Olga Craveiro[1,2], Joaquim Macedo[3], Henrique Madeira[2]

([1]) ESTG, Instituto Politécnico de Leiria

([2]) CISUC, DEI, Universidade de Coimbra

([3]) DI, Universidade do Minho

PorTexTO
Sistema para anotação/extracção de expressões temporais

Encontro do Segundo HAREM

Aveiro, 7 de Setembro de 2008

Resumo

- Motivação
- Características
- Arquitectura
- Funcionamento
- Exemplo
- Conclusões e Trabalho Futuro

Motivação

- Identificação das expressões temporais (entidades mencionadas temporais) mais utilizadas em textos de língua portuguesa

- Algoritmo simples e rápido que possa ser utilizado em ambientes de tempo real

PorTexTO

- Designação "*POR*tuguese *T*emporal *EX*pressions *TO*ol"
- Utiliza padrões de expressões temporais
  - Padrões criados através de identificação das co-ocorrências
  - A definação dos padrões é feita com REGEX
- Processamento do texto é realizado frase a frase, em que a divisão das frases é realizada com o atomizador da Linguateca (módulo Perl Lingua::PT::PLNbase)

Arquitectura

REGEX

[Dd]e (oc_DATA\d{1,5})(do ano passado)?
[Dd]esde há (anos|décadas|dias|meses|semanas|séculos)
[Nn]o (mês|ano) (passado|anterior|seguinte)
[Ee]m (oc_MES\d{15})(?: (deste ano)|passado|último|(do ano passado))?
(...)

Palavras chave Temporais

amanhã
ano
anos
anteontem
anualmente
cedo
década
Dezembro
(...)

Corpus

(...) Foi uma violência anunciada: o líder do Sinn Fein -- o braço político do IRA -- falara poucos dias antes num «`show' espectacular» como resposta à iniciativa anglo-irlandesa lançada pelos primeiros-ministros da Grã-Bretanha e da República da Irlanda com a sua «declaração» de 15 de Dezembro do ano passado. Mas a campanha terrorista foi só parte da resposta. (...)

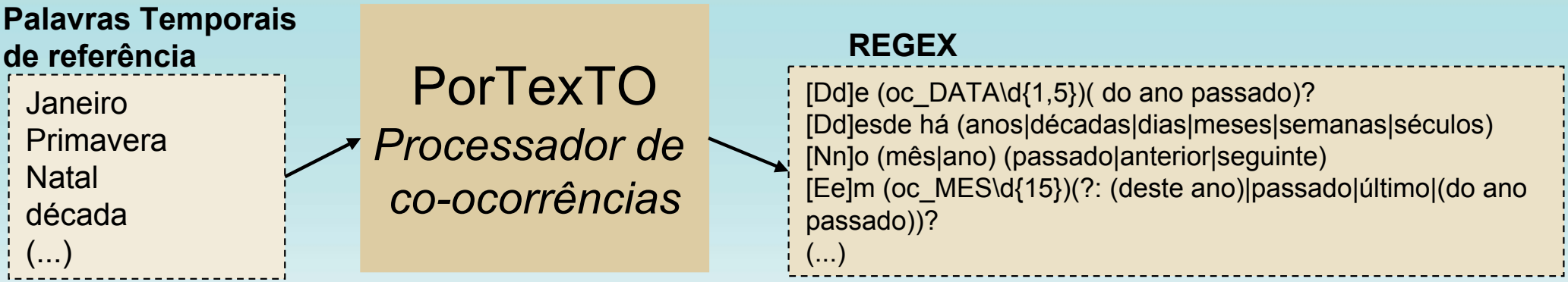
PorTexTO
Anotador

Corpus anotado

(...) Foi uma violência anunciada: o líder do Sinn Fein -- o braço político do IRA -- falara poucos dias antes num «`show' espectacular» como resposta à iniciativa anglo-irlandesa lançada pelos primeiros-ministros da Grã-Bretanha e da República da Irlanda com a sua «declaração» **<EM ID="2" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">*de 15 de Dezembro do ano passado*</EM>**. Mas a campanha terrorista foi só parte da resposta. (...)

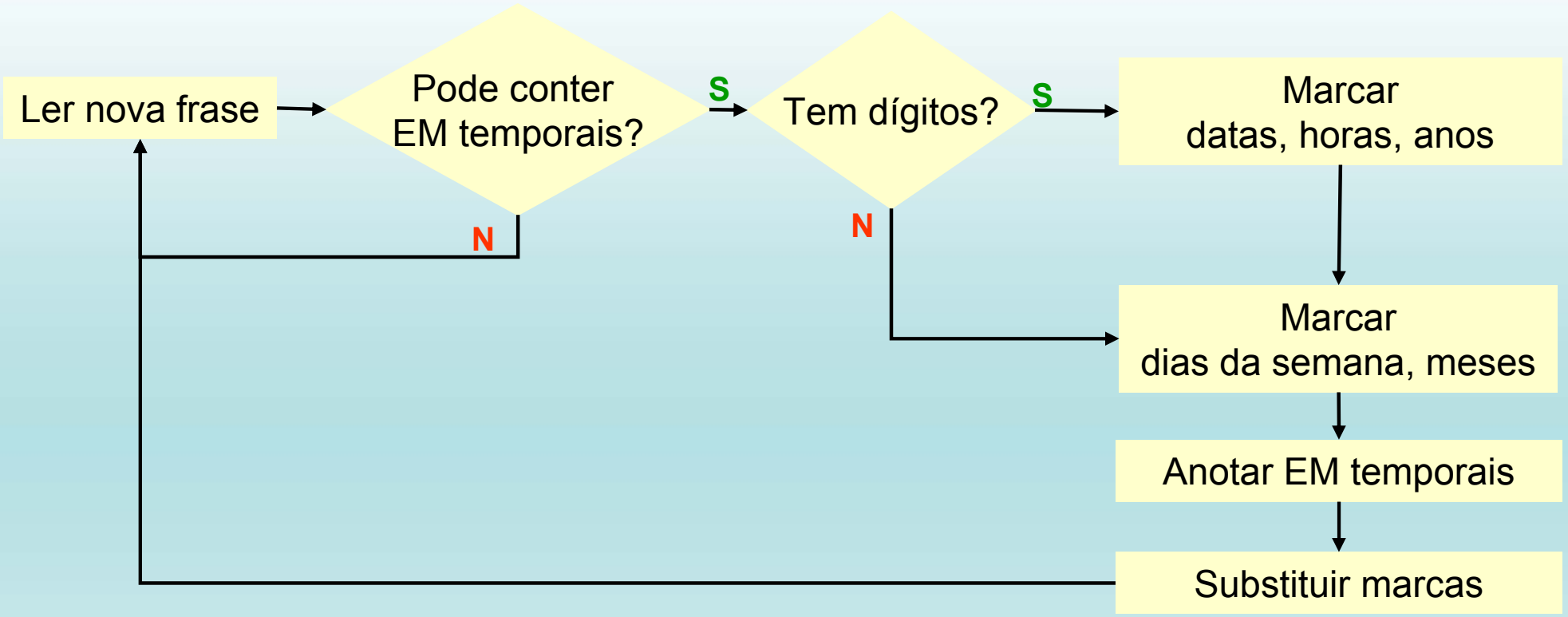
Funcionamento (1)

- REGEXs criadas com base nas co-ocorrências existentes em palavras de referência
- Palavras de referência utilizadas:
  - Meses do ano, dias da semana, estações do ano
  - Festividades: **Natal, Páscoa, Carnaval, Entrudo**
  - Medidas temporais: **década, período, século, ano, ...**



Funcionamento (2)

- Processamento realizado frase a frase

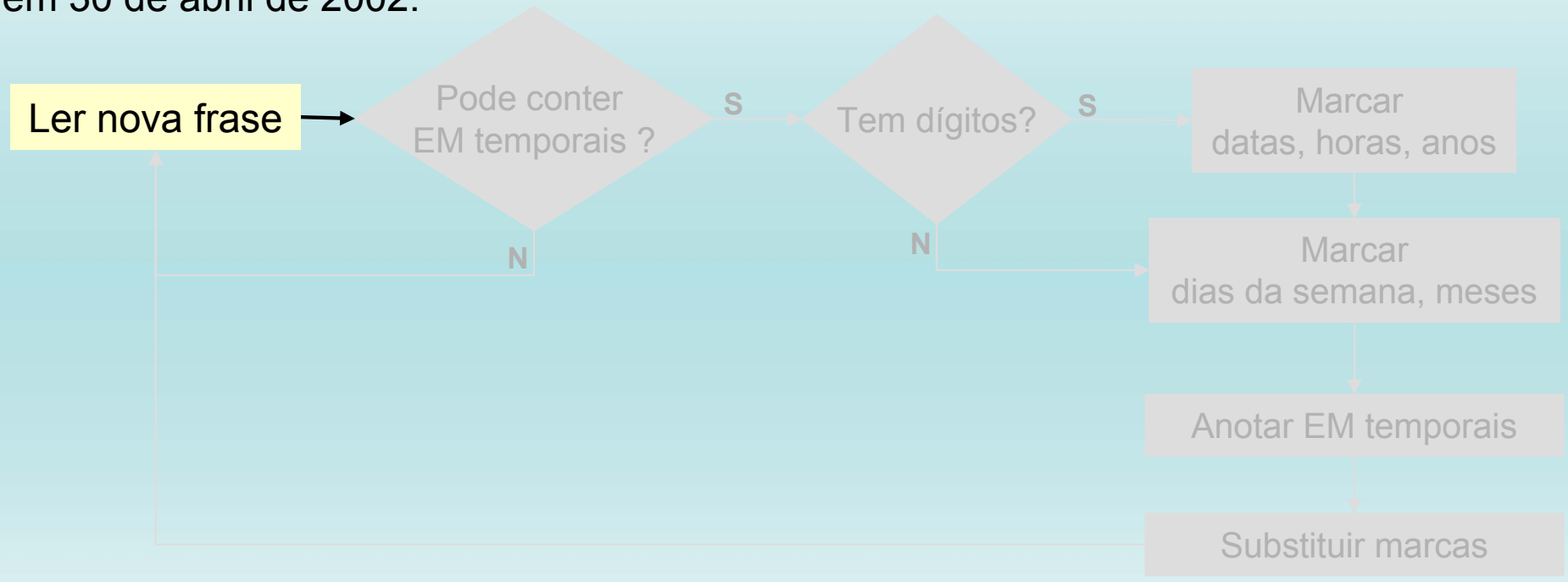


Exemplo (1)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

Frase = "A missão científica da nave foi concluída em 30 de abril de 2002."

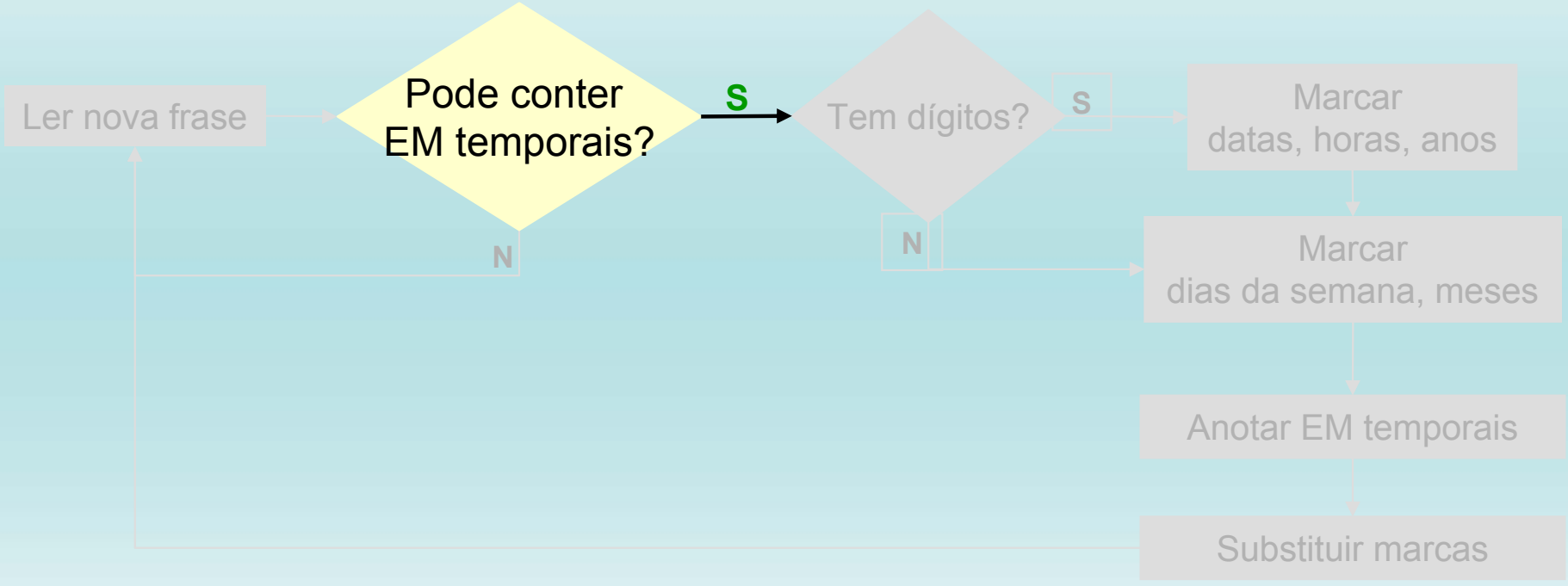


Exemplo (2)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

Frase = "A missão científica da nave foi concluída em 30 de abril de 2002."

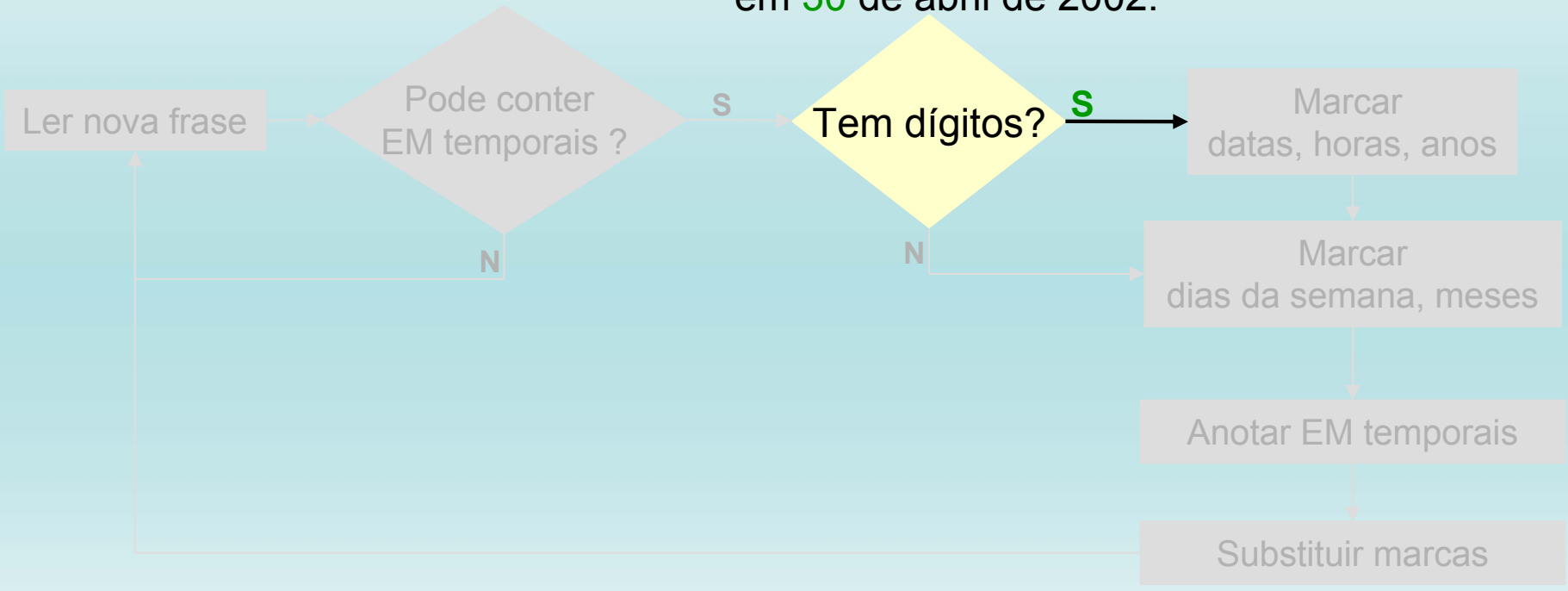


Exemplo (3)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

Frase = "A missão científica da nave foi concluída em 30 de abril de 2002."

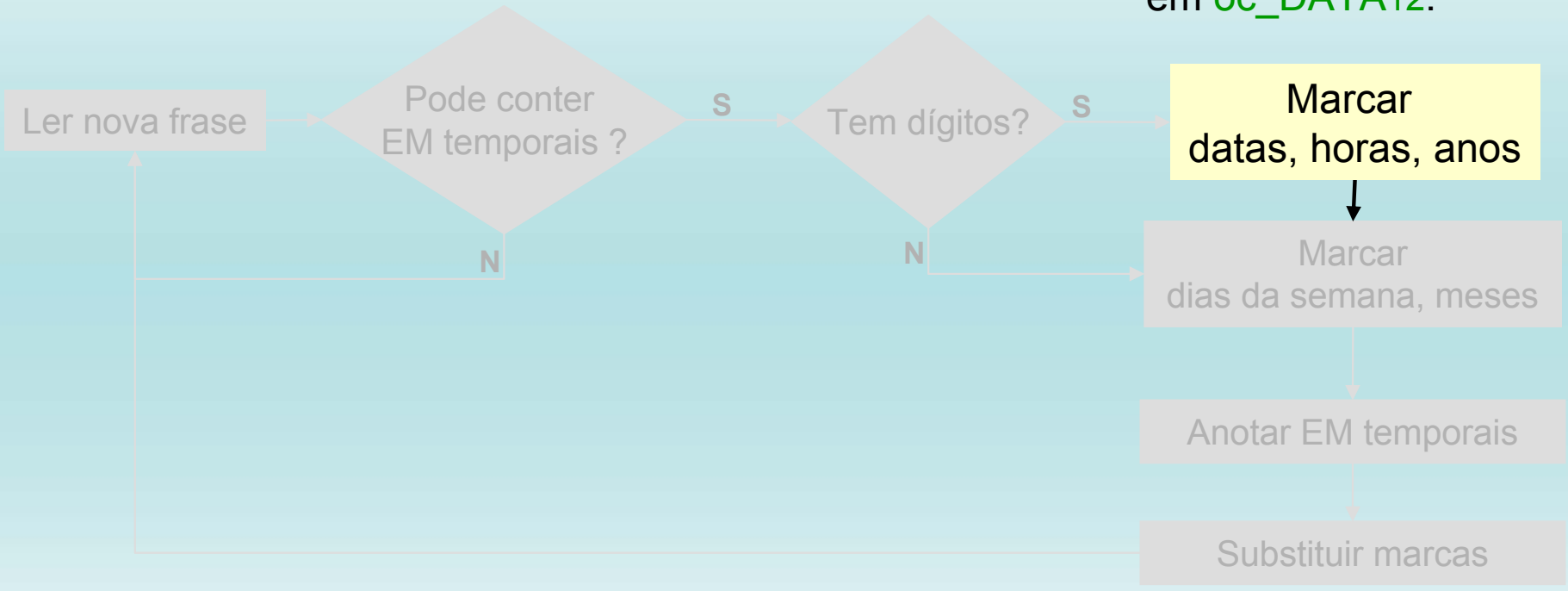


Exemplo (4)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

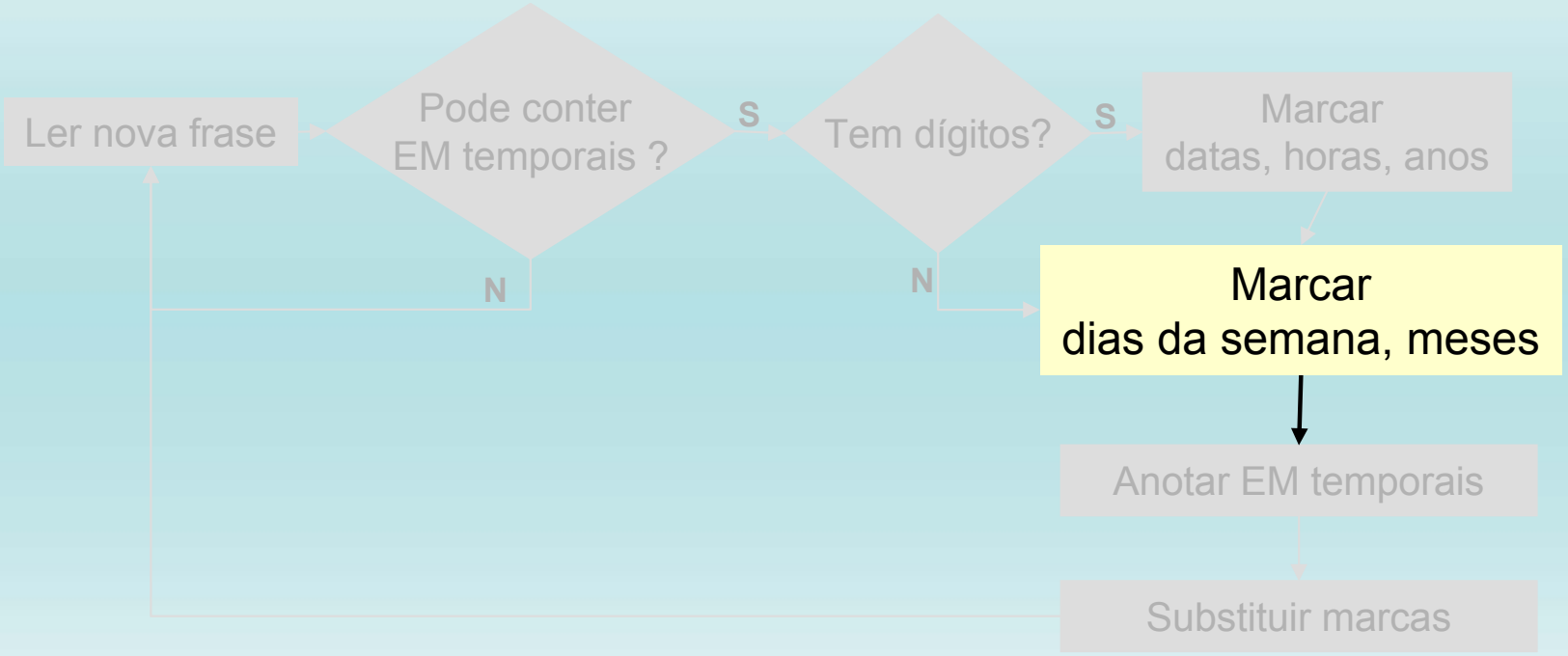
Frase = "A missão científica da nave foi concluída em oc_DATA12."



Exemplo (5)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

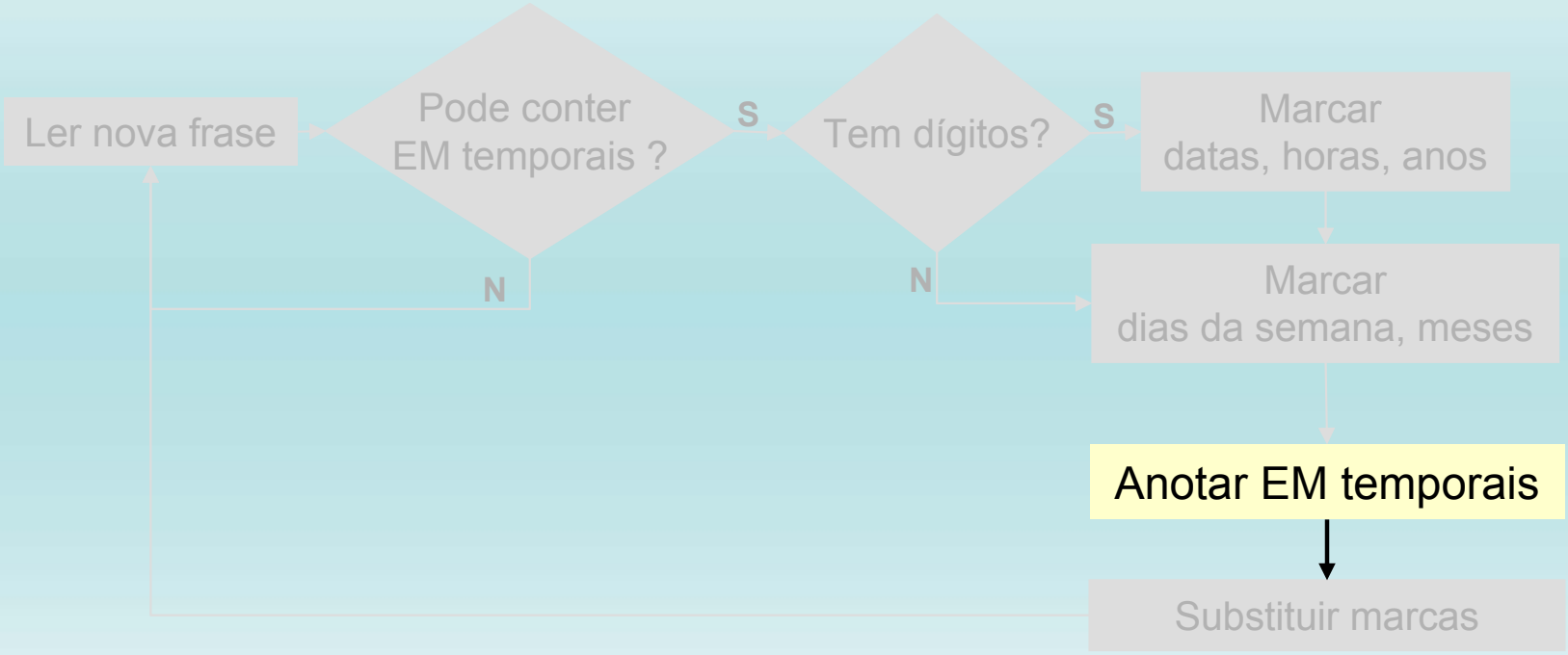


Frase = "A missão científica da nave foi concluída em oc_DATA12."

Exemplo (6)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

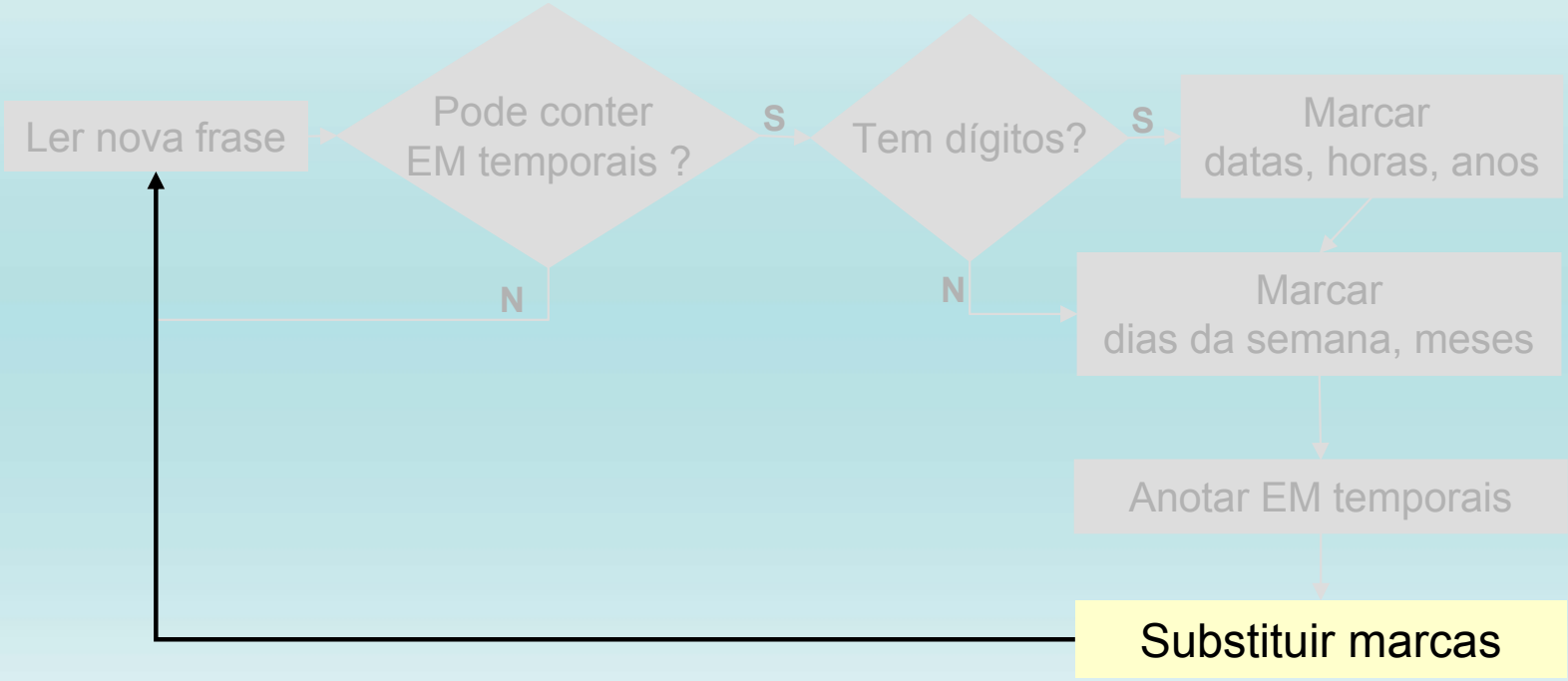


Frase = "A missão Científica da nave foi concluída
<EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">
em oc_DATA12
</EM>."

Exemplo (7)

Corpus anotado

(...) A missão científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">em 30 de abril de 2002</EM>. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)



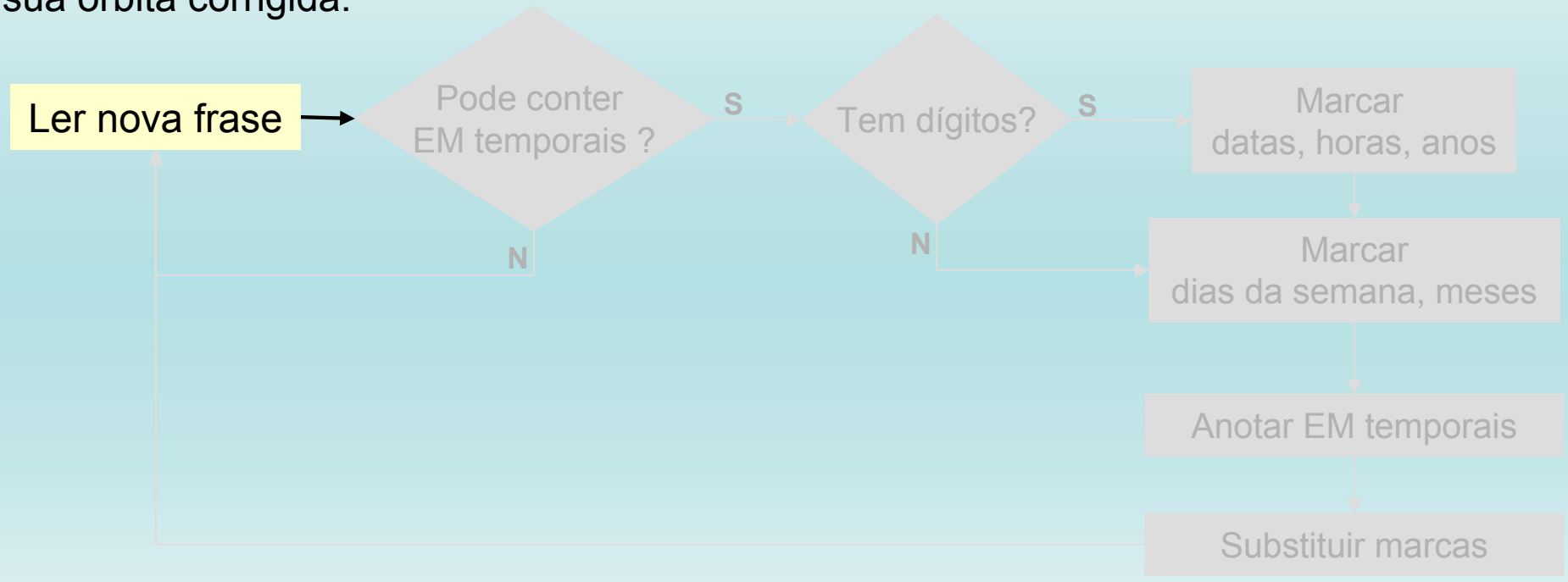
Frase = "A missão Científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA"> em 30 de abril de 2002 </EM>."

Exemplo (8)

Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

Frase = "Depois disso, o satélite deixou de ter sua órbita corrigida."



Exemplo (9)

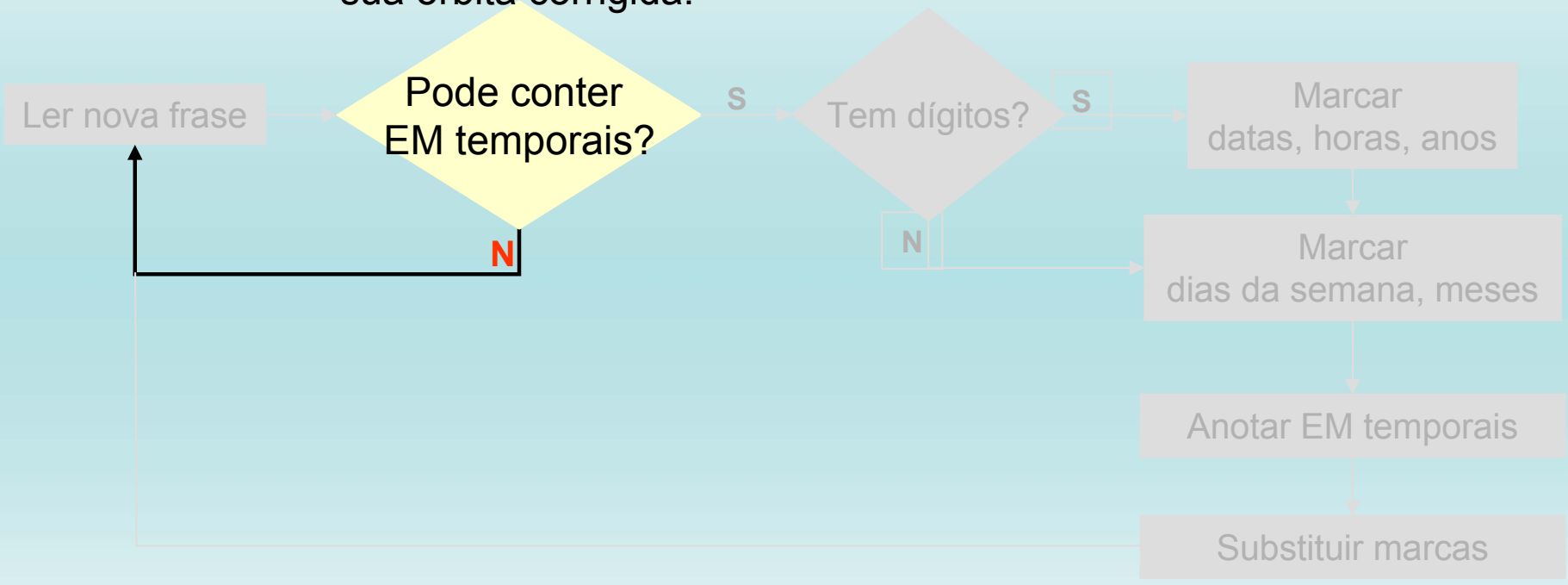
Corpus

(...) A missão científica da nave foi concluída em 30 de abril de 2002. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

Corpus anotado

(...) A missão científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">em 30 de abril de 2002</EM>. Depois disso, o satélite deixou de ter sua órbita corrigida. (...)

Frase = "Depois disso, o satélite deixou de ter sua órbita corrigida."



Conclusões e Trabalho Futuro

- A simplicidade e a rapidez do sistema foram conseguidas na participação do HAREM
- Os resultados excederam as expectativas, tratando-se de um sistema ainda prematuro
  - Participação somente na categoria TEMPO
  - Precisão ≈ 70%
  - Abrangência ≈ 55%
  - Tempo de execução ≈ 00:03:20

Conclusões e Trabalho Futuro

- Ainda há muito para fazer!
  - Alargar o processamento a expressões temporais complexas (com mais de uma referência temporal), de modo a melhorar a abrangência do sistema
  - Ultrapassar o limite de n=2, na determinação das co-ocorrências
  - Tornar mais automatizada a criação dos padrões das expressões temporais
  - Testar o sistema em outras línguas

Obrigada.

olga.craveiro@gmail.com