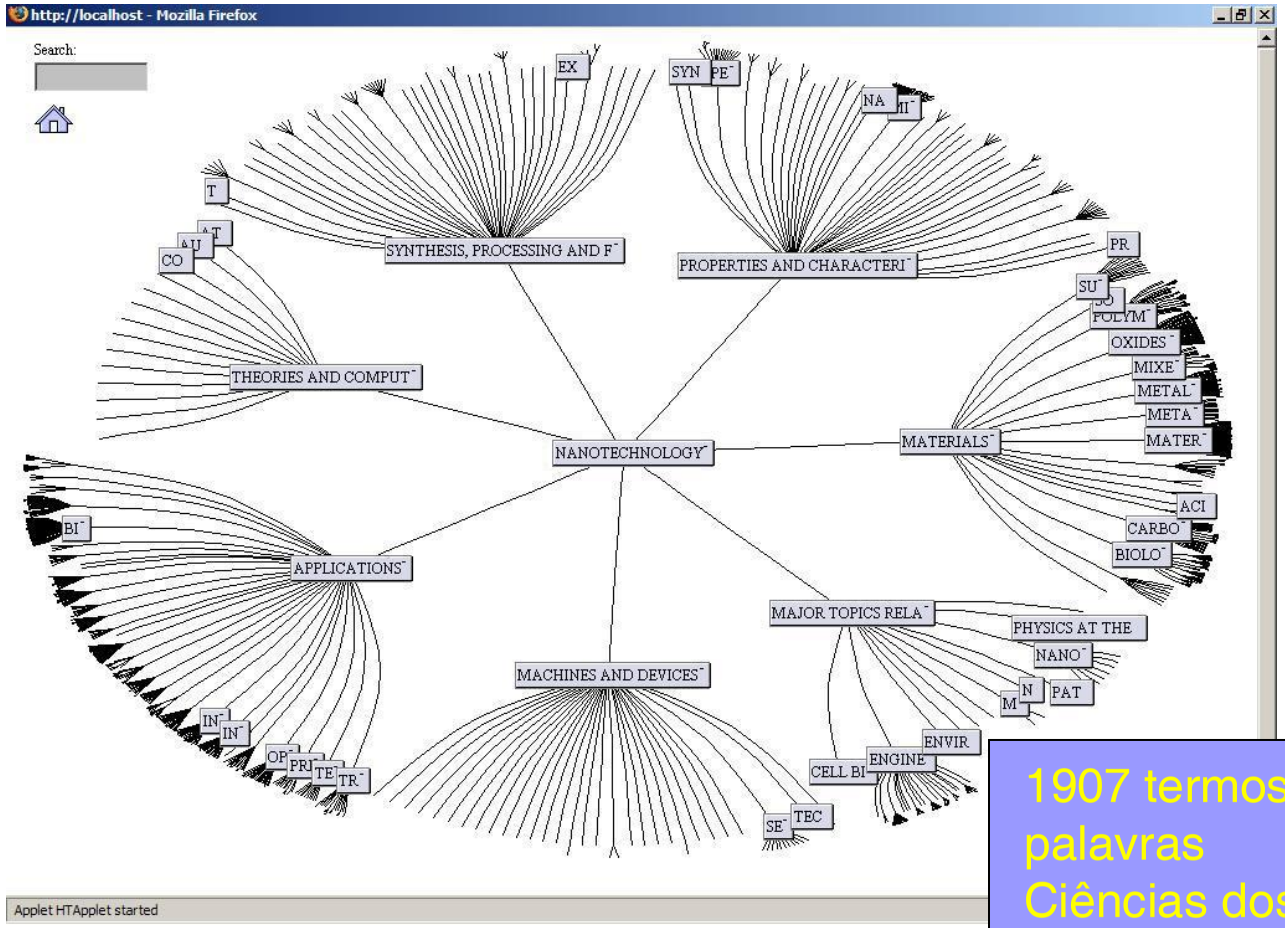
Exemplo de definição, indicadores estruturais enfatizados

Projetos financiados do NILC

- Desenvolvimento de uma estrutura conceitual (ontologia) para a área de Nanociência e Nanotecnologia (N&N), 2005
  - Gênero técnico-científico

- Terminologia em Língua Portuguesa da Nanociência e Nanotecnologia: Sistematização do Repertório Vocabular e Elaboração de Dicionário-Piloto, 2006-2008.
  - Gênero informativo, científico de divulgação, técnico-científico.

N&N Inglês



Interface Web de busca e visualização hiperbólica

Disponível em: www.nilc.icmc.usp.br/nanotech

6 grandes entradas:
1. Synthesis, Processing and Fabrication
2. Materials
3. Properties and Characterization techniques
4. Machines and Devices
5. Theories and Computational methods
6. Applications

1907 termos; corpus 2,5 milhões de palavras
Ciências dos Materiais, Biociências, Física e Química Teóricas, Engenharia Eletrônica e Ciência da Computação

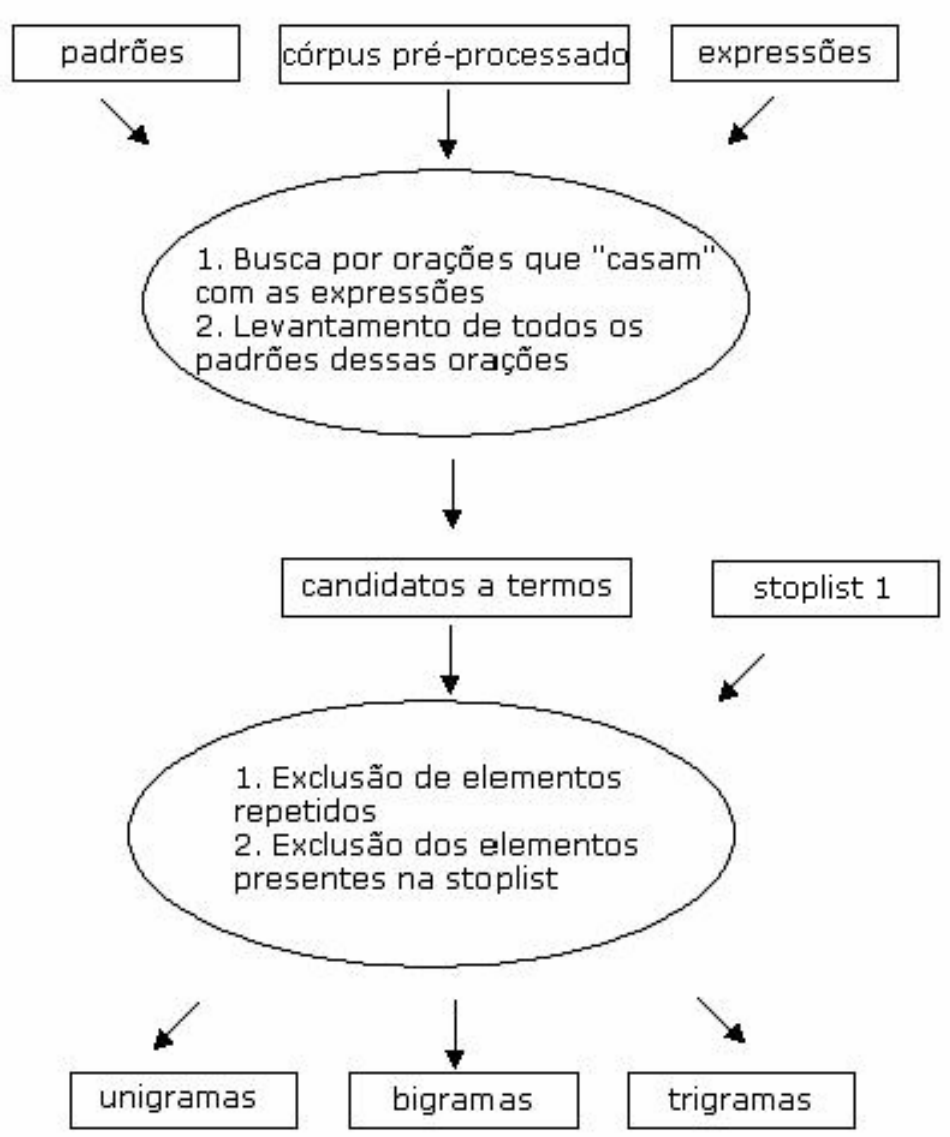
N&N Português

- Córpus de 2 milhões de palavras
  - Uso de um cabeçalho para os textos, seguindo o padrão do projeto Lácio-Web
- Previsão de 500 entradas para o dicionário piloto

Referências

- SANTOS, E. T. ; BARROS, L. N.; VALENTE, V. C. P. N. Projetando uma Ontologia de Geometria Descritiva. **Anais do IV Congresso Internacional de Engenharia Gráfica nas Artes e no Desenho (GRAPHICA 2001),** p.918-928, nov. 2001, São Paulo, SP.

- CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. *What are ontologies, and why do we need them?* IEEE Intelligent Systems, p.20-25, January/February 1999.



Método ExPorTer_lingüístico (Teline, 2004)

Informação Mútua (IM), Log-likelihood (LL) e Dice

- IM uma medida da quantidade de informação que uma variável contém sobre uma outra, sendo ela a redução da incerteza de uma variável randômica devido ao conhecimento da outra.
  - usada inicialmente para extração de colocações.
    - Existe uma sobreposição entre as colocações e os termos técnicos: as colocações têm uma composicionalidade limitada, e os termos técnicos aceitam um número limitado de modificadores.
  - Quando todas as ocorrências de x e y são adjacentes umas às outras, a informação mútua é a maior,
    - deteriorando-se em contas de baixa freqüência.
- LL, por se apresentar mais robusta para eventos de baixa freqüência, é utilizada a fim de amenizar o problema da informação mútua quando esta apresenta contagens de baixa freqüência.
- Coeficiente *dice* depende apenas da freqüência do bigrama e das palavras do bigrama.
  - Diferentemente do que ocorre com a informação mútua, essa medida não depende do tamanho da amostra