

**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

ADRIANO ARAÚJO SANTOS

RISO-TT – EXTRAÇÃO DE EXPRESSÕES TEMPORAIS EM TEXTOS



**CAMPINA GRANDE, PARAÍBA, BRASIL.
22 DE ABRIL DE 2013**

ADRIANO ARAÚJO SANTOS

RISO-TT – EXTRAÇÃO DE EXPRESSÕES TEMPORAIS EM TEXTOS

Dissertação submetida à Coordenação do Curso de Pós-graduação em Ciência da Computação, da Universidade Federal de Campina Grande – Campus I - como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Ulrich Schiel

Área de concentração: Ciência da Computação

Linha de Pesquisa: Recuperação de Informação

**CAMPINA GRANDE, PARAÍBA, BRASIL.
22 DE ABRIL DE 2013**

DIGITALIZAÇÃO:
SISTEMOTECA - UFCG

FICHACATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

S237r Santos, Adriano Araújo.
Riso-TT – Extração de expressões temporais em textos / Adriano Araújo Santos. – Campina Grande, 2013.
84 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2013.

"Orientação: Prof. Dr. Ulrich Schiel.
Referências.

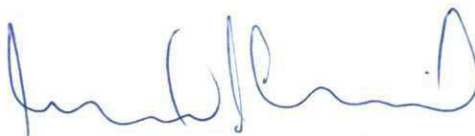
1. Extração Temporal. 2. Reconhecimento de Padrões Temporais.
3. Processamento de Linguagem Natural. I. Schiel, Ulrich.
II. Título.

CDU 004:681.5(043)

"RISO-TT EXTRAÇÃO DE EXPRESSÕES TEMPORAIS EM TEXTOS"

ADRIANO ARAUJO SANTOS

DISSERTAÇÃO APROVADA EM 24/05/2013



ULRICH SCHIEL, Dr., UFCG
Orientador(a)



CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Examinador(a)



ED PORTO BEZERRA, D.Sc, UFPB
Examinador(a)

CAMPINA GRANDE - PB

RESUMO

A necessidade de gerenciar a grande quantidade de documentos digitais existentes na atualidade, associada à incapacidade humana de analisar todas essas informações em tempo hábil, fez com que as pesquisas e o desenvolvimento de sistemas na área de automatização de processos para a gestão de informação crescessem, no entanto, essa atividade não é trivial. A maioria dos documentos disponíveis não tem estrutura bem definida (padronizada), o que torna difícil a criação de mecanismos computacionais que automatizem a análise das informações e gera a necessidade de se promoverem atividades intermediárias de conversão de informações em linguagem natural em informações estruturadas. Para isso, são necessárias atividades de reconhecimento de padrões nominais, temporais e espaciais. No tocante a essa pesquisa, o objetivo principal foi criar um mecanismo de reconhecimento de padrões temporais.

Heuristicamente, foi criado um dicionário de regras baseadas em associações de padrões temporais e desenvolvido um módulo de marcação e recuperação de padrões temporais em uma arquitetura extensível e flexível, chamado RISO-TT, que implementa esse mecanismo de reconhecimento de padrões temporais. Esse módulo faz parte do projeto de pesquisa RISO (Recuperação da Informação Semântica de Objetos Textuais). Foram realizados dois experimentos para avaliar a eficiência do RISO-TT. O primeiro, com o intuito de verificar a extensibilidade e a flexibilidade do módulo RISO-TT, e o segundo, para analisar a eficiência da abordagem proposta com base em uma comparação com duas ferramentas consolidadas no meio acadêmico (HeidelTime e SuTime). O RISO-TT obteve resultados superiores aos concorrentes no processo de marcação de expressões temporais, comprovados por meio de testes estatísticos.

Palavras-chave: extrator temporal; reconhecimento de padrões temporais; processamento de linguagem natural.

ABSTRACT

The necessity of managing the large amount of digital existing documents nowadays, associated to the human inability to analyze all this information in a fast manner, led to a growth of research in the area of system development for automation of the information management process. Nevertheless, this is not a trivial task. Most of the available documents do not have a standardized structure, hindering the development of computational schemes that can automate the analysis of information, thus requiring jobs of information conversion from natural language to structured information. For such, syntactic, temporal and spatial pattern recognition tasks are needed. Concerning the present study, the main objective is to create an advanced temporal pattern recognition mechanism. We created, heuristically, a rules dictionary of temporal patterns, developing a module in an extendable and flexible architecture for retrieval and marking. This module, called RISO-TT, implements this pattern recognition mechanism and is part of the RISO project (Retrieval of Semantic Information from Textual Objects). Two experiments were carried out in order to evaluate the efficiency of this approach. The first one was intended to verify the extendability and flexibility of the RISO-TT architecture and the second one to analyze the efficiency of the proposed approach, based on a comparison between the developed module and two consolidated tools in the academic community (Heideltime and SuTime). RISO-TT outperformed the rivals in the temporal expression marking process, which was proved through statistical tests.

Keywords: temporal extractor; temporal pattern recognition; natural language processing.

AGRADECIMENTOS

O Senhor é o protetor da minha vida, de quem terei medo?

Só sabe o valor de um Amigo quem tem Um. O meu maior amigo é Deus. Ele sempre esteve presente em minha vida. Segurou minha mão, quando estive perdido sem saber para onde ir. Segurou meu corpo, quando tive fome, sem força, e não tinha o que comer. Segurou minhas lágrimas, quando a dor foi mais forte do que pude segurar. Teve paciência... me ouviu... me orientou... me ajudou. Foi determinístico, quando perdi a minha mãe (no dia das mães de 2011) no momento em que estava pagando as disciplinas do Mestrado. E soube suprir a mim e a minha família, quando passei meses sem receber salário na empresa em que trabalhei. Portanto, com Ele, do que terei medo?

Se existisse alguma palavra de agradecimento que representasse o amor que tenho para com minha mãe (Dona Marta) e meu pai ("Seu" Lula), eu até que escreveria aqui... mas, como agradecer a quem sempre foi exemplo em sua vida, motivação e força para lutar? Poucas pessoas sabem, mas somos de origem muito humilde e passamos por muitas provações... Mas sempre superamos. Sempre! Vejam vocês: muitos pensavam que eu seria um marginal. Hoje, estou defendendo o Mestrado e já estou cursando o Doutorado. Meu objetivo sempre foi que meus pais se orgulhassem de mim. Tenho certeza de que consegui.

Agradeço pelo apoio e pela credibilidade que recebo da minha namorada, Raelma Patriota, que tem aguentado os meus abusos constantes, pois, como vocês sabem, pesquisa é dedicação.

Tenho irmãos de sangue e irmãos por escolha. Não poderia deixar de agradecer a quem sempre esteve comigo, mas não tenho como citar todos. Sendo assim, vou listar alguns que representarão todos: Luciene A. Santos, Marta Milene, Alan Robson, Danilo Abreu, Larissa Lynda, Ermison Sousa, Joanna Marques, Diego Loureiro, Zé Gildo, Manoel Neto, Roberta Macedo, Tiago Silva (Baiano), Nayara, Otacílio Lacerda, Isabel Nunes, Vera Medeiros e Tércio. Obrigado, meus irmãos! Agradeço aos mestres: Cláudio Baptista, Jacques Sauv e, Joseana Fechine e Ant o Moura. Conhecimento e oportunidade s o coisas impag veis. Obrigado!

Por fim, e n o menos importante, agradeço ao GRANDE MESTRE Ulrich Schiel. Primeiro, por ter acreditado em mim no in cio de tudo isso. Porque vim para a UFCG como um estrangeiro, e ele me deu a oportunidade de adentrar o Mestrado. Segundo, por todas as experi ncias vivenciadas ao longo da minha forma o inicial como cientista. E por fim, pela paci ncia e o bom humor de sempre. Valeu, MESTRE!

GLOSSÁRIO

ATEL - Automatic Temporal Expression Labeler

DARPA - Defense Advanced Research Projects Agency

DSTO - Defence Science and Technology Organization

ET - Extrator Temporal

IEEE - The Institute of Electrical and Electronics Engineers

PLN - Processamento de Linguagem Natural

REM - Reconhecimento de Entidades Mencionadas

RISO - Recuperação da Informação Semântica de Objetos Textuais

TERN - Time Expression Recognition and Normalisation

TERSEO - Temporal Expression Recognition System applied to Event Ordering

TIDES - Temporal Annotation Guidelines

TimeML - Markup Language for Temporal and Event Expressions

LISTA DE FIGURAS

Figura 1: Processamento de informação no Chronos	25
Figura 2: Processamento de informação no TERSEO.....	28
Figura 3: Estrutura geral do RISO	34
Figura 4: Processos do RISO-VTD	36
Figura 5: Módulos do RISO-IS	37
Figura 6: Módulos de consulta RISO-CS	39
Figura 7: Arquitetura do RISO-TT	42
Figura 8: Histograma dos dados do Heideitime	56
Figura 9: Histograma dos dados do RISO-TT	57
Figura 10: Histograma dos dados do SUTime	58

LISTA DE TABELAS

Tabela 1: Tabela de exemplos de padrões temporais	43
Tabela 2: Tabela de exemplos de regras	43
Tabela 3: Tabela de <i>tokens</i> ideais, <i>tokens</i> checados e <i>tokens</i> encontrados	54
Tabela 4: Tabela de valores de precisão, cobertura e <i>F-measure</i>	55
Tabela 5: Tabela de testes de normalidade	56
Tabela 6: Tabela de intervalos de confiança	57

*“Devemos partir do princípio de que a passagem
do tempo é algo totalmente constante e imutável.
A única coisa que podemos fazer é aproveitá-lo
bem ou mal...”*

Ulrich Schiel

SUMÁRIO

	CAPÍTULO 1 – INTRODUÇÃO	11
1.1	OBJETIVOS	13
1.2	MOTIVAÇÃO	14
1.3	METODOLOGIA	14
1.3.1	Caráter da Pesquisa	14
1.3.2	Estudo Experimental	15
1.4	ESTRUTURA DA DISSERTAÇÃO	15
	CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA	17
2.1	PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)	17
2.2	EXTRAÇÃO DE INFORMAÇÃO	18
2.3	EXPRESSÕES TEMPORAIS (ET)	20
2.3.1	Normalização de expressões temporais	21
2.3.2	Esquemas de anotação de expressões temporais	22
2.3.2.1	Translingual Information Detection, Extraction and Summarization (TIDES)	22
2.3.2.2	TimeML	23
2.3.2.3	TIMEX	23
2.3.3	WikiWars	24
	CAPÍTULO 3 – TRABALHOS RELACIONADOS	25
3.1	ATEL	25
3.2	CHRONOS	25
3.3	TempEx	26
3.4	GUTime	27
3.5	DANTE	27
3.6	TERSEO	28
3.7	TimexTag	29
3.8	TIPSem	30
3.9	HEIDELTIME	31
3.10	SUTime	31
3.11	CONSIDERAÇÃO	32
	CAPÍTULO 4 – O PROJETO RISO	35
4.1	ESTRUTURA GERAL	35

4.2	CRIAÇÃO DE VOCABULÁRIOS TEMÁTICOS (RISO-VTD)	36
4.3	INDEXAÇÃO (RISO-IS)	38
4.4	CONSULTAS SEMÂNTICAS (RISO-CS)	39
4.5	CONSIDERAÇÕES FINAIS SOBRE O PROJETO RISO	40
	CAPÍTULO 5 – RISO TEMPORAL TAGGER (RISO-TT)	41
5.1	APRESENTAÇÃO	41
5.2	ARQUITETURA	45
5.3	ESTRUTURAS RESULTANTES DO PROCESSAMENTO DO RISO-TT	46
5.4	PSEUDOCODIGO DO RISO-TT	46
5.5	FORMALIZAÇÃO DO RISO-TT	47
5.5	EXEMPLO DE UM DOCUMENTO PROCESSADO NO RISO-TT	49
	CAPÍTULO 6 – VALIDAÇÃO E VERIFICAÇÃO	53
6.1	ROTEIRO DO ESTUDO EXPERIMENTAL	53
6.1.1	Seleção das variáveis	54
6.1.2	Design de experimento	55
6.2	VERIFICAÇÃO	56
6.3	VALIDAÇÃO	57
6.3.1	Análise dos dados	57
6.3.2	Análise dos dados	58
6.3.3	Ameaças à validade	59
	CAPÍTULO 7 – CONCLUSÕES	65
	REFERÊNCIAS	69
	ANEXO A – Exemplo de documento da WikiWars	72
	ANEXO B – Arquivo de configuração do RISO-TT	77
	ANEXO C – Exemplos de padrões mapeados para o RISO-TT	78
	ANEXO D – Arquivo de regras do RISO-TT	82
	ANEXO E – Exemplos de arquivos processados pelo RISO-TT	85

CAPÍTULO 1

INTRODUÇÃO

A grande quantidade de documentos digitais existentes na atualidade, fruto do acesso irrestrito e da liberdade de publicação fornecida às pessoas, por meio da Web (BAEZA-YATES; RIBIERO-NETO, 1999), fez com que a capacidade humana de analisar todas as informações existentes fosse superada. Com isso, cresceu a necessidade de automatizar o acesso, a pesquisa e a gestão da informação a fim de gerar valiosas fontes de conhecimentos (LIORENS, 2011).

Um ser humano, geralmente, entende seu idioma nativo com pouco (ou nenhum) esforço consciente e, para isso, basta que conheça as regras do idioma aliado ao conhecimento do mundo para entender um texto (TIDES, 2005). Os computadores, por sua vez, podem processar informações estruturadas ou semiestruturadas. E como a maior parte das informações existentes e disponíveis não é estruturada (LIORENS, 2011), o desafio atual é permitir que computadores processem informações em linguagem natural e converta-as em informação estruturada, para que seja possível o maior nível de automatização de processos computacionais.

Assim, o Processamento da Linguagem Natural (PLN) surge como uma possível solução para esse desafio, pois se caracteriza como um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua. Dentre essas técnicas, existe o Reconhecimento de Entidades Mencionadas (REM) (FERNEDA, 2011), cujo objetivo é de localizar e classificar elementos atômicos em um texto, de acordo com um conjunto predefinido de categorias.

A extração de informação (EI) é a tarefa de encontrar informações a partir de grandes volumes de documentos ou textos, estruturados ou livres (ZAMBENEDETTI, 2002). Para Zambenedetti (2002), uma tecnologia de extração de informação bem desenvolvida permitiria criar rapidamente sistemas de extração para novas tarefas cujo desempenho teria o mesmo nível de tarefas realizadas por um humano – esse nível ainda não foi alcançado.

Essa dissertação se destina ao processo de extração de expressões temporais de textos, uma atividade que se tornou um grande campo de pesquisa e

desenvolvimento dentro da Ciência da Computação, motivada pelo grande número de aplicações que exploram informações temporais extraídas de textos. Como exemplos dessas aplicações, podem ser citados aplicações de perguntas e respostas automáticas e sistemas de sumarização de textos. Com o uso das técnicas de extração de expressões temporais, aplicativos executam atividades em um nível mais complexo de automação (SAQUETE, 2010).

Segundo Schiel (1996), a pesquisa sobre o tempo, em Sistemas de Informação, pode ser dividida em: a) Modelagem temporal, que estuda os tempos quando objetos existem ou eventos ocorrem em um sistema e as relações temporais existentes entre ele; b) Modelagem comportamental, que estuda a descrição e a representação de eventos, bem como a sua influência nos estados de um sistema, permite que se preveja a ocorrência de ações e de reações futuras, automatizando, ao máximo, o funcionamento dos Sistemas de Informação, e reduz a interferência humana.

Existem diversas abordagens para o reconhecimento de expressões temporais em textos de acordo com os recursos disponíveis para o idioma ao qual ela se aplica e com os requisitos do sistema (SAQUETE, 2010). Saquete (2010) classifica como principais abordagens: a) Baseadas em regras; b) Aprendizagem de máquina e c) Combinação entre regras e máquina de aprendizagem. Porém, independentemente da abordagem adotada, a saída do processamento dos textos será um esquema de anotações temporais padronizados. Os esquemas TIDES 2005 (MANI et al., 2001) e *TimeML* (PUSTEJOVSKY et al., 2003b) são os mais adotados.

Com base nas informações apresentadas, foi realizada uma avaliação das ferramentas de marcação de expressões temporais e técnicas de extração de expressões que atendessem aos requisitos do projeto RISO (Recuperação da Informação Semântica de Objetos Textuais) de:

- a) Estender as funcionalidades de extração de sintagmas nominais já implementadas no módulo *RISO-Extractor* (BISPO, 2012) para reconhecer expressões temporais e que fosse desenvolvido na linguagem *Python*;
- b) Ter arquitetura flexível a regras e extensível de forma transparente e de fácil adaptação a novos padrões e regras;

- c) Reconhecer expressões temporais compostas, tais como: a) intervalos temporais abertos, semiabertos e fechados; b) expressões explícitas e implícitas; c) associações entre expressões regulares e sintagmas nominais e d) qualquer nova expressão encontrada nos textos e que ainda não foi mapeada e;
- d) Prover independência de ferramentas ou *software* de terceiros.

Entre as ferramentas avaliadas, não foi possível identificar uma que tivesse as características necessárias para o projeto no qual essa pesquisa está inserida (Projeto *RISO*). Por isso, foi necessário desenvolver um extrator de expressões temporais, intitulado *RISO Temporal Tagger* (RISO-TT) para suprir as necessidades elencadas. Ademais, foi realizado um experimento para comprovar as características de extensibilidade e flexibilidade do sistema, se o RISO-TT apresentava alguma vantagem competitiva e se trazia contribuições para a área pesquisada.

Com a análise dos resultados verificamos que o RISO-TT é, de fato, o extrator de expressões temporais flexível e extensível e, ao compará-lo com os *Heideltime* e *SUTime* (extratores temporais conceituados na literatura) observou-se, por meio de testes estatísticos, que o RISO-TT obteve melhor resultado em seu processo de marcação.

1.1 OBJETIVOS

O objetivo principal dessa dissertação foi criar um dicionário de padrões temporais e regras associativas extensíveis, baseados em heurística, para o reconhecimento de expressões temporais em textos, culminando com o desenvolvimento de um módulo de recuperação e marcação de expressões temporais extensível e flexível, que implementa as regras propostas.

Para isso, foram elencados os seguintes objetivos específicos:

1. Criar padrões e regras com base em heurísticas para o reconhecimento de expressões temporais;
2. Desenvolver um Extrator Temporal flexível e extensível que reconhece sintagmas temporais em textos, baseado nas regras e padrões criados;

3. Criar um normalizador temporal que converta as expressões temporais extraídas em intervalos correspondentes;
4. Avaliar a extensibilidade dos padrões e regras;
5. Avaliar o Extrator Temporal.

1.2 MOTIVAÇÃO

O reconhecimento de expressões temporais é uma subatividade do processo de Reconhecimento de Entidades Mencionadas e o reconhecimento dessas informações pode ser utilizado em diversos sistemas computacionais, tais como: bancos de dados textuais, tradutores automáticos, indexação e recuperação de informação, sistemas de informação geográfica e sistemas especialistas. O desenvolvimento do RISO-TT atende a uma necessidade de processamento de expressões temporais na indexação de documentos digitais do projeto RISO.

1.3 METODOLOGIA

Os passos realizados para a pesquisa e a solução do problema exposto estão a seguir.

1.3.1 CARÁTER DA PESQUISA

O presente trabalho tem caráter de pesquisa aplicada quanto à sua natureza, pois assume o desenvolvimento de uma solução a partir da criação de um dicionário de expressões temporais com base em conhecimento heurístico.

Foi definida uma abordagem quantitativa para coleta de dados, onde foram avaliados a quantidade de expressões mapeadas dos documentos da Corpora WikiWars.

Esta pesquisa também é caracterizada como exploratória, pois um dos objetivos é expor os problemas encontrados nas ferramentas disponíveis.

1.3.2 ESTUDO EXPERIMENTAL

Foram realizados dois experimentos científicos. O primeiro, para a verificação da extensibilidade e flexibilidade do RISO-TT e o segundo, com relação ao processo de comparação entre extratores temporais.

Para o primeiro experimento, foram realizados três ensaios sobre a corpora WikiWars com três versões diferentes de configuração das regras, e o ajuste realizado em cada versão foi baseado em padrões não encontrados nos resultados do ensaio.

Para o segundo experimento, foram analisadas as características das ferramentas e a disponibilidade de cada uma, foram selecionadas as ferramentas HeideTime e SUTime, por se tratar de ferramentas com características similares ao RISO-TT e estarem disponíveis (seja para acesso on-line ou para a instalação e a configuração da ferramenta *in loco*).

Para cada ferramenta, foram realizadas as seguintes atividades:

- a) *Download* do aplicativo;
- b) Busca por artigos científicos e informações da especificação do desenvolvimento;
- c) Instalação no ambiente de teste;
- d) Estudo de funcionalidades das ferramentas;
- e) Processamento dos documentos da WikiWars;
- f) Obtenção dos resultados do processamento;
- g) Comparação dos resultados dos documentos processados com o gabarito proposto pela Corpora;

1.4 ESTRUTURA DA DISSERTAÇÃO

No capítulo 2 é apresentada a fundamentação teórica para essa pesquisa, considerando-se, desde a grande área de Processamento de Linguagem Natural (PLN), até a necessidade do reconhecimento de expressões temporais. O capítulo 3 apresenta um conjunto de ferramentas e técnicas relacionadas à área de reconhecimento de expressões temporais. Nos capítulos 4 e 5, são apresentados o

projeto RISO (Recuperação da Informação Semântica de Objetos Textuais) como um todo e o detalhamento do RISO-TT, ferramenta desenvolvida com a finalidade de implementar a proposta dessa dissertação. Os experimentos, a verificação e a validação são apresentados no capítulo 6. Por fim, no capítulo 7, as conclusões obtidas com os resultados dos experimentos, bem como os trabalhos futuros.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

Esse capítulo apresenta um conjunto de informações que possibilitam o melhor entendimento da área pesquisada e a motivação do desenvolvimento dessa pesquisa.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O Processamento de Linguagem Natural (PLN) é um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua e que se tornou uma área fortemente pesquisada na atualidade (FERNEDA, 2011). Uma de suas principais motivações está em aperfeiçoar a relação homem-máquina, utilizando conceitos de Inteligência Artificial, Teoria da Computação, Compiladores e Linguística Computacional (OLIVEIRA, 2012).

Oliveira (2012) classifica as aplicações da PLN em: a) baseadas em textos, sistemas que procuram documentos específicos em uma base de dados, tradutores de documentos e sistemas que resumem textos e b) baseadas em diálogos, interfaces de linguagem natural para os bancos de dados, os sistemas tutores e os sistemas que interpretam e respondem a comandos expressados em linguagem escrita ou falada.

No tocante às etapas do PLN, Oliveira (2010) as define como:

- a) **Análise morfológica:** identifica palavras e/ou expressões específicas em uma sentença, utilizando-se de delimitadores. As palavras identificadas são classificadas de acordo com seu tipo de uso ou categoria gramatical;
- b) **Análise sintática:** constrói árvores de derivação para cada sentença com as relações existentes entre suas palavras. Para isso, o analisador sintático faz uso de gramáticas regulares, livres de contexto ou sensíveis ao contexto.

- c) **Análise semântica:** determina o sentido das palavras que foram reagrupadas pelo analisador sintático. Essa etapa apresenta grandes desafios, como o tratamento da ambiguidade de estruturas complexas - anáforas, elipses, zeugma, metáfora e metonímia.

2.2 EXTRAÇÃO DE INFORMAÇÃO

Dentre as atividades mais relevantes do PLN, existe a Extração de Informação (EI). Essa atividade propõe localizar padrões específicos de dados em textos, a fim de extrair informações estruturadas e relevantes (KUSHMERICK; THOMAS, 2003).

Segundo Fernandes (2011), os sistemas de EI têm dois domínios específicos: um dicionário de padrões de extração e um dicionário semântico. O dicionário de padrões pode ser gerado manual ou automaticamente, e o dicionário semântico, quase sempre, é construído manualmente, por causa do seu vocabulário específico.

O processo de EI, geralmente, é confundido com o processo de Recuperação da Informação (RI), no entanto, eles são processos distintos. Bispo (2012) refere que o processo de EI está relacionado à extração de trechos relevantes de um documento, e a RI seleciona documentos relevantes em um conjunto de documentos de acordo com parâmetros de consulta.

Melo et al (2000) apresentam a estrutura de um documento como:

- a) **Estruturado:** dados são organizados em blocos semânticos (entidades), entidades similares são mantidas de forma agrupada (relações ou classes), entidades de um mesmo grupo têm as mesmas descrições (atributos), e as descrições para todas as entidades de um grupo (esquema) apresentam o mesmo formato, o mesmo tamanho, estão todas presentes e seguem a mesma ordem; (Ex: tabelas e listas)
- b) **Semiestruturado:** os dados são organizados como entidades semânticas, entidades similares são mantidas de forma agrupada, entidades em um mesmo grupo podem não ter os mesmos atributos, a ordem dos atributos não é, necessariamente, importante, nem todos os atributos podem ser obrigatórios; o tamanho e o tipo de um atributo podem variar dentro de um

mesmo grupo, os dados são organizados como entidades semânticas, entidades similares são mantidas de forma agrupada, entidades em um mesmo grupo podem não ter os mesmos atributos, a ordem dos atributos não é, necessariamente, importante, nem todos os atributos podem ser obrigatórios, e o tamanho e o tipo de um atributo podem variar dentro de um mesmo grupo.

- c) Não estruturados: dados podem ser de qualquer tipo, não seguem, obrigatoriamente, nenhum formato ou sequência, não seguem regras e não são previsíveis (Ex: documentos quaisquer).

Bispo (2012) apresenta os seguintes tipos de abordagem adotados pelo processo de EI:

- a) Baseada em dicionário: Utiliza uma lista de termos para identificar ocorrências no texto, relacionando os termos existentes nos dicionários com os encontrados no texto;
- b) Baseada em regras: Utiliza regras (baseadas em heurísticas ou não) para a extração de informações sobre o domínio desejado;
- c) Baseada em aprendizagem de máquina: Utiliza o processo de automatização de atividades de processamento de informações para obter regras a serem usadas em um novo domínio.

O Reconhecimento de Entidades Mencionadas (REM) (ROMÃO, 2007) é uma subatividade da área de extração de informação, que tem o objetivo de localizar e classificar elementos atômicos em textos, como: nomes de pessoas, valores monetários, lugares e expressões temporais (tema abordado nesta dissertação). Essa atividade é utilizada frequentemente em aplicações que realizam perguntas e respostas de forma automática e para se obterem informações estruturadas em documentos não estruturados.

Cardoso (2006) afirma que a atividade de REM não é trivial, já que as EMs podem ser vagas (desconexas com o contexto) ou ambíguas (mais de um sentido), uma característica intrínseca à língua humana, razão por que é necessário analisar

o contexto em que a EM se insere a fim de definir corretamente sua categoria semântica.

Romão (2007) classifica as estratégias utilizadas no REM em: a) independentes de língua e b) dependentes de língua. A diferença básica entre as duas abordagens é de que a estratégia REM dependente de língua utiliza recursos específicos da língua e não é aplicável a outra língua.

Os sistemas que são independentes de língua se baseiam na análise automática das cadeias de caracteres que compõem um texto sem utilizar listas de palavras nem qualquer informação sobre a segmentação das partes do discurso. Esses sistemas utilizam três características distintas: a) TIMEX (frases e expressões temporais), b) NUMEX (frases e expressões numéricas) e c) ENAMEX (nomes próprios, locais e organizações) e faz uso de *Corpora* (conjunto de documentos semanticamente organizados) de vários idiomas (ROMÃO, 2007).

Os sistemas que dependem do idioma utilizam características linguísticas tais como lexicais (com conhecimento sobre parte do discurso do texto) ou regras gramaticais e de contexto específico para obter melhores resultados no processamento da informação.

2.3 EXPRESSÕES TEMPORAIS (ET)

Biber et al. (1999) definem como uma expressão temporal (ET) qualquer expressão que denote:

- um acontecimento com dimensão temporal;
- um objeto com existência temporal ou
- qualquer expressão textual que denomina informações do tipo datas do calendário, horários do dia, períodos de tempo, durações, intervalos.

Biber et al. (1999) afirmam que as ET transmitem diferentes tipos de informação relacionada com o tempo, como posição, frequência, duração e relacionamento. Marsic (2011) enfatiza que as ET detêm uma vasta gama de relações gramaticais e, geralmente, são sinalizadas por uma ou mais palavras de tempo, chamadas de gatilhos lexicais, a saber: a) substantivos: Século, ano, mês,

dia, fim de semana, minutos, futuro, passado; b) nomes próprios: Brasil, Itália, Nova Iorque, Luís XIV, St. Antônio etc.; c) adjetivos: passado, presente, futuro, próximo, medieval, mensais; d) advérbios: atualmente, então, semanal, hoje, ontem, amanhã, essa noite; e) padrões de tempo: 09h00min, 26/12/2002; f) números: 4 (João chegou às 4h).

2.3.1 Normalização de expressões temporais

O processo de normalização é uma atividade importante para o processamento de informações temporais, porque é responsável pela transformação das expressões temporais em valores (ZHAO; JIN; YUE, 2010). Segundo Hagège, Baptista e Mamede (2010), a normalização das ET consiste em representar o valor de uma expressão temporal de uma forma que permita a realização de cálculos. Para exemplificar a normalização de uma expressão temporal, imagine a expressão temporal "01 de maio de 2012", que pode ser normalizada no formato TIMEX2, como: `<TIMEX2 VAL = "2012-05-01"> 01 de maio de 2012</TIMEX2>`.

Hagège, Baptista e Mamede (2010) concluem que o processo de normalização de expressões temporais é a ação de percorrer todos os nós das entidades temporais e converter para o formato adequado, levando em consideração os seguintes casos especiais:

- a) Expressões com elementos em numeração romana (Ex: XX, XV);
- b) Expressões que envolvem unidade de tempo não representada nos padrões normais (Ex: semana, trimestre e quinzena);
- c) Expressões com frações de tempo (Ex: dois dias e meio);
- d) Expressões que indiquem informalidade para indicar hora (Ex: 10 pras 6h);
- e) Expressões não numéricas (Ex: duas horas da tarde);
- f) Expressões com advérbios temporais (Ex: ontem, amanhã etc.);
- g) Expressões com marcação em valores relativos (Ex: No próximo mês).

2.3.2 Esquemas de anotação de expressões temporais

Nessa subseção, apresentam-se os esquemas de anotação de expressões temporais mais comuns utilizados pelas ferramentas de marcação de expressões temporais.

2.3.2.1 Translingual Information Detection, Extraction and Summarization (TIDES)

Trata-se de um projeto desenvolvido pela DARPA (*Defense Advanced Research Projects Agency*), com o objetivo principal de desenvolver tecnologias capazes de ajudar pessoas que utilizam o idioma inglês a se comunicarem com pessoas de idiomas diferentes (TIDES, 2002; MARSIC, 2011). Para tanto, eles investem em projetos que possibilitem avanços reais nas áreas de detecção, extração, resumos e tradução de forma automática.

Dentre os projetos mantidos pela TIDES (TIDES, 2002), existe um conjunto de práticas destinadas aos trabalhos de extração temporal e normalização dessas expressões e princípios direcionados a: a) Marcadores humanos, responsáveis por criar corporas temporais com base em anotações das expressões temporais, e b) Programadores de Sistemas, que buscam desenvolver sistemas de marcação automática de expressões temporais. Loureiro (2007) classifica o TIDES como um guia que reúne um conjunto de padrões de anotação que tem como finalidade identificar e normalizar expressões temporais de forma a serem utilizadas em sistemas que requeiram a automatização da comunicação.

Os princípios da marcação temporal do TIDES (FERRO et al., 2005) classificam as expressões temporais em:

- a) Precisas: expressões que representam uma data de calendário (Ex: *October 3, 1990*);
- b) Imprecisas: expressões vagas (Ex: *the last year*);
- c) Modificadas: expressões quantificadas (Ex: *approximately*);
- d) Frequências: expressões que representam a frequência com que algo ocorre (Ex: *every day*);
- e) Não específicas: expressões que não se referem a um tempo específico.

Com relação à classificação, existem dois tipos de expressões:

- a. Genéricas: expressões que especificam uma classe de entidades temporais no lugar de um tempo específico (Ex: *I like october*);
- b. Indefinidas: expressões indefinidas (Ex: *on a monday*).

2.3.2.2 TimeML

O TimeML é uma especificação empregada para a análise de eventos e de expressões temporais em linguagem natural (TIMEML, 2012). O TimeML foi projetado para resolver quatro problemas em evento e marcação de expressões temporais: a) imprecisão da data de eventos; b) ordenação de eventos em relação um ao outro; c) raciocínio contextualmente subespecificado com expressões temporais e d) raciocínio sobre a persistência de eventos. Ele é responsável por especificar os padrões TIMEX¹.

O TimeML é um trabalho que se refere ao processo de identificação de expressões temporais, pois não se propõe apenas a classificar e normalizar as expressões temporais, mas também a marcar informações dos processos relacionados a tempo, aspecto e modalidade (HAGÈGE; BAPTISTA; MAMEDE, 2010).

2.3.2.3 TIMEX

O padrão TIMEX (TIMEML WORK GROUP, 2009) é usado para marcar expressões temporais, tais como: *16h00min*, *outubro* e *na semana passada*. Trata-se de um esquema para classificar expressões temporais. Os primeiros sistemas de anotação de expressões temporais incluem TIMEX e TIMEX2. Porém a versão atual é chamada de TIMEX3.

¹ TimeX: <http://www.timexportal.info/>

Os exercícios para evolução da área são realizados no *SemEval*² (*International Workshop on Semantic Evaluations*), com o objetivo de avançar a pesquisa sobre o processamento da informação temporal (MARTÍNEZ, 2011).

2.3.3 WikiWars

O processamento de informações temporais tem se tornado comum em muitas aplicações da PLN, e as atividades de reconhecimento de padrões temporais e de normalização tem conquistado um espaço significativo, fazendo com que corporas com marcações temporais fossem criadas (MAZUR, 2010). No entanto, as corpora existentes na literatura apresentam limitações relacionadas ao comprimento dos textos e sua estrutura, o que impacta diretamente na variedade das expressões temporais (MAZUR, 2010).

Baldwin (2002), Ahn et al. (2005) e Mazur e Dale (2010) sugerem que muitas expressões temporais em documentos, principalmente em notícias, podem ser interpretadas como simples relações com a data de criação do documento e que, conseqüentemente, existem corporas que não são ideais para o desenvolvimento de sistemas cuja intenção seja de trabalhar com narrativas históricas ou que necessitem do reconhecimento de fatos.

Com base nas informações apresentadas e tendo em vista a evolução dos sistemas de extração de expressões temporais, conforme será apresentado na seção de trabalhos futuros, a corpora escolhida para o experimento do *RISO-TT* foi a *WikiWars*. Ela consiste em 22 documentos extraídos da *Wikipedia*, que descrevem o percurso histórico das guerras, no idioma inglês, com um total de 120,000 *tokens* e 2.681 expressões temporais marcadas no padrão TIMEX2.

A *WikiWars* foi apresentada por Mazur e Dale (2010), que justificam o pequeno número de documentos com a qualidade do processo de construção da mesma que considera narrativas longas, que dão origem a fenômenos temporais mais complexos do que são encontrados em simples documentos. Os autores ainda afirmam que o próprio formato sugere um novo desafio para as ferramentas do seguimento, devido à grande quantidade de expressões temporais existentes em cada documento.

² SemEval: <http://www.cs.york.ac.uk/semeval-2013/>

CAPÍTULO 3

TRABALHOS RELACIONADOS

Nesse capítulo, são apresentados, em ordem alfabética, os trabalhos relacionados a sistemas de marcação de expressões temporais.

3.1 ATEL

Desenvolvido no Centro de Pesquisa Linguagem e Educação Computacional (anteriormente conhecido como o Centro de Pesquisa de Língua Falada), da Universidade do Colorado, o sistema ATEL (*Automatic Temporal Expression Labeler*) (HACIOGLU et al. 2005) adota o modelo da cadeia de Markov para detectar expressões temporais nos idiomas inglês e chinês.

Esse sistema utiliza uma base de treinamento disponibilizado pelo TERN³ (*Time Expression Recognition and Normalisation*) com um conjunto de termos temporais, e para cada sentença encontrada em um documento processado por ele, o termo é marcado com as seguintes *tags*:

- a) (*) – para um *token* de expressão temporal;
- b) (* - para o início de uma expressão temporal;
- c) *) – para o final de uma expressão temporal.

3.2 CHRONOS

Desenvolvido no ITC-IRST⁴, *Centro per la Ricerca Scientifica e Tecnológica*, Povo, Itália, o sistema *Chronos* (NEGRI; MARSEGLIA, 2005) tem uma abordagem baseada em regras, que separa a identificação de expressões temporais em reconhecimento (detecção) e interpretação dos valores (normalização).

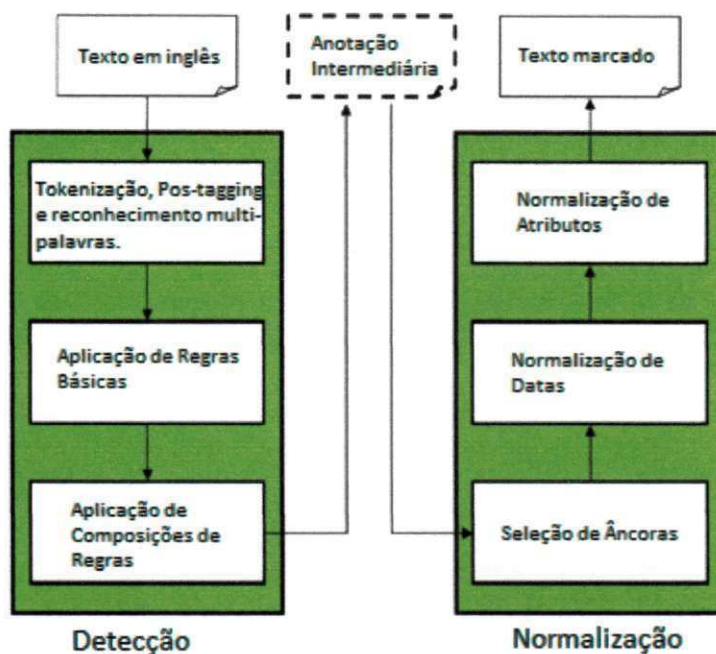
A estrutura do *Chronos* é baseada em análise linguística (tokenização, *tagging* e reconhecimento de padrões). A aplicação de uma regra gera uma marcação intermediária que contém todas as informações relevantes para a fase de

³ TERN - <http://www.lsi.upc.edu/~nlp/meaning/documentation/3rdYear/WP3.6.pdf>

⁴ ITC-IRST: <http://itc.fbk.eu>

normalização e, por fim, a fase de normalização transforma as marcações em valores definidos pelo TIMEX2. Esse processo é apresentado na figura 1.

Figura 1: Processamento de Informação no *Chronos*.



Fonte: (NEGRI; MARSEGLIA, 2005).

3.3 TempEx

Desenvolvido na MITRE Corporation⁵, com um programa Perl para o reconhecimento e a interpretação de expressões de tempo, conforme especificado nas orientações TIMEX2 2001, o TempEx se caracteriza como um dos primeiros do gênero (MANI ;WILSON, 2000). Ele reconhece tempos absolutos (Ex: “15 de março de 2013”) e relativos (Ex: Nasceu depois da II Guerra Mundial), e o cálculo de normalização é baseado na data de publicação do documento, o que significa dizer que o algoritmo utiliza a “metainformação” do próprio documento para calcular a normalização dos tempos relativos (MANI e WILSON, 2000).

O TempEx é citado frequentemente em artigos, sites e outras ferramentas, porém os links que disponibilizam o código-fonte estão “quebrados” e não foi possível executar e analisar o mesmo.

⁵ Mitre Corporation: <http://www.mitre.org/about/>

3.4 GUTime

Desenvolvido na Universidade de *Georgetown*, o GUTime (VERHAGEN et al, 2005) é uma extensão do *tagger* TempEx (MANI; WILSON, 2000). Sua função é de reconhecer e normalizar expressões temporais no padrão TIMEX3.

Uma característica importante do GUTime é que ele possibilita o deslocamento em expressões temporais, fazendo com que cálculos sejam realizados com base em uma data de entrada (VERHAGEN et al, 2005), ou seja, se uma expressão temporal for encontrada no início de uma frase, ele o calcula a normalização dessa data, e se uma nova expressão temporal for encontrada e se referir a uma semana depois da primeira, ele calcula o seu valor com base na primeira. Essa característica é muito importante para sistemas que desejam realizar análises de sequência temporal de fatos.

Outra característica importante do TempEx é que ele incorporou um conjunto de expressões ACE TIMEX2 (inexistente no TempEx), que inclui a duração, uma variedade de modificadores temporais e formatos europeus de data (VERHAGEN et al, 2005).

3.5 DANTE

O sistema DANTE (*Detection And Normalisation of Temporal Expressions*) (MAZUR; DALE, 2007) foi desenvolvido no Centro de Tecnologia da Linguagem da Universidade Macquarie, baseado no padrão TIMEX2 2005 v.1.1. Em sua fase inicial, foi financiado pelo Departamento de Defesa da Austrália, *Defense Science and Technology Organization* (DSTO)⁶. Sua arquitetura é modularizada e constituída, basicamente, por dois módulos: reconhecimento e interpretação.

O módulo de reconhecimento de expressões temporais foi desenvolvido utilizando-se a gramática JAPE (CUNNINGHAM et al, 1999), que consiste em um conjunto de regras < *condição, ação* >. Segundo Faria e Girardi (2010), uma gramática JAPE tem dois lados: o esquerdo e o direito. O lado esquerdo da

⁶ Defence Science and Technology Organization: <http://www.dsto.defence.gov.au/>

gramática contém uma expressão regular a ser encontrada, e o lado direito descreve a ação a ser tomada sobre a expressão encontrada.

O módulo de interpretação percorre frase a frase de um documento à procura de padrões existentes, de acordo com um padrão pré-definido (base de conhecimento). Esse módulo foi desenvolvido em Java e dispõe de um conjunto de funções para cálculos de datas e de horários. Apesar das informações em sites de que esse sistema está disponível sob demanda, porém o gestor do projeto não retornou ao e-mail de solicitação encaminhado.

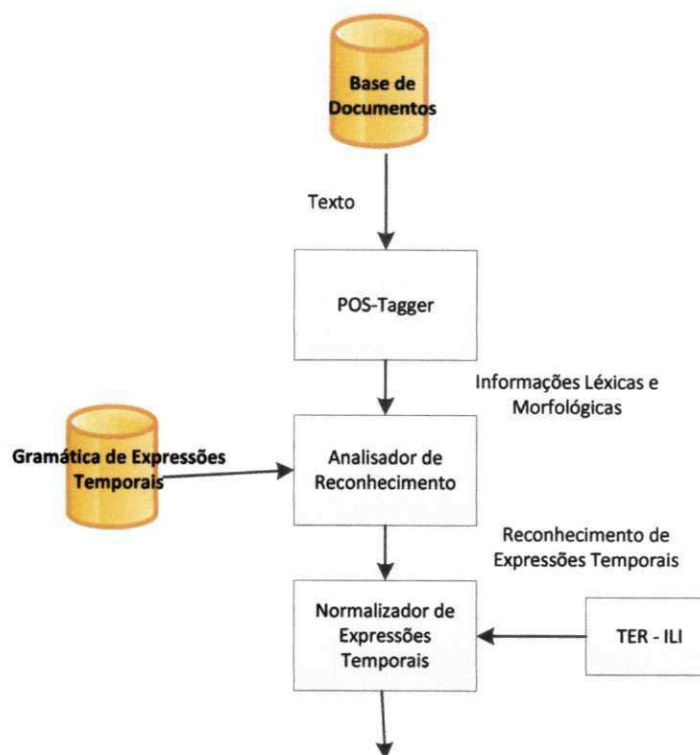
3.6 TERSEO

O TERSEO (*Temporal Expression Recognition System applied to Event Ordering*) foi desenvolvido pelo Grupo de Processamento de Linguagem Natural e Sistemas de Informação da Universidade de Alicante, parcialmente financiado pelo governo espanhol. Ele gera anotações no padrão TIMEX2, conforme especificado nas orientações para 2005 TIMEX2, inglês e espanhol (SAQUETE, 2010).

Inicialmente, segundo Saquete (2010), o TERSEO foi desenvolvido como um sistema de base de conhecimento, a fim de reconhecer automaticamente e normalizar expressões temporais em textos espanhóis.

Um documento é processado por um *Pos-Tagger*, que classifica os termos em todo o texto. Depois disso, esse documento marcado é processado por um Parser de reconhecimento de expressões, que utiliza uma gramática de expressões temporais. O processo é concluído com a normalização das expressões temporais encontradas. Este processo está apresentado na figura 2:

Figura 2: Processamento de informação no TERSEO (SAQUETE, 2010).



O TERSEO utiliza a tradução das expressões temporais como modelos temporais - já definidos na primeira versão - para obter, de forma automática, as expressões temporais de outros idiomas (SAQUETE, 2010). O sistema resultante foi chamado de MT-TERSEO e tem como objetivo final obter uma arquitetura semelhante à *Wordnet*⁷, com uma unidade central com as regras de normalização e de reconhecimento para os diferentes idiomas (que é dependente de idiomas) e relacioná-los com a unidade TERILI. Esse sistema está disponível on-line para avaliação.

3.7 TimexTag

O TimexTag foi desenvolvido na Universidade de Amsterdam – Holanda - e gera anotações no padrão TIMEX2, conforme especificado nas orientações TIMEX2

⁷ Wordnet - <http://wordnet.princeton.edu/>

2005 sob a GNU LGPL⁸. Trata-se de um sistema modular de reconhecimento e interpretação de expressões temporais no idioma Inglês (AHN; VAN RANTWIJK; DE RIJKE, 2007).

A arquitetura do TimexTag utiliza o analisador *Charniak*⁹ e o LIBSVM¹⁰, mas existe uma versão do TimexTag disponível em versão *on-line*, que dispensa a instalação do ambiente para teste. No entanto, o código disponibilizado para *download* para configurar o ambiente apresentou problemas de dependência de componentes. Um e-mail foi enviado para o administrador do projeto, mas não houve retorno.

3.8 TIPSem

Esse sistema foi desenvolvido pelo Grupo de Pesquisa em Processamento de Linguagem Natural e Sistema de Informação da Universidade de Alicante - Espanha - e está disponível nos idiomas inglês e espanhol e foi desenvolvido com base no padrão TIMEX3 (LLORENS; SAQUETE; NAVARRO, 2010).

O TIPSem lida com seis diferentes tarefas relacionadas ao tratamento de informação temporal multilíngue proposto por TempEval-2 (*Evaluating Events, Time Expressions, and Temporal Relations*)¹¹. Essas atividades são classificadas em A, B, C, D, E e F. A atividade A consiste em definir as extensões temporais; a B, na classificação de eventos definidos pela *TimeML (Markup Language for Temporal and Event Expressions)*¹², e as demais estão relacionadas à categorização de diferentes *links* temporais.

As informações sobre o projeto TIPSem estão disponíveis no site da universidade¹³, porém não foi possível realizar download do código-fonte do projeto para a realização de testes no momento do experimento.

⁸ GNU - <http://www.gnu.org/copyleft/lesser.html>

⁹ Charniak : <ftp://ftp.cs.brown.edu/pub/nlparser/>

¹⁰ LibSVM: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

¹¹ TempEval-2: <http://www.timeml.org/tempeval2/>

¹² TimeML: <http://timeml.org/site/index.html>

¹³ TIPSem: <http://gplsi.dlsi.ua.es/demos/TIMEE/>

3.9 HEIDELTIME

O *HeidelTime* é um sistema baseado em regras destinadas a apoiar vários idiomas e foi desenvolvido pelo *Institute of Computer Science* da *Ruprecht-Karls-University Heidelberg*¹⁴. Ele utiliza o padrão de anotação do TIMEX3, que faz parte da linguagem de marcação *TimeML* (com foco no atributo "value"). Existe, atualmente, a versão nos idiomas inglês e alemão.

A marcação das expressões temporais depende do domínio no qual os documentos estão inseridos. Os domínios existentes na versão mais atual são: notícias, narrativas (artigos da Wikipédia), coloquial (SMS, *tweets*) e científicos (estudos biomédicos) (STROTGEN; GERTZ, 2010).

Sua estrutura é definida por um programa de marcação temporal e pelos recursos (padrões de normalização, informações e regras). Então, criar recursos para outros idiomas é transparente (STROTGEN; GERTZ, 2010).

3.10 SUTime

O SUTime é uma biblioteca para reconhecer e normalizar expressões temporais desenvolvida pela Universidade de Stanford¹⁵. É um sistema desenvolvido em Java e baseado em regras determinista projetado para ser extensível.

Em seu desenvolvimento foi utilizado o *framework* TokensRegex¹⁶, uma estrutura genérica para definining padrões sobre o texto e mapeamento de objetos semânticos e faz uso de expressões regulares para o reconhecimento das expressões temporais (CHANG E MANNING , 2012).

Suas principais características são:

- a) Extração temporal em textos: Utiliza o padrão TIMEX3 e parte do TimeML para o processamento e marcação das expressões temporais em textos;
- b) Representação de objetos temporais como classes Java: Realiza um mapeamento das expressões tempoais em representações lógicas e

¹⁴ Institute of Computer Science Ruprecht-Karls da Univerity Heidelberg : <http://dbs.ifi.uni-heidelberg.de/index.php?id=129>

¹⁵ Stanford: <http://nlp.stanford.edu>

¹⁶ TokenRegex: <http://nlp.stanford.edu/software/tokensregex.shtml>

estruturas de dados, por acreditar que são estruturas mais fáceis de manipulação. Para isso, o SUTime faz uso de uma biblioteca desenvolvida em Java chamada Joda-Time¹⁷.

- c) Resolução de expressões temporais: Realiza o reconhecimento de expressões temporais referentes a expressões temporais relativas.

3.11 CONSIDERAÇÃO

Existem vários estudos e ferramentas que se destinam ao tratamento de expressões temporais, mostrando que se trata de uma área relevante. Foram observadas várias características comuns a todas as ferramentas pesquisadas, enfatizando o uso das mesmas tecnologias no desenvolvimento (ou reuso de ferramentas de terceiros) até a mesma arquitetura de software.

Geralmente, as pesquisas são patrocinadas por órgãos governamentais e/ou universidades (grupos de pesquisa) especializadas na área de processamento de linguagem natural e, apesar da maior parte das ferramentas se destinarem aos idiomas inglês e espanhol, existem ferramentas específicas para outros idiomas¹⁸. Os documentos utilizados nessa pesquisa de Mestrado são escritos no idioma inglês, já que todo o projeto RISO, em sua fase atual, está concentrado nesse idioma. As primeiras versões das ferramentas destinadas ao tratamento de expressões temporais eram, apenas, *scripts* escritos em Perl ou outras linguagens de processamento de textos (MANI; WILSON, 2000). Porém, na atualidade, a maioria das ferramentas é desenvolvida em Java.

Outra característica encontrada em quase todas as pesquisas estudadas é a utilização de padrões fixos, como o TIMEX. Existem organizações que padronizam as expressões temporais que devem ser encontradas nos documentos e utilizam *software* de terceiros em sua estrutura. E como a maior parte dessas ferramentas utiliza os mesmos princípios, os resultados são semelhantes. Foi observado que as ferramentas que se destacam, na atualidade, estendem os padrões de reconhecimento de expressões temporais tradicionais para outras técnicas de análise (associações de padrões) e de arquitetura. O *Database Systems Reseach*

¹⁷ Joda-Time — joda-time.sourceforge.net

¹⁸ Timex Portal: <http://www.timexportal.info/systems>

Group da Heidelberg University, por exemplo, emprega a arquitetura do UIMA (*Unstructured Information Management*),¹⁹ desenvolvido e disponibilizada pela Apache²⁰ na criação do extrator HeidelbergTime.

Além das características já mencionadas, observou-se que muitas das ferramentas de extração temporal não são flexíveis e extensíveis, como é necessário no projeto RISO, porque, quando se trata de ferramentas mais elaboradas, existem dependências estruturais de ferramentas de terceiros, seja um *POS-tagger*, *frameworks* ou qualquer ferramenta que auxilie no processamento de texto.

Antes do desenvolvimento do RISO-TT foi realizada uma análise das soluções disponíveis na literatura que tivessem as características mencionadas anteriormente por se tratar de uma necessidade para o projeto RISO. Nessa análise, foi observado que a ferramenta que mais se aproximou das necessidades do projeto foi o HeidelbergTime, por se tratar de uma solução que permite a inserção de novas regras, apesar de utilizar software de terceiros em sua arquitetura, o que dificulta sua utilização.

Outro ponto que justifica a escolha por não se utilizar o HeidelbergTime foi o requisito funcional do projeto RISO de que o extrator temporal fosse desenvolvido em *Python*. Existem módulos do projeto RISO já desenvolvidos em *Python* e que serão integrados nas fases futuras de seu desenvolvimento.

Foi encontrado um projeto de extrator temporal em *Python*, chamado TERNIP²¹. Porém a ferramenta não atendia aos nossos propósitos, devido à complexidade analisada para uma mudança arquitetural do TERNIP e porque as características das ferramentas de extração de expressões temporais analisadas na literatura se apresentam de forma semelhante em suas características, tal como já referido no início dessa subseção.

O RISO-TT é independente de padrões fixos e de *software* de terceiros (módulos complementares), diferentemente da maioria dos demais extratores de expressões temporais encontrados, que implementam os mesmos padrões e/ou se utilizam de *software* de terceiros (*POS-Taggers*, ferramentas de PLN etc.) em sua arquitetura. Ambos os casos podem causar problemas futuros, visto que os padrões

¹⁹ UIMA: <http://uima.apache.org/>

²⁰ Apache: <http://apache.org/>

²¹ TERNIP: <https://github.com/jnicklas/turnip>

evoluem, e se houver uma mudança arquitetural considerável, pode acarretar grandes problemas para o desenvolvimento. Além disso, o uso de *software* de terceiros põe em risco a continuidade e a evolução do projeto, já que os desenvolvedores não controlam totalmente todos os módulos do projeto.

Por fim, o uso das associações complexas mencionadas no início dessa subseção é um requisito funcional do projeto RISO, pois o módulo subsequente ao RISO-TT necessita do máximo de informações das sentenças extraídas para o processo de enriquecimento semântico dos documentos. Essa fase do projeto não será abordada nesse trabalho, pois se trata de outro projeto de pesquisa e dissertação de Mestrado.

A proposta do RISO-TT é de disponibilizar uma ferramenta flexível e extensível, com base em regras bem definidas oriundas de observações (heurística) de expressões temporais encontradas em um conjunto de documentos. A arquitetura do RISO-TT foi definida para facilitar a configuração e a criação de novos padrões e regras. Para tanto, foram utilizados arquivos de configuração no formato XML, pois, caso existisse um padrão temporal não mapeado antes do processamento, ele fosse inserido para, no próximo processamento, o novo padrão fosse encontrado.

CAPÍTULO 4

O PROJETO RISO

Esse capítulo se destina a apresentar o projeto RISO (Recuperação da Informação Semântica de Objetos Textuais), do qual o RISO-TT, descrito nessa pesquisa, faz parte como um dos componentes estruturais do sistema.

4.1 ESTRUTURA GERAL

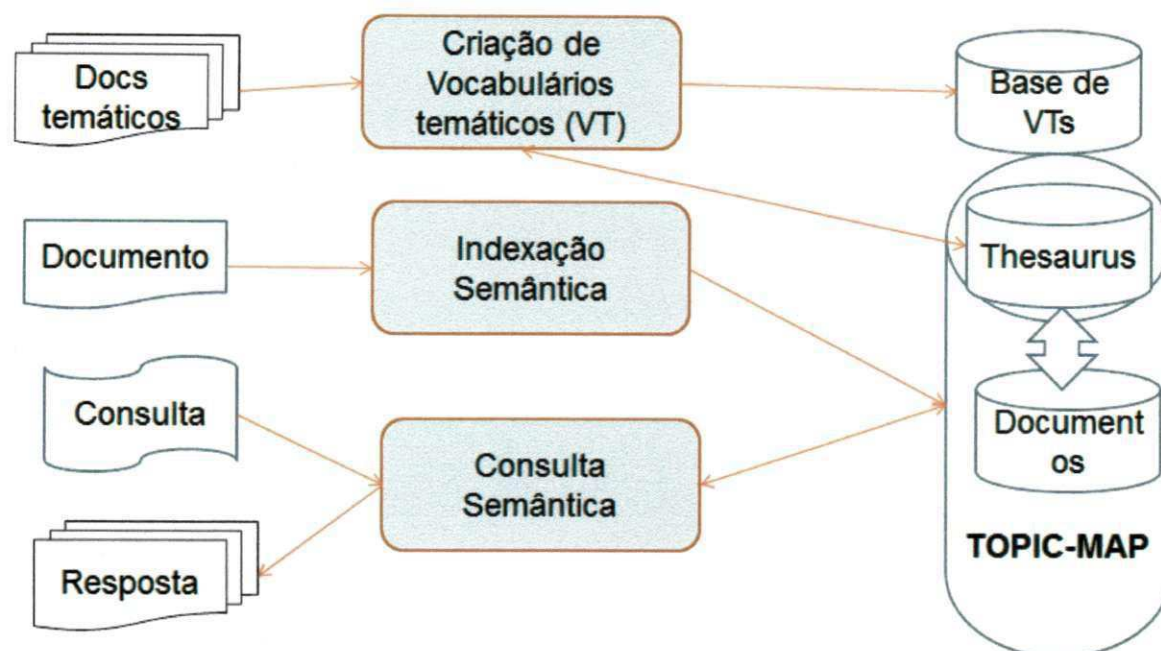
O projeto RISO (Recuperação da Informação Semântica de Objetos Textuais) tem como objetivo criar um ambiente de indexação e recuperação semântica de documentos, que possibilite uma recuperação mais acurada e melhore o fator de precisão dos resultados mediante a diminuição da ambiguidade do sentido dos termos (BISPO, 2012).

O RISO é um projeto de pesquisa que está sendo desenvolvido pelo grupo de Sistemas de Informação e Bancos de Dados (SINBAD), do DSC/CEEI/UFCEG, composto por vários professores, pesquisadores e bolsistas de iniciação científica e coordenado pelo Prof. Dr. Ulrich Schiel. O RISO está dividido, estruturalmente, em três componentes independentes: a) Criação de vocabulários temáticos (RISO-VTD); b) Indexação semântica (RISO-IS); e c) Consulta semântica (RISO-CS). A figura 3 apresenta essa estrutura.

Um conjunto de documentos temáticos é colocado como entrada para o processamento e extração dos vocabulários temáticos. Como saída desse processamento, é gerado uma base de vocabulários temáticos. Ou seja, vocabulários que caracterizam um domínio de conhecimento.

Os vocabulários temáticos são utilizados no processo de indexação de novos documentos. Isso porque, esses vocabulários são utilizados no processo de enriquecimento semântico dos novos documentos.

Figura 3: Estrutura geral do RISO (BISPO, 2012)

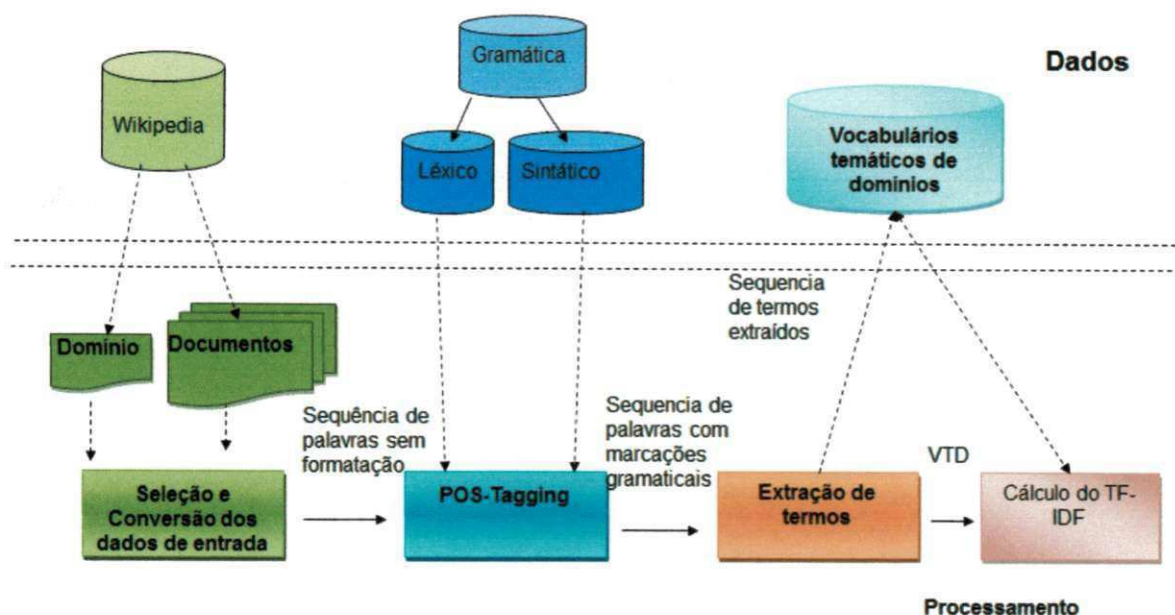


Uma vez indexados e enriquecidos, os documentos ficarão disponíveis para que os usuários realizem pesquisas semânticas por meio do módulo de Consulta.

4.2 CRIAÇÃO DE VOCABULÁRIOS TEMÁTICOS (RISO-VTD)

Com o intuito de classificar previamente novos documentos por domínios do conhecimento, com vistas a promover uma indexação semântica melhor dos termos contidos nele, a primeira etapa do processo está relacionada à criação de vocabulários temáticos de domínio (VTD). Cada domínio do conhecimento terá um VTD que reflete seu vocabulário. Cada VTD é criado por meio da extração dos termos de um número significativo de documentos do domínio correspondente. Essa extração é baseada em heurísticas de associações sintáticas e morfológicas das palavras (BISPO, 2012). A figura 4 mostra as quatro etapas desse processo: preparação dos documentos fonte; *POS-Tagging*; determinação de termos e cálculo do TF-IDF.

Figura 4: Processos do RISO-VTD (BISPO, 2012)



Inicialmente, foram determinados os principais domínios de conhecimento e selecionados os seus documentos existentes na Wikipédia. O segundo passo consistiu em processar os documentos e marcá-los com o uso de um *POS-Tagger*²² cuja finalidade é o reconhecimento gramatical dos termos que compõem os documentos. Empregou-se o *POS-Tagger* do software livre MontyLingua²³, que foi estendido com diversas novas heurísticas, para atender aos objetivos do projeto. De posse dos sintagmas marcados, foram determinados os termos significativos. Esse dois módulos são denominados de RISO-EXT. Por último, a importância de cada termo em seu domínio foi determinada pelo cálculo do seu TF-IDF (MANNING; RAGHAVAN; SCHUTZE, 2008).

Para validar esse trabalho, foi desenvolvido um módulo de cálculo da similaridade entre vetores de termos, baseado no modelo vetorial de Salton (1983), denominado SVSim (BISPO, 2012), o qual foi reaproveitado no componente seguinte de indexação de documentos.

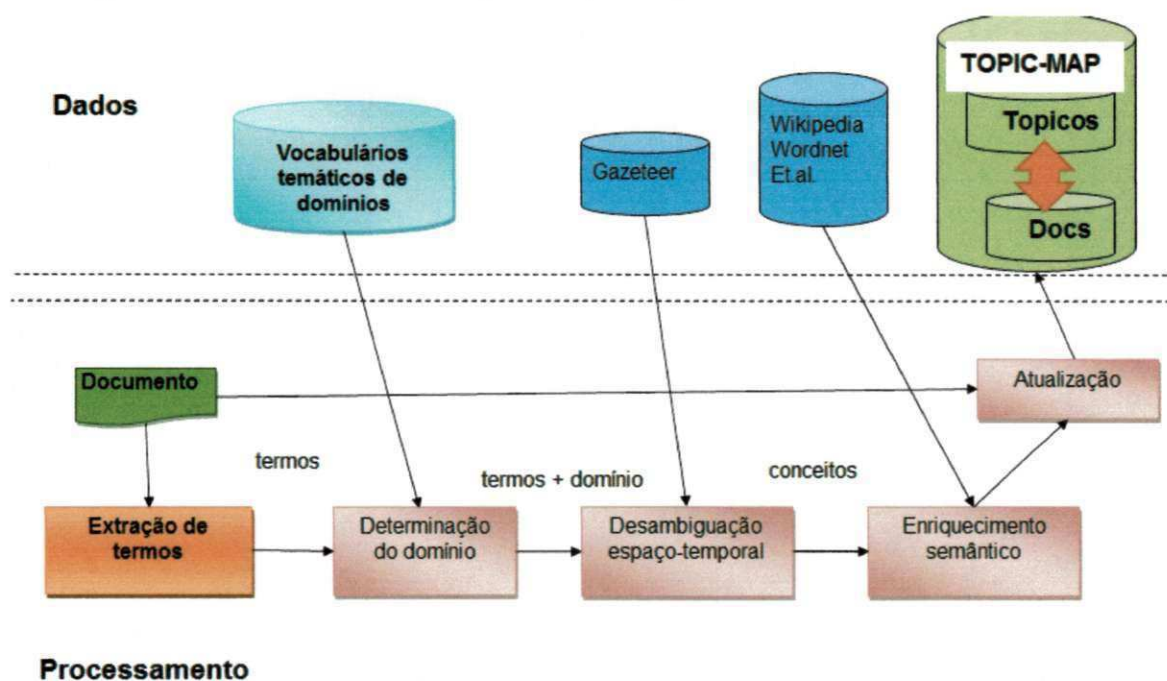
²² Pos-Tagger é um programa utilizado para reconhecer marcações de palavras em texto, de acordo com uma parte específica da fala, com base em sua definição, em seu contexto e na relação com as palavras adjacentes.

²³ MontyLingua: <http://web.media.mit.edu/~hugo/montylingua/>

4.3 INDEXAÇÃO (RISO-IS)

O segundo componente do projeto - o RISO-IS - está relacionado à fase de indexação de documentos. Como a principal característica do projeto RISO é a recuperação semântica de documentos, que só é possível com uma indexação preliminar adequada das fontes, esse componente tem especial importância na Arquitetura. O principal objetivo desse módulo é de transformar termos (ou sintagmas) existentes nos documentos em conceitos bem definidos. Na figura 5, são mostrados os cinco principais módulos do RISO-IS.

Figura 5: Módulo do RISO-IS (BISPO, 2012)



A primeira atividade, com a chegada de um novo documento, é a extração de termos, utilizando-se o RISO-EXT. Com os vetores extraídos, é determinado o domínio do documento comparando esse vetor com os VTD pelo SVSim. A próxima etapa é a de desambiguação conceito-espaço-temporal, que procura detectar expressões com características temporais e/ou espaciais no texto e remover ambiguidades conceituais. Esse módulo está dividido em três etapas: extração temporal, reconhecimento espacial e desambiguação conceitual.

O módulo de reconhecimento espacial se destina a identificar termos espaciais nas sentenças. Para isso, utilizará um *Gazetteer* (dicionário de termos geográficos) para avaliar se um termo é espacial ou não e, se for, realizará uma correspondência entre termos marcados pelo RISO-TT e por termos espaciais. Com a associação dessas atividades, será possível determinar características espaço-temporais dos termos existentes no documento.

A última etapa é a de analisar termos sem referência espaço-temporal que são ambíguos. Para remover essa ambiguidade, o sistema utiliza o domínio obtido na segunda etapa desse módulo e os termos vizinhos ao termo ambíguo para compará-los com as vizinhanças dos diversos conceitos correspondentes no thesaurus de conceitos (tópicos), detectando a vizinhança mais próxima.

Após a transformação de um termo em conceito, no módulo de enriquecimento semântico, serão acrescentadas relações semânticas linguísticas, obtidas de fontes externas, e o resultado é inserido no mapa de tópicos com as devidas conexões com o novo documento. Esse trabalho está em fase de conclusão e culminará em outra dissertação de Mestrado.

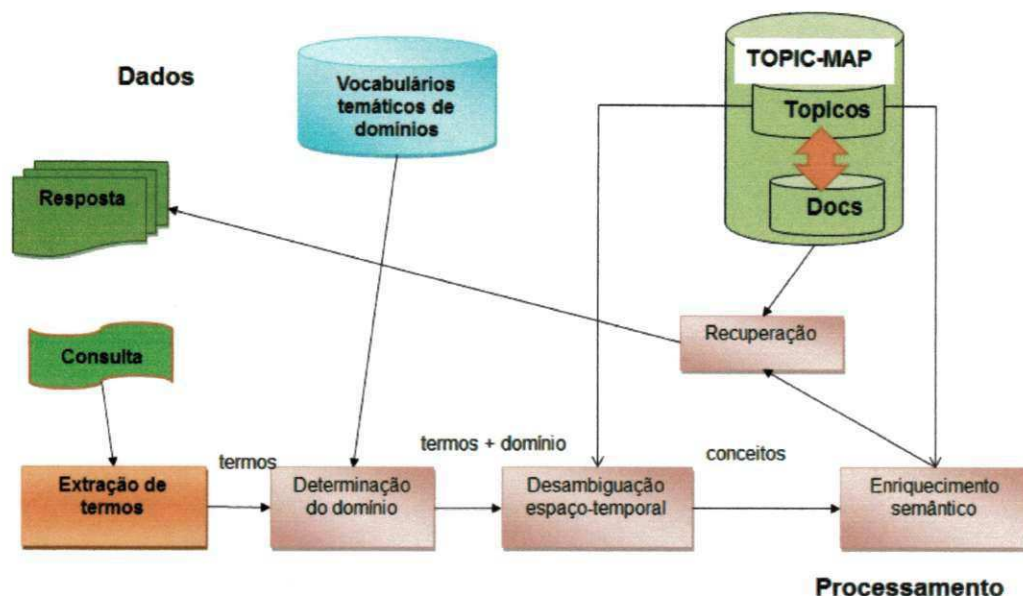
4.4 CONSULTAS SEMÂNTICAS (RISO-CS)

O último componente é o módulo de consultas. Trata-se de um *front-end* a ser desenvolvido com a finalidade de possibilitar consultas inteligentes dos usuários.

A partir de uma entrada inicial dada pelo usuário, a interface de comunicação realizará uma busca no Mapa de tópicos de todos os conceitos que tenham alguma relação (sintática ou semântica) com o termo definido e trará opções de desambiguação – conceitual, espacial e temporal - e enriquecimento semântico como sugestão. Com base no contexto determinado dos termos da consulta, o usuário poderá determinar, com precisão, a real necessidade de informação, fazendo com que os resultados estejam de acordo com essa necessidade.

Com isso, espera-se que os resultados sejam mais precisos, já que eles estarão diretamente relacionados ao domínio, ao sentido, ao tempo e ao espaço que o usuário deseja. Esse módulo ainda está sendo desenvolvido por um aluno de mestrado, no entanto, sua arquitetura está definida na figura 6.

Figura 6: Módulos de consulta RISO-CS (BISPO, 2012)



4.5 CONSIDERAÇÕES FINAIS SOBRE O PROJETO RISO

O RISO é um projeto de grande proporção e contribuição científica para a área de Recuperação da Informação. Ele se propõe a ser uma opção diferente para outros sistemas de recuperação da informação existentes, com a tentativa de transformar uma pesquisa sintática em semântica. Para que isso seja possível, os termos representativos, tanto na indexação de documentos quanto no processamento de uma consulta, são convertidos em conceitos, o que promove um enriquecimento desses conceitos por outros semanticamente relacionados por relações de hiper-/hiponímia, mero-/holonímia, acronímia, entre outras. Como o enriquecimento ocorre a partir de conceitos, e não, termos, espera-se evitar a introdução de "ruídos" com termos não desejados que degenerem a qualidade da resposta.

CAPÍTULO 5

RISO TEMPORAL TAGGER (RISO-TT)

Nesse capítulo, é apresentada a arquitetura, a estrutura e as características do módulo RISO *Temporal Tagger* (RISO-TT).

5.1 APRESENTAÇÃO

O RISO-TT é o extrator de expressões temporais do projeto RISO. Baseado em regras de associações de padrões temporais e inspirado no padrão TIMEX3, ele se difere das demais ferramentas de extração temporal por considerar associações gramaticais mais complexas em seu processo de identificação das expressões temporais.

Consideramos Expressões Temporais Compostas os resultados de associações gramaticais, que determinam períodos de tempo, e não, apenas, *tokens* temporais cuja análise é puramente sintática. Uma expressão temporal complexa é mais precisa por possibilitar o entendimento específico do tempo que se deseja representar.

Podemos exemplificar como Expressões Temporais Compostas estruturas formadas por intervalos fechados (*de 5 de março de 2010 até 20 de março de 2011; entre outubro e dezembro etc.*), intervalos semiabertos (*a partir de 5 de março de 2010*), associações gramaticais formadas por preposições, advérbios, números e *tokens* temporais (*em dezembro, no início de maio, entre de fevereiro e junho, meses depois etc.*), partes da semana e qualquer outra relação entre termos de expressão semântica temporal.

Essas associações complexas são representadas por associações entre sintagmas temporais e sintagmas gramaticais, que são inseridos em um dicionário de regras, as quais são organizadas com base em prioridade, que é heurísticamente definida pela quantidade de sintagmas associados na formação da expressão temporal para o processamento dos documentos, a fim de reconhecer, marcar e normalizar os valores temporais. Um sintagma é uma unidade formada por uma ou várias palavras que, juntas, descrevem um conceito. A ordem de marcação é

definida pelo número de sintagmas em uma frase. Quanto mais sintagmas, maior a prioridade da expressão sobre as outras.

Para exemplificar o que são Expressões Temporais Compostas e a necessidade de organizar o dicionário de regras em ordem de prioridade, temos a expressão temporal *“from December 10, 2011 to December 10, 2012”*. Essa frase está se referindo a um tempo específico entre as datas de *December 10, 2011* até *December 10, 2012*, que pode ser representada por $10/12/2011 < X < 10/12/2012$, em que *X* determina o intervalo temporal ao qual a expressão se refere. Nos outros extratores temporais pesquisados, a frase citada seria marcada com duas expressões temporais a) *December 10, 2011* e b) *December 10, 2012*. O período temporal ao qual se refere a expressão completa (intervalo entre as datas) seria desconsiderado. No exemplo dado, um documento contendo *“from December 10, 2011 to December 10, 2012”* teria mais prioridade do que o que contnha só *“December 10, 2011”* ou *“December 10, 2012”*.

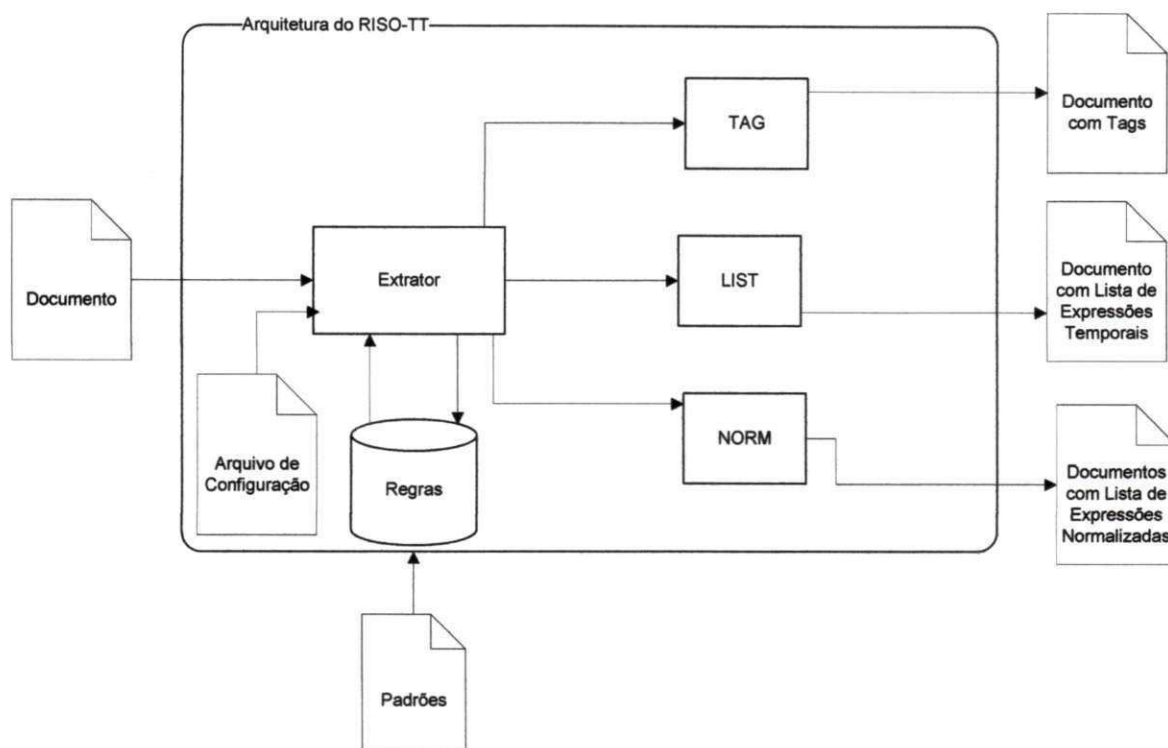
5.2 ARQUITETURA

O RISO-TT foi projetado para se tornar uma ferramenta extensível e flexível. Estas características são possíveis pelo uso de um arquivo de configuração que é utilizado pelo programa para identificar expressões temporais. Portanto, para acrescentar novos padrões e regras ou modificar os existentes, basta atualizar este arquivo, sem ter que alterar o programa.

Para o usuário acrescentar um novo padrão ou regra esse novo padrão ou regra terá que ser acrescido em seus arquivos de configuração, tendo-se o cuidado de analisar a ordem de prioridade definida no dicionário. Para que essa modificação seja observada, o RISO-TT deverá ser executado novamente. A figura 7 mostra a arquitetura do RISO-TT.

O módulo central do RISO-TT é o Extrator. Ele recebe um documento como entrada para o processamento das informações e, com base nas regras definidas no arquivo de regras e de acordo com os padrões mapeados nos arquivos de padrões, realiza o processamento do texto para a extração das expressões temporais. O Extrator acessa as regras e padrões de acordo com os caminhos contidos no Arquivo de Configuração.

Figura 7: Arquitetura do RISO-TT



O processamento das informações do documento significa que o Extrator irá procurar, em todo o texto do documento, expressões temporais reconhecidas por alguma regra e, uma vez encontrada, uma marcação é realizada de acordo com a função de processamento (TAG, LIST ou NORM) escolhida pelo usuário no momento do processamento.

Cada função de processamento resulta em um arquivo específico. A função TAG criará um documento com todas as expressões temporais marcadas entre tags no formato `<RISOTime type=>expressão</RISOTime>`. A função LIST gera, como resultado do processamento do documento, um arquivo com o vetor das expressões temporais encontradas e as respectivas regras utilizadas. A função NORM gera um arquivo com a lista de expressões temporais normalizadas, de acordo com `<alguma norma!>`.

Considerando-se essas características e necessidades, a arquitetura do RISO-TT é baseada em um arquivo de configuração no formato XML (*eXtensible*

Markup Language) que determina o acesso a um arquivo de padrões e um de regras. Esses arquivos estão disponíveis nos Anexos B, C e D.

Um padrão é um conjunto de denominações de termos (temporais ou gramaticais) agrupados semanticamente, que possam atribuir valor a uma expressão temporal. Os padrões são utilizados na formação das regras. Para descrever preposições, advérbios, as estações do ano, datas, horas e expressões regulares são utilizados padrões. Na tabela 1, são apresentados exemplos de padrões e, no Anexo C, há uma lista mais extensa dos padrões utilizados no RISO-TT.

Tabela 1: Exemplos de padrões temporais

Sigla (Classe de Padrões)	Exemplo
I (Intervalos)	"from" dia mes "until" dia mês
EPT (Estrutura Pré-temporal)	in the beginning, in the start of, in mid-[A-Za-z]*
EBT (Estrutura Básica Temporal)	mes dia", " ano, mes", " ano, mês, ano
Pre (Preposições)	In, on, at, from
UT (Unidade Temporal)	Weekends, week-end, minute, hours
Adv (Advérbio)	Early, later, past
DE (Datas Especiais)	Branch Sunday, Thursday of Mysteries, Presidents Day
H (Padrões de Hora)	^([0-1][0-9] [2][0-3]):([0-5][0-9])\$
Dia (Padrões de Dia)	Eleventh, twenty-second, second
Mês (Meses)	mid-february, january, oct
Ano (Padrões para Ano)	((?<=\\s)\\d{4})^\\d{4}

Uma regra é uma sequência ordenada de padrões (temporais e/ou nominais) que caracteriza a formação de expressões temporais e considera a posição dos termos que formam uma expressão. Por exemplo, a regra "*adv dia mês ano*", que é formada pelos padrões *adv*, *dia*, *mês* e *ano*, é diferente da regra "*adv mês dia ano*". Muitas ferramentas de extração temporal buscam, apenas, por *tokens* comumente conhecidos, tais como: datas, horas, datas especiais e/ou estruturas simples de expressões temporais (Ex: *today*, *now*, *week*). A tabela 2 apresenta um conjunto de regras definidas para o RISO-TT; no Anexo D, está disponível o arquivo de regras completo.

Tabela 2: Exemplos de regras

Regras	Descrição
EPT-EBT-Adv	estrutura_pre_temporal estrutura_basica_temporal adverbio
EPT-EBT-UT	estrutura_pre_temporal estrutura_basica_temporal unidade_temporal
EPT-UT-Adv	estrutura_pre_temporal unidade_temporal adverbio
Pre-EBT-Adv	preposicoes estrutura_basica_temporal adverbio
Pre-EBT-UT	preposicoes estrutura_basica_temporal unidade_temporal
EPT-EMT-Adv	estrutura_pre_temporal estrutura_minima_temporal adverbio
EPT-EMT-UT	estrutura_pre_temporal estrutura_minima_temporal unidade_temporal
Pre-EA-A	preposicoes estacao_do_ano ano
EPT-EBT	estrutura_pre_temporal estrutura_basica_temporal
Pre-EBT	preposicoes estrutura_basica_temporal
Pre-EMT	preposicoes estrutura_minima_temporal
EPT-UT	estrutura_pre_temporal unidade_temporal

Com o uso das regras, é possível encontrar expressões temporais mais complexas e, também, com valores semânticos mais precisos. Isso porque possibilitou que as estruturas complexas entre as relações gramaticas e as expressões temporais clássicas fossem reconhecidas como uma única expressão, visando aumentar o processo de reconhecimento de diversos padrões. Com isso, expressões como “*from* December 10, 2011 *to* December 10, 2012” são classificadas como uma única expressão temporal, e não, como vários *tokens* temporais.

Devida a quantidade de relações existentes entre os padrões na formação das regras utilizadas no reconhecimento dos padrões temporais, o processamento dos documentos é lento. A justificativa para isso é de que uma única regra que combina dois padrões com n e m possíveis instanciações gera um conjunto de $n \times m$ associações a serem verificadas. Por exemplo, dados os padrões *Preposições* = (*until, since, after, before*) e *Mês* = (*january, february, march, april, may, june, july, august, september, october, november, december*), a regra *Preposições X mês*, gerará o produto cartesiano entre esses padrões, na ordem de disposição dos termos presentes nos padrões (Ex: *until january, until february, ..., before december*).

5.3 ESTRUTURAS RESULTANTES DO PROCESSAMENTO DO RISO-TT

Como saídas de um documento processado pelo RISO-TT, conforme visto na figura 7 , são gerados (vide Anexo E):

- a) Documento marcado (TAG): A partir de um documento sem marcação é criado um documento marcado com as *tags* “RISOTime” e com o atributo “*type*” seguido *pela* regra utilizada (Ex: `<RISOTime type=Pre-EBT>On September 1, 1939</RISOTime>`).
- b) Relação de regras aplicadas e expressões temporais encontradas (LIST): Uma lista de todas as expressões temporais encontradas no documento (Ex: EBT-N -> *from 499 to 493 BC*).
- c) Expressões normalizadas (NORM): As expressões temporais em LIST são convertidas em valores normalizados (Ex: *On September 1, 1939 <--> 1-09-1939*). Os cálculos numéricos para o processamento do valor normalizado das expressões temporais do tipo *data*, hora e minutos são realizados por meio da função *date*²⁴ da linguagem Python, que realiza essa normalização de forma transparente. Além disso, um grupo de Expressões Temporais Compostas foi mapeado para que valores de intervalos fossem calculados (Ex: *from May 10, 2010 to May 10, 2013 < -- > 10/05/2010 > X < 10/05/2012*, onde *X* representa o intervalo temporal). Existem padrões para os quais não foi conseguido gerar uma forma normal. Neste caso é colocada a observação “Padrão ainda não foi mapeado”

5.4 PSEUDO-CODIGO DO RISO-TT

O módulo *Extrator* do RISO-TT realiza a marcação e extração das expressões temporais encontradas em textos de acordo com um conjunto de regras definidas em um dicionário de expressões temporais. O pseudo-código a seguir, demonstra a execução desta atividade.

²⁴ Função *Date* da Linguagem Python: <http://pythonhelp.wordpress.com/2012/07/10/trabalhando-com-datas-e-horas-em-python-datetime/>

CONFIGURAÇÃO:

Leitura do arquivo de configuração;

Para cada regra no arquivo de configuração faça:

Pegue o nome da regra e seu caminho //caminho significa o caminho em que o arquivo está no sistema de arquivo

Abra o arquivo das regras

Para cada símbolo encontrado no arquivo faça:

Leia o símbolo, pegue seu tipo e sua expressão

Salve no mapa de regras o símbolo como chave e a lista de sub-expressão como valor

Para cada arquivo de padrão faça:

Leitura do arquivo do padrão

As expressões são unidas em uma única expressão lógica usando o operador "OU"

MÉTODOS (TAG, LIST ou NORM):

Ler (entrada) //entrada significa o documento que é recebido como parâmetro

C = Pega primeiro caractere da entrada

Para cada regra R no mapa de regras faça:

Para cada expressão E em R faça:

Match de E a partir de C

Se não encontrar

Vai para próxima regra

Se encontrar

Se a função for TAG:

Substitui E por E entre as tags da Regra.//<Risotemp valor=EBT> E

</Risotemp>

Se a função for LIST:

Extrai a expressão E e insere numa lista.

Se a função for NORM:

Extrai a expressão E, normaliza o valor da expressão em uma data e

adiciona o resultado em uma lista.

Desloca C para direita (na entrada) em 1 ou no mesmo número de caracteres do match

SAÍDA

SE ainda tem entrada

Vá para início

SENAO

Encerra

5.5 FORMALIZAÇÃO DO RISO-TT

Um **padrão temporal** p é composto por um nome (geralmente um mnemonico (como EBT, Adv, etc) e uma lista ordenada $\langle t_1, t_2, \dots, t_n \rangle$, onde cada t_i é um *itemset* de termos relacionados semanticamente. Os termos de

um *itemset* podem ser substantivos, adjetivos, advérbios, padrões de tempo, números e expressões regulares, que expressam direta ou indiretamente uma frase. Por exemplo, expressões como '*the end of the [0-9]{3} BC*' ou '*mes dia "of" ano*' são padrões.

Uma **regra temporal** r é uma lista ordenada $\langle p_1, p_2, \dots, p_n \rangle$ de padrões temporais. As associações desses padrões temporais formam as expressões temporais que são organizadas em um dicionário D , utilizado como fonte de consulta no processo de reconhecimento das expressões temporais nos documentos. Por exemplo a regra "EPT-EBT-Adv" associa 'estrutura_pre_temporal' 'estrutura_basica_temporal' e "adverbio</expressao>" enquanto a regra "Pre-A-Adv" associa 'preposições' 'ano' e 'adverbio'.

Uma **Corpora** C é um conjunto de documentos. Para $C = \{d_1, d_2, \dots, d_n\}$ cada documento d_i é formado por uma sequência de símbolos. Para cada documento d existe um gabarito completo de expressões temporais $g(d) = \{et_1, et_2, \dots, et_n\}$ onde n representa o número total de termos temporais existentes no documento. Um gabarito completo de expressões temporais pode ser determinado por uma análise realizada por especialistas humanos.

Um processo automático de extração de expressões temporais de um documento d é um mecanismo que utiliza padrões e regras pré-determinados para determinar um conjunto de expressões temporais $g'(d)$. A qualidade desse processo é definida pela intersecção de $g'(d)$ e $g(d)$. Isso significa dizer que, quanto maior o número de itens que estão em $g'(d)$ e que também estão presentes em $g(d)$, melhor o processo.

O processo de extração de expressões temporais pode ser descrito da seguinte forma. Um documento textual d pode ser considerado uma sequência de símbolos formados por letras ou sinais especiais de um idioma, ou seja, $d = (s_1, \dots, s_n)$. Toda subsequência $t = (s_{k_1}, \dots, s_{k_m})$ de d é chamada de um **trecho** em d . Seja T o conjunto de todos trechos de D , o problema posto é encontrar o subconjunto de T que contém todos trechos que representam expressões temporais. Estas formarão o gabarito $g(d)$.

Seja P um conjunto de padrões temporais e R um conjunto de regras temporais. Pela definição, cada regra pode ser escrita como $r = \langle p_1, \dots, p_n \rangle$ em que p_1, \dots, p_n são padrões. Um padrão p por sua vez é dado por $p = \langle n, t_1, t_2, \dots, t_n \rangle$, um

nome e n *itemsets*. Cada *itemset* é formado por trechos que determinam palavras ou outras formações de uma língua.

Uma **instanciação** *in* de uma regra $r = \langle p_1, \dots, p_n \rangle$ (denotada por $in \angle r$) é uma sequência de itemsets $\langle it_1, \dots, it_n \rangle$, tal que cada $it_i \in p_i$, para $i = 1, \dots, n$. Ou seja, instâncias dos padrões que formam uma regra, alinhados na ordem da regra, formam uma instanciação dela.

Uma **extração de expressões temporais** de um documento D consiste em encontrar um conjunto $ET \subseteq T$ tal que para todo $et \in ET$, temos $et \angle r$ para algum $r \in R$. Essa extração é maximal, se para todo $in \in T - ET$, *in* não é uma expressão temporal, ou seja, não existe $r \in R$ tal que $in \angle r$. Caso se encontre um trecho em D que expresse alguma informação temporal mas que não é detectado pelo sistema de extração, pode-se tentar estender os padrões P ou as regras R a fim de conseguir reconhecer este trecho. Isto é chamada uma extensão do processo de extração.

Uma forma 'bruta' de encontrar ET é de, para cada símbolo s_i em D , analisar todos os trechos (s_{i1}, \dots, s_{im}) a partir de s_i e verificar se alguma regra de R se aplica a este trecho.

Para o processo de extração de expressões temporais foram implementadas três funções: TAG, LIST e NORM.

A função TAG altera o documento d recebido como entrada, acrescentando, antes e depois de cada expressão temporal et encontrada, as respectivas tags referentes à regra correspondente. Ou seja, esta função substitui cada expressão temporal et por $\langle tag \rangle et \langle tag \rangle$, onde o valor de $\langle tag \rangle$ é atribuído de acordo com a regra que reconheceu a expressão temporal et .

A função LIST gera uma tabela no formato "regra -> expressão temporal" baseada nas expressões temporais encontradas pela função TAG, Todas as expressões temporais encontradas são listadas na ordem em que a expressão for encontrada no documento.

A última função de processamento é a NORM. Essa função baseia-se em um conjunto de funções temporais de normalização *norm*: $et \rightarrow etn$. Isso significa dizer que, para cada expressão temporal et encontrada no documento d uma expressão temporal normalizada etn será resultante.

Normalizar uma expressão temporal é encontrar um valor no formato de data (ex: 06/10/1982) que a represente. Para isso, é importante o mapeamento das regras mapeadas e, também, os cálculos necessários para que a normalização seja possível. Isso porque, cada símbolo encontrado numa expressão temporal deve ser reconhecido para que a normalização aconteça. Ex: a data 06/10/1982, deve ser reconhecida como 06 – dia, 10 – mês e 1982 – ano.

5.6 EXEMPLO DE UM DOCUMENTO PROCESSADO NO RISO-TT

O texto seguinte foi extraído do documento “16_SpanishCivilWar” da WikiWars. As expressões temporais estão destacadas em negrito.

... **On 7 March**, the Nationalists launched the Aragon Offensive. **By 14 April** they had pushed through to the Mediterranean, cutting the Republican-held portion of Spain in two. The Republican government tried to sue for peace **in May**, but Franco demanded unconditional surrender; the war raged on. **In July**, the Nationalist army pressed southward from Teruel and south along the coast toward the capital of the Republic at Valencia but was halted in heavy fighting along the XYZ Line, a system of fortifications defending Valencia. The Republican government then launched an all-out campaign to reconnect their territory in the Battle of the Ebro, **from 24 July until 26 November**...

Existe um conjunto de expressões temporais, como *on 7 March*, *By 14 April*, *in May*, *In July* e *from 24 July until 26 November* que, exceto a última, representam expressões temporais simples. Porém, a expressão *from 24 July until 26 November* não seria reconhecida por qualquer extrator, por se tratar de uma expressão temporal composta.

O RISO-TT reconhece Expressões Temporais Compostas por meio de regras reconhecendo associações de sintagmas temporais e nominais.

O texto apresentado como exemplo, ao ser processado pelo RISO-TT, tem as seguintes saídas:

a) Documento marcado (TAG):

<RISOTime type=Pre-EBT>On 7 March</RISOTime>, the Nationalists launched the Aragon Offensive. <RISOTime type=Pre-EBT>By 14 April</RISOTime>, they had pushed through to the Mediterranean, cutting the Republican-held portion of Spain in two. The Republican government tried to sue for peace <RISOTime type=Pre-EBT>in May</RISOTime>, but Franco demanded unconditional surrender; the war raged on. <RISOTime type=Pre-EBT>In July</RISOTime>, the Nationalist army pressed southward from Teruel and south along the coast toward the capital of the Republic at Valencia but was halted in heavy fighting along the XYZ Line, a system of fortifications defending Valencia. The Republican government then launched an all-out campaign to reconnect their territory in the Battle of the Ebro, <RISOTime type=I>from 24 July until 26 November</RISOTime>.

Onde:

Pre-EBT é a regra Preposições + Estrutura Básica Temporal.

I é a regra que agrupa padrões de Intervalos Temporais.

b) Vetor temporal (LIST):

Pre-EBT -> On 7 March

Pre-EBT -> By 14 April

Pre-EBT -> in May

Pre-EBT -> In July

I -> from 24 July until 26 November

c) Vetor normalizado (NORM):

by 22 February <--> 22-02-XXXX

On 7 March <--> 7-03-XXXX

By 14 April <--> 14-04-XXXX

in May <--> Padrão ainda não foi mapeado

In July <--> Padrão ainda não foi mapeado

from 24 July until 26 November <--> 24-07 < X < 26-11

Onde:

XXXX: representa um ano desconhecido.

< X <: representa um intervalo entre duas datas.

Para cada expressão encontrada em um texto, deverá haver regras específicas que determinam a normalização desta expressão. Quando esta regra ainda não estiver inserida nas regras para a normalização, o RISO-TT emitirá a mensagem "*Padrão ainda não foi mapeado*".

Todos os três tipos de processamento podem ser utilizados nos processos de indexação, busca e em sistemas que necessitem reconhecer expressões temporais para a tomada de decisão de forma automática. Para isso, basta que as ferramentas reconheçam as regras mapeadas no RISO-TT e implementem suas regras.

Na indexação, toda expressão temporal se torna um índice do documento e poderá ser recuperado. Por exemplo, se uma pesquisa por documentos que se referem a "7 de março" for submetida a um engenho de busca que utilize de indexação temporal dos documentos, caso exista algum documento indexado com este valor, este documento deverá ser retornado. Além disso, pesquisas mais trabalhadas (ex: por intervalos de tempo) poderão ser realizadas. Para isso, se faz necessária a normalização das expressões temporais encontradas.

A versão atual do RISO-TT poderá ser expandida para associar os tempos aos sintagmas nominais. Com essa associação, pode-se inferir que *Aragon offensive* ocorreu em "7 de março". Logo uma consulta por 7 de março terá um retorno mais preciso, pois retornará os conceitos '*Aragon offensive*', e '*Nationalists*'. Inversamente poder-se-á pesquisar por time-of ('*Aragon offensive*') e receber o 7-03-XXXX, onde XXXX representa o ano.

CAPÍTULO 6

VALIDAÇÃO E VERIFICAÇÃO

Nesse capítulo são apresentados o roteiro experimental realizado no RISO-TT e a verificação e validação dos resultados alcançados.

6.1 ROTEIRO DO ESTUDO EXPERIMENTAL

O roteiro do estudo experimental a seguir, se refere ao processo de reconhecimento de expressões temporais do RISO-TT.

- 1) **Tema do experimento:** reconhecimento de padrões temporais em documentos;
- 2) **Área de estudo:** recuperação de informação;
- 3) **O problema:** identificar corretamente expressões temporais em documentos;
- 4) **Importância do problema:** o processamento e o reconhecimento dessas informações podem ser utilizados em diversos sistemas computacionais, a saber: tradutores automáticos, indexação, recuperação de informação, processamento sintático e semântico, sistemas especialistas etc.
- 5) **Objetivos:**
 - Pergunta I:** A arquitetura proposta para RISO é flexível e extensível?
 - a. **Hipótese nula:** Não. Não é possível adaptar ou inserir novos padrões e regras, caracterizando a arquitetura desenvolvida como fixa.
 - b. **Hipótese alternativa:** Sim. O modelo se mostra flexível e extensível pois permite acrescentar novos padrões e regras. Caso expressões para estas novas regras sejam encontradas nos documentos processados elas serão identificadas e mapeadas.
 - Pergunta II:** Os resultados das marcações do RISO-TT foram melhores do que os resultados das ferramentas semelhantes?
 - a. **Hipótese nula:** Não. Os resultados encontrados não foram superiores aos dos concorrentes.

b. **Hipótese alternativa:** Sim. Os resultados encontrados foram superiores aos dos concorrentes selecionados.

6.1.1 Seleção das variáveis

- a) Variáveis independentes: Documentos sem marcação (D), padrões e regras.
- b) Variáveis dependentes: Documentos Marcados com tags temporais (D'), vetor temporal (VT) e vetor normalizado (VN).

6.1.2 Design do experimento

O *design* do experimento é formado por extratores temporais e documentos.

Experimento I

Pergunta I: A arquitetura proposta para RISO é flexível e extensível?

Extrator temporal

- a) RISO-TT: Extrator temporal desenvolvido para o projeto RISO.

Regras

Em relação à análise da capacidade de extensibilidade da ferramenta RISO-TT, serão realizados três ensaios complementares e será investigado se existem padrões ou regras temporais que ainda não mapeados no RISO-TT. Depois disso, caso seja encontrado algum novo padrão ou regra, esses serão inseridos nas regras e nos padrões do RISO-TT, e outra rodada do ensaio foi realizado. O resultado desse ensaio foi analisado a fim de verificar se os novos padrões foram reconhecidos conforme a hipótese alternativa referente à pergunta II.

Documentos

O corpus selecionado é composto por 22 documentos extraídos da Wikipédia, que descrevem o percurso histórico das guerras, no idioma inglês, com um total de 120,000 *tokens* e 2.681 expressões temporais marcadas no padrão TIMEX2.

Experimento II

Pergunta II: Os resultados das marcações do RISO-TT foram melhores do que os resultados das ferramentas semelhantes?

Design

Para o experimento foi definido 66 ensaios com base no cálculo abaixo:

1 (extrator temporal) * 3 (regras) * 22 (Documentos) = 66 ensaios para o experimento comparativo entre os resultados dos processamentos.

Extratores temporais

- a) RISO-TT: Extrator temporal desenvolvido para o projeto RISO;
- b) *Stanford Temporal Tagger* – SUTime: Extrator temporal desenvolvido pela Universidade de Stanford;
- c) *Heideltime Temporal Tagger*: Extrator temporal desenvolvido pelo *Database Systems Reseach Group* da Universidade de Heidelberg.

Documentos

Foi utilizada a mesma Corpora definida para o primeiro experimento.

Design

Para o experimento foi definido 66 ensaios com base no cálculo abaixo:

3 (extratores temporais) * 22 (Documentos) = 66 ensaios para o experimento comparativo entre as ferramentas temporais.

Variáveis de resposta

Para cada documento do corpus, existe um gabarito com todas as marcações temporais encontradas e marcadas por um conjunto de especialistas humanos. O resultado de cada ensaio realizado com os extratores será comparado com os do gabarito para o cálculo de precisão, cobertura e *f-measure*.

6.2 VERIFICAÇÃO

O processo de verificação foi responsável por responder ao questionamento sobre a extensibilidade e a flexibilidade do RISO-TT (Pergunta I acima), que diz: “O modelo proposto no desenvolvimento do RISO *Temporal Tagger* é flexível e extensível?”.

Para responder a essa questão, foram realizados três ensaios sobre a corpora WikiWars com três versões diferentes de configuração das regras, e o ajuste realizado em cada versão foi baseado em padrões não encontrados nos resultados do ensaio.

Para exemplificar esse processo, imagine que um documento D tem um conjunto de expressões temporais (ET_D), e esse documento foi processado por um marcador temporal (TT), resultando no documento D' . Esse documento D' conterá um conjunto de expressões temporais marcadas por TT definidas como $MT_{D'} = \{m_1, m_2, \dots, m_n\}$, onde cada m_i é uma expressão marcada com base em uma regra temporal p_i determina um conjunto de expressões temporais. Tem-se, nesse caso, $MT_{D'} \subseteq ET_D$.

Ao analisar o documento D' , observou-se que existem expressões temporais et em ET que não foram marcadas por TT , e isso ocorre porque a regra r , que corresponde àquelas expressões temporais, não pertence ao grupo de regras R , ou seja, $r \notin R$. Para que em D seja marcada a expressão et , é preciso determinar uma nova regra r que falta para identificar et . Uma vez que a nova regra foi definida,

roda-se um novo ensaio com $R'=R \cup \{r\}$ obtendo um novo D' determinando $MT_{D'}$ tal que $et \in M_{D'}$.

Para exemplificar esse processo, no primeiro ensaio do experimento, o documento "09_GrecoPersian" foi processado com a primeira versão das regras. Porém, foi observado que os padrões de datas com três caracteres numéricos e símbolo BC não foram encontrados (Ex: 100 BC). Sabendo disso, foi criado um padrão chamado "datas_de_0_a_999" e criado regras (semelhantes às regras das datas convencionais). Depois de um novo processamento do mesmo documento, foi observado que todas as expressões temporais existentes no documento "09_GrecoPersian" e que foram mapeados no padrão "datas_de_0_a_999" foram extraídos normalmente.

Esse processo foi repetido três vezes (com base em outros padrões) e foi observado que as expressões que não eram reconhecidas antes, passaram a ser reconhecidas após a inserção das regras correspondentes. Isso prova que o RISO-TT é extensível, porquanto não foi necessário alterar a arquitetura do RISO-TT para que os novos padrões fossem reconhecidos, além da própria inserção deles e de regras nos arquivos de configuração.

6.3 VALIDAÇÃO

Para validar o desenvolvimento do RISO-TT e analisar o seu desempenho com ferramentas do mesmo segmento, foi realizado um experimento comparativo. Para isso, foram selecionadas duas ferramentas de marcação temporal. Os ensaios foram realizados, e os resultados computados e comparados com o gabarito de marcações ideais disponível pela própria corpora *WikiWars*.

6.3.1 Análise dos dados

A corpora *WikiWars* apresenta um gabarito com os resultados das marcações temporais ideais de todos os documentos que a compõem. Esse gabarito foi comparado com os resultados dos processamentos (documentos marcados) das ferramentas escolhidas para o experimento (Heidelttime, SUTime e RISO-TT) nos

seguintes aspectos: *tokens* checados - tc, *tokens* encontrados - te e as medidas de precisão, cobertura e *f-measure*.

A tabela 3 é formada por dados de *tokens* ideais, *tokens* checados e *tokens* encontrados. Os *tokens* ideais são formados pela marcação do gabarito da WikiWars. Os *tokens* checados são formados pela intersecção entre os *tokens* encontrados no gabarito da WikiWars e os *tokens* marcados nos documentos. Os *tokens* encontrados são formados por todos os *tokens* temporais encontrados por cada ferramenta, incluindo os *tokens* marcados de forma errada (Ex: *may* (verbo) e marcado como *May* (mês)) e demais erros de marcação.

Tabela 3: Tabela de tokens ideais, tokens checados e tokens encontrados

Documentos	Tokens checados			Tokens encontrados			
	Tokens ideais	Heidel-time	Su-Time	RISO-TT	Heidel-time	Su-Time	RISO-TT
01	170	144	143	165	167	161	166
02	265	206	204	254	231	240	258
03	75	63	55	72	74	70	79
04	147	126	123	141	145	143	149
05	245	172	163	212	203	199	226
06	149	110	95	129	126	123	134
07	247	203	174	204	228	217	211
08	175	133	121	157	154	150	167
09	129	30	80	108	58	107	121
10	57	15	38	47	17	47	51
11	102	68	72	90	102	92	95
12	98	86	78	90	92	89	92
13	104	87	81	100	103	93	104
14	71	52	53	58	64	62	61
15	78	64	47	68	72	55	72
16	63	53	54	56	64	58	69
17	130	92	83	97	107	109	103
18	110	88	87	98	95	93	102
19	62	47	47	57	63	55	60
20	106	92	87	99	103	96	110
21	29	24	25	28	28	28	30
22	69	54	48	58	64	59	62

Com base nas informações encontradas, foram calculadas as precisões, a cobertura e a *f-measure* das amostras.

A precisão é calculada por:

$$\text{Precisão} = \frac{|\text{Tokens Checados}|}{|\text{Tokens Encontrados}|}$$

A cobertura é calculada por:

$$\text{Cobertura} = \frac{|\text{Tokens Checados}|}{|\text{Tokens Ideais}|}$$

A *f-measure* é calculada por:

$$F\text{-measure} = \frac{2 * (\text{Precisão} * \text{Cobertura})}{(\text{Precisão} + \text{Cobertura})}$$

Os resultados são apresentados na tabela 4:

Tabela 4: Tabela de valores de precisão, cobertura e *f-measure*

Docu- mentos	Precisão			Cobertura			F-Measure		
	HT	SU	R-TT	HT	SU	R-TT	HT	SU	R-TT
01	0,8622	0,8882	0,994	0,8470	0,8411	0,9705	0,8545	0,8640	0,9821
02	0,8917	0,85	0,984	0,7773	0,7698	0,9584	0,8306	0,8079	0,9713
03	0,8513	0,7857	0,911	0,84	0,7333	0,96	0,8456	0,7586	0,9350
04	0,8689	0,8601	0,946	0,8571	0,8367	0,9591	0,8630	0,8482	0,9527
05	0,8472	0,8191	0,938	0,7020	0,6653	0,8653	0,7678	0,7342	0,9002
06	0,8730	0,7724	0,963	0,7382	0,6375	0,8657	0,8	0,6985	0,9116
07	0,8903	0,8018	0,967	0,8218	0,7044	0,8259	0,8547	0,75	0,8908
08	0,8636	0,8067	0,94	0,76	0,6914	0,8971	0,8085	0,7446	0,9181
09	0,5172	0,7477	0,893	0,2325	0,6201	0,8372	0,3208	0,6779	0,864
10	0,8823	0,8085	0,922	0,2631	0,6666	0,8245	0,4054	0,7307	0,8703
11	0,6666	0,7826	0,947	0,6666	0,7058	0,8823	0,6666	0,7422	0,9137
12	0,9347	0,8764	0,978	0,8775	0,7959	0,9183	0,9052	0,8342	0,9473
13	0,8446	0,871	0,962	0,8365	0,7788	0,9615	0,8405	0,8223	0,9615
14	0,8125	0,8548	0,951	0,7323	0,7464	0,8169	0,7703	0,7969	0,8787
15	0,8888	0,8545	0,944	0,8205	0,6025	0,8717	0,8533	0,7067	0,9066
16	0,8281	0,931	0,812	0,8412	0,8571	0,8888	0,8346	0,8925	0,8484
17	0,8598	0,7615	0,942	0,7076	0,6384	0,7461	0,7763	0,6945	0,8326
18	0,9263	0,9355	0,961	0,8	0,7909	0,8909	0,8585	0,8571	0,9245
19	0,7460	0,8545	0,95	0,7580	0,7580	0,9193	0,752	0,8034	0,9344
20	0,8932	0,9063	0,9	0,8679	0,8207	0,9339	0,8803	0,8613	0,9166
21	0,8571	0,8929	0,933	0,8275	0,8620	0,9655	0,8421	0,8771	0,9491
22	0,8437	0,8136	0,935	0,7826	0,6956	0,84058	0,8120	0,75	0,8854

Com base nos dados apresentados na tabela 4, o RISO-TT apresentou melhores resultados do que as demais ferramentas. Por exemplo, nos documentos 9 e 10 o HeideTime e o SUTime apresentaram baixos valores de precisão e cobertura e o motivo disso foi a falta de padrões temporais referentes aos padrões temporais existentes nos respectivos documentos. Destacamos que a superioridade do RISO-TT foi possível devida a quantidade de regras mapeadas e a diversidade de relações existentes entre os padrões.

Foi observado, também, que devido a alguns padrões não serem reconhecidos pelas ferramentas concorrentes, os resultados foram baixos em alguns documentos. Exemplo disso, o processamento do documento de número 10 pela ferramenta HeideTime.

6.3.2 Resultados

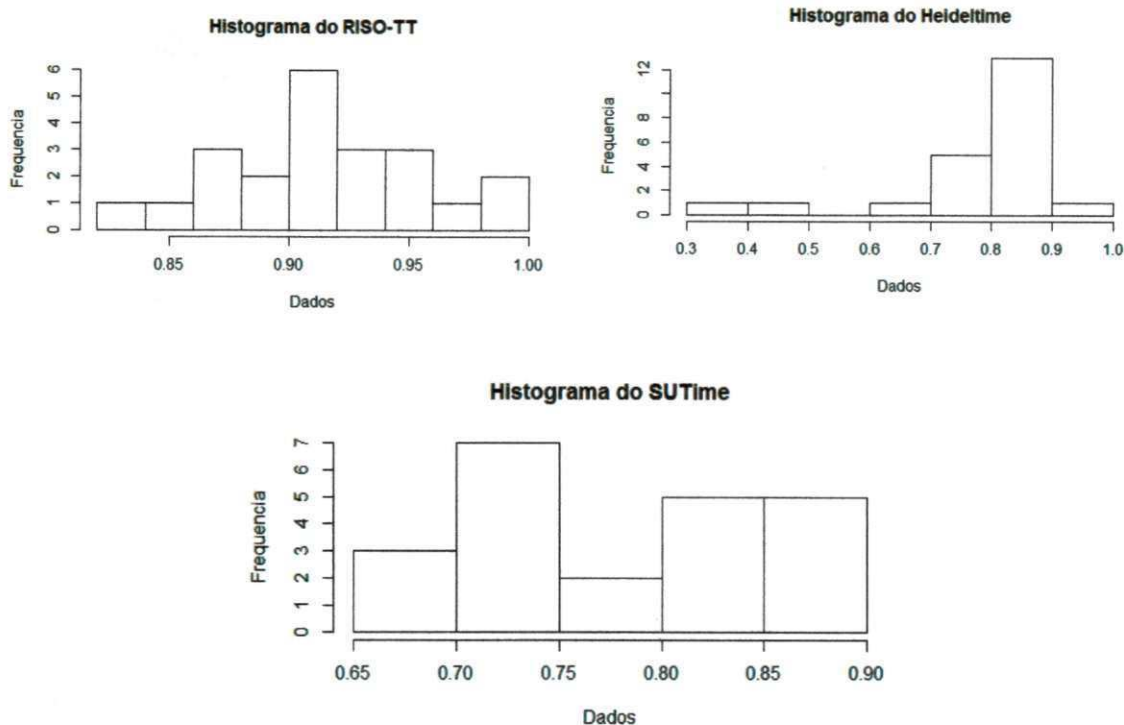
A primeira atividade realizada após a execução dos ensaios foi avaliar a normalidade dos dados resultantes. Para isso, foi executado o teste *Shapiro-Wilk*, por meio do qual a normalidade dos dados é verificada. Neste teste, segundo Royston (1995), se o valor do *p-value* for maior do que 0,1, os dados são normais e, também, quanto maior o valor de *W* maior a confiabilidade do teste.

Com a execução do teste *Shapiro-Wilk* foram obtidos os resultados mostrados na Tabela 5 e, nos histogramas da Figura 8, são apresentadas as distribuições dos dados referentes às ferramentas HeideTime, SUTime e RISO-TT e servem para avaliar se a distribuição dos dados são normais. No histograma do HeideTime existe uma calda na distribuição dos dados que caracteriza uma amostra não normalizada e este valor se confirma com o *p-value* da Tabela 5.

Tabela 5: Tabela de testes de normalidade

Shapiro-Wilk		
Amostra	W	P-Value
HeideTime	0.652	0,00508
SUTime	0.942	0.2176
RISO-TT	0.9831	0.9574

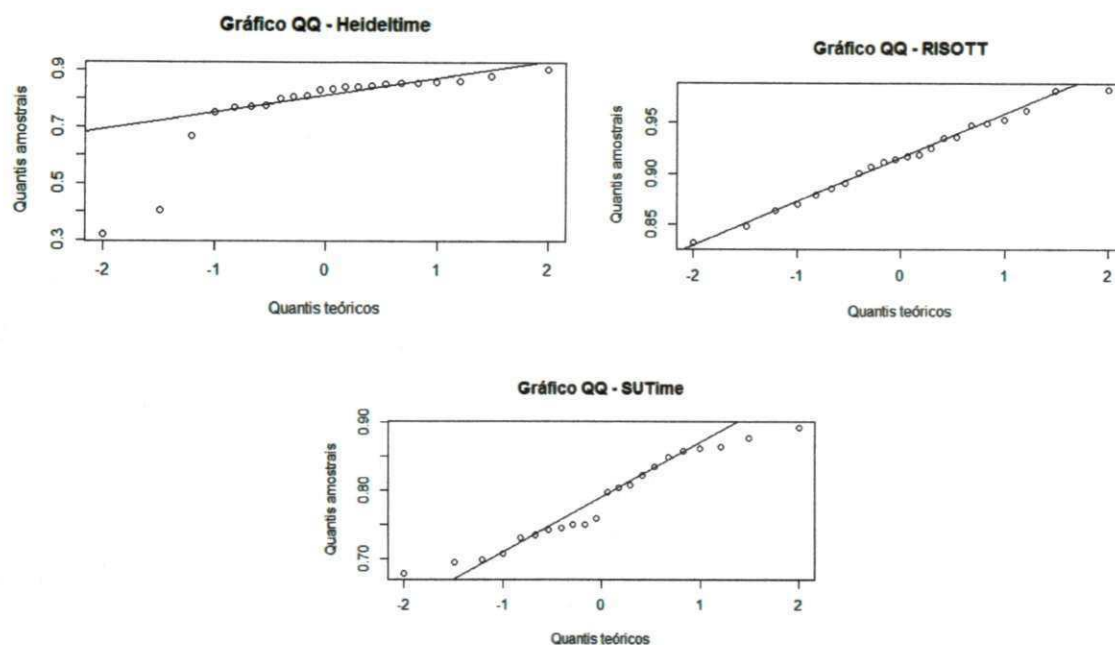
Figura 8: Histogramas dos dados dos extratores temporais



O p -value obtido dos dados *HeidelTime*, assim como as distribuições dos dados apresentadas nos histogramas, caracterizam uma amostra não normalizada, conforme já mencionado. No entanto, avaliando os dados, observou-se que trata-se de três amostras que, no processo de marcação de documento, não obtiveram um bom resultado. Essas amostras apresentam expressões temporais em um formato para datas com três caracteres numéricos (Ex: 200 DC) e esse formato não foi reconhecido pela ferramenta *HeidelTime*. Portanto, por se tratar de um número de documentos não expressivo, podem ser considerados *outliers*.

Para que não haja dúvidas sobre o processo de avaliação dos dados, foram elaborados os gráficos Q-Q (Quantil-Quantil), que também possibilitam a análise da distribuição dos dados. Conforme o comportamento dos dados apresentados na figura 9, o comportamento dos dados observados nos histogramas foi confirmado, isso porque as amostras (pontos) tangenciam a reta traçada no gráfico representando a distribuição normal e, também, observa-se um distanciamento de três amostras no gráfico do *HeidelTime*.

Figura 9: Gráficos Q-Q



Com base no intervalo de confiança de 95%, procedeu-se o teste estatístico. Para tanto, foram calculados a média amostral, o desvio-padrão e o erro-padrão, cujos resultados estão expostos na tabela 6. Com o cálculo desses valores é possível comparar os resultados encontrados. Para tanto, foi criado o gráfico de *boxplot* dos intervalos de confiança apresentado na Figura 10. Esses gráficos, também, são utilizados para analisar se as distribuições são normais. Os pontos do gráfico devem ficar próximos à reta que corta o gráfico.

Tabela 6: Tabela de intervalos de confiança

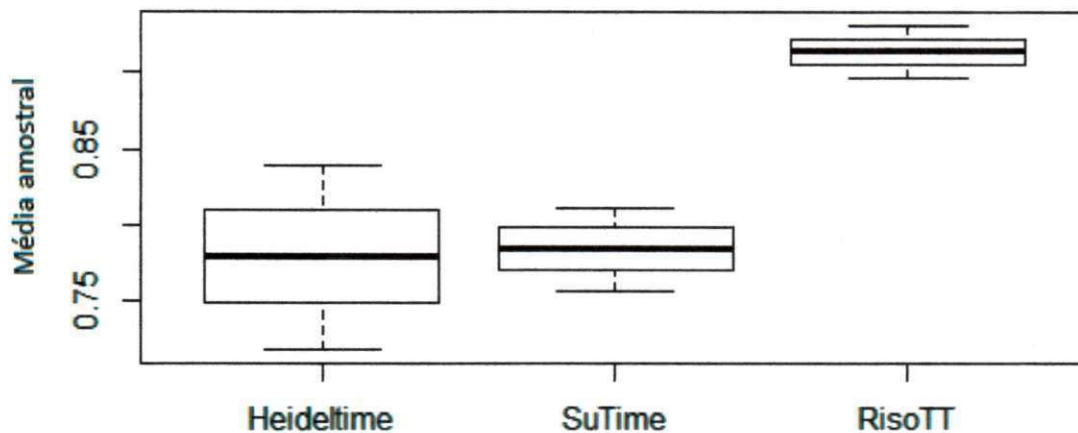
Intervalos de confiança			
Amostra	Média amostral – Erro-padrão	Média amostral	Média amostral + Erro-padrão
Heideltime	0.7187674	0.7792521	0.8397369
SuTime	0.7566779	0.7842650	0.8118521
RisoTT	0.8970079	0.9139180	0.9308280

Com os dados estatísticos encontrados e apresentados na figura 10, é possível afirmar que não existe uma superioridade entre as ferramentas Heidelberg

e SUTime. Isso é possível devido a presença de uma intersecção entre os intervalos de confiança dos seus *boxplot*. Esse fato pode ser explicado porque ambas as ferramentas utilizam o mesmo padrão de marcação e se baseiam, apenas, na base de conhecimento existente no padrão.

Com base no gráfico correspondente apresentado na figura 10, no entanto, é possível afirmar que existe uma diferença estatisticamente comprovada de que os resultados do RISO-TT são superiores às demais ferramentas analisadas, isso porque não existe intersecção e ainda apresenta valores superiores em seu *boxplot* com as demais amostras.

Figura 10: *BoxPlot* do Intervalo de confiança



Provavelmente essa superioridade se deve ao número de relações existentes entre os padrões temporais mapeados e suas relações definidas pelas regras na ferramenta RISO-TT. O processamento das informações, devido a esse número de relações existentes no dicionário de regras, fez com que as marcações das expressões temporais sejam mais detalhadas do que as demais.

Com a possibilidade de extensão do RISO-TT, os padrões que não foram mapeados puderam ser inseridos na formação das regras, isso possibilitou que qualquer padrão não mapeado fosse inserido para o processamento dos documentos.

A ferramenta SuTime também reconheceu algumas das expressões temporais em um formato para datas com três caracteres numéricos (Ex: 200 DC, conforme já mencionado anteriormente) assim como o RISO-TT, mas não

reconheceu os intervalos entre essas datas. Já o HeideTime não reconheceu o padrão mencionado.

6.3.3 Ameaças à validade

Para analisar o experimento foram considerados os quatro tipos de ameaça: externa, interna, construção e de conclusão. A validação externa relaciona-se à veracidade aproximada das conclusões e à generalização para o mundo real; a validade interna verifica se o resultado obtido é consequência da manipulação que foi feita, e não, de outro fator externo; a validade de construção se refere a problemas na elaboração e no controle do experimento; e a validade de conclusão trata da correlação entre o que foi medido e as conclusões obtidas.

Validade externa: O experimento foi realizado com um único conjunto finito de dados (*Corpora WikiWars*). Os resultados encontrados representam o escopo finito, e nada poderá ser afirmado com uma variação de dados diferentes sem que haja um novo experimento.

Validade interna: Os resultados encontrados podem ter sofrido influência de fatores não avaliados no experimento, como os de qualidade dos documentos e das marcações realizadas pelos especialistas. Nenhum ajuste aos parâmetros-padrão das ferramentas de marcação foi feito para uma nova avaliação.

Validade de construção: A técnica utilizada, as tecnologias e a estrutura do conteúdo utilizado no experimento podem ter características particulares e que não se aplicam a todos os casos analisados.

Validade de conclusão: Não foi analisado o uso de outras ferramentas de marcação de documentos nem todas as ferramentas disponíveis na literatura. Isso significa que não se garante que os resultados se aplicam a todas as demais tecnologias. O motivo pelo qual não foi possível a análise foi a não disponibilidade das ferramentas.

CAPÍTULO 7

CONCLUSÕES

O desenvolvimento do RISO-TT faz parte do projeto de pesquisa RISO. Esse experimento teve o intuito de avaliar se a metodologia de desenvolvimento utilizada em sua construção seria extensível e flexível e se houve alguma contribuição no cenário das ferramentas de marcação de expressões temporais. Para isso, foi utilizada a coleção *WikiWars*, pois a base dos experimentos realizados em todos os módulos do projeto RISO são os documentos da Wikipédia e por se tratar de uma Corpora com uma variação de padrões temporais e documentos expressivos.

No processo de verificação e validação desta pesquisa, foram consideradas duas perguntas: I) se a ferramenta era extensível e flexível, e II) se os resultados das marcações das expressões temporais são estatisticamente melhores do que as concorrentes *HeidelTime* e *SUTime*, ferramentas desenvolvidas pelo *Database Systems Research Group* da Universidade de Heidelberg e Universidade de Stanford, respectivamente.

Para a pergunta I, foram rodados três ensaios e coletados os dados. Depois de uma análise, verificou-se que algumas expressões temporais não tinham sido mapeadas e que o motivo desse resultado foi que esses padrões não eram reconhecidos. Uma vez identificados, os padrões foram inseridos nas regras do RISO-TT, e uma nova rodada foi realizada, tendo como resultado a marcação dos novos padrões mapeados, o que prova que a ferramenta é extensível e flexível, já que se adapta a qualquer padrão desejado.

Para a pergunta II, os resultados das marcações dos documentos resultantes do processamento dos documentos da *WikiWars* pelo *HeidelTime*, *SUTime* e *RISO-TT* foram comparados ao gabarito existente na própria Corpora da *WikiWars*. Seus resultados foram tabulados, analisados e comparados estatisticamente, o que resultou na superioridade do marcador temporal RISO-TT. Essa superioridade é atribuída ao fato de haver um maior número de regras para reconhecimento das relações entre as expressões temporais e sintagmas nominais. Assim, com essa análise, os objetivos propostos foram atingidos.

Contribuições

Considerando as informações apresentadas ao longo dessa dissertação, as principais contribuições do RISO-TT são:

- a) Arquitetura flexível e extensível baseados em padrões e regras configuráveis por arquivo XML;
 - a. O Heidelberg e o SuTime podem ser extensíveis, no entanto não tem uma arquitetura flexível, pois dependem de software de terceiro em sua arquitetura.
- b) Independência de software de terceiro;
- c) Reconhecimento das expressões temporais baseadas na análise de prioridades das regras;
- d) Possibilidade de criação de estruturas complexas de associações gramaticais e temporais (Expressões Compostas);
- e) Estende os padrões do mercado com a possibilidade de arranjos e associações com outras expressões não temporais;
- f) Normalização de padrões temporais complexos, considerando intervalos entre *tokens* temporais.

Limite de escopo

Mesmo sabendo das contribuições apresentadas pelo RISO-TT, existem características que não foram levadas em consideração para essa etapa do projeto. As principais características mapeadas são:

- a) Tempo de processamento;
 - i. Devido ao número de padrões existentes no RISO-TT, o processamento dos documentos se torna mais lento do que as ferramentas comparadas.
- b) Eventos e objetos temporais

- i. A identificação de eventos e de objetos com dimensão temporal representa um campo vasto na identificação e classificação de sintagmas a ser explorado.
- c) Normalização
- i. Um número significativo de expressões temporais ainda não foi normalizado. Esta atividade está nos trabalhos futuros.
- d) Ambiguidade;
- i. O tratamento de ambiguidade semântica está relacionado aos trabalhos futuros.

Trabalhos futuros

Considerando todos os aspectos abordados nesta pesquisa, foram elencadas as seguintes atividades para trabalhos futuros:

- a) Reconhecimento de objetos e eventos temporais
 - a. Para cada sintagma descrevendo um objeto ou evento em uma frase contendo uma expressão temporal, determinação de sua respectiva temporalidade a partir das expressões temporais.
- b) Análise e reconhecimento de padrões espaciais:
 - a. A evolução do RISO-TT, que trabalha com expressões temporais, para um marcador de expressões Espaço-Temporais denominado como RISO-ET. Esse projeto já está em andamento e será tema para a tese de um novo aluno de Mestrado e um projeto de extensão na Faculdade FACISA.
- c) Conclusão do módulo de normalização de expressões temporais:
 - a. No que diz respeito ao processo de normalização das expressões temporais, a ferramenta ainda precisa ser concluída.
- d) Reconhecimento e tratamento de ambiguidade das expressões temporais encontrada no documento.
- e) Tentar determinar algoritmos para melhorar a eficiência do processamento do casamento entre regras e expressões.

Empacotamento do RISO-TT

O projeto do RISO-TT está disponível no site do projeto RISO (<https://sites.google.com/a/copin.ufcg.edu.br/riso-t/projeto>). Para as demais ferramentas utilizadas nessa pesquisa, acesse:

- a) HeidelTime: <https://code.google.com/p/heideltime/>
- b) SUTime: <http://nlp.stanford.edu/software/sutime.shtml>
- c) WikiWars: <http://www.timexportal.info/wikiwars/>

REFERÊNCIAS

AHN, D.; VAN RANTWIJK, J.; DE RIJKE, M. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In: **Proceedings NAACL-HLT**, April 2007.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. Boston: Addison-Wesley Longman, 1999.

BISPO, M.C.T. **Criação de Vetores Temáticos de Domínios para a Desambiguação Polissêmica de Termos**. 2012. 130 f. Dissertação (Mestrado em Computação) – Universidade Federal de Campina Grande - UFCG, Campina Grande. 2012.

CARDOSO, N. **Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas**. In: II Simpósio Doutoral da Linguatca, 2006. Faculdade de Ciências - Universidade de Lisboa, 10-11 abr., 2006.

CUNNINGHAM, H. **JAPE: a Java Annotation Patterns Engine**. Research Memorandum CS-99-06, Department of Computer Science, University of Sheffield, May, 1999.

CHANG, A. X.; MANNING C. D. **SUTIME: A Library for Recognizing and Normalizing Time Expressions**. 8th International Conference on Language Resources and Evaluation (LREC 2012), 2012.

FARIA, C., GIRARDI, R. Um processo semi-automático para o povoamento de ontologias a partir de fontes textuais. In: **iSys - Revista Brasileira de Sistemas de Informação**, v.. 3, 2010.

FERNEDA, E. (18.04.2011) **Processamento da linguagem natural**. Disponível em: <<http://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/MRI-06%20-%20Processamento%20da%20Liguagem%20Natural.pdf>>. Acesso em: 12/11/2012.

FERRO, L et al. **TIDES - 2005 Standard for the Annotation of Temporal Expressions**, MITRE Corporation. 2005.

HACIOGLU, K; CHEN, Y; DOUGLAS, B.. Automatic time expression labeling for english and chinese text. In: GELBUKH, A. F. **Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing'05, Lecture Notes in Computer Science**, p. 548-559, Mexico City, Mexico, February. Springer, 2005.

HAGÈGE, C.; BAPTISTA, J.; MAMEDE, N. Caracterização e processamento de expressões temporais em português. In: **Linguamática**, v. 2, n. 1, p. 63-77, abr, 2010.

HAGÈGE, C.; BAPTISTA, J.; MAMEDE, N. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. In: MOTA, C.; SANTOS, D. (Eds.) **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o segundo HAREM**. São Paulo: Linguatca, 2008. IEEE - Institute of Electrical and Electronics Engineers. (2012) Draft standard for e language reference. Disponível em: http://www.ieee1647.org/downloads/D0.1-PDF-files/temporal_expressions.pdf. Acesso em: 05 jan, 2013.

LLORENS, H. S. **A Semantic Approach to Temporal Information Processing**. 2011. (PhD Dissertation) - University of Alicante, Departamento de Lenguajes y Sistemas Informáticos, Alicante, 2011.

LLORENS, H. S., E. e NAVARRO, B. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In: **Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics - ACL 2010**, p. 284–291, Uppsala, Sweden, 15-16 July, 2010.

LOUREIRO, J. M. S. **Reconhecimento de entidades mencionadas (obra, valor, relações de parentesco e tempo) e normalização de expressões temporais**. Lisboa: Instituto Superior Técnico da Universidade de Lisboa, 2007.

MANI, I. e WILSON, G. Processing of News. In: **Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, (ACL 2000)**, p. 69-76. 2000.

MANNING, C. D.; RAGHAVAN, P; SCHUTZE, H. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008.

MARSIC, G. **Temporal Processing of News: annotation of temporal expressions, verbal events and temporal relations**. 2011. (PhD Thesis). University of Wolverhampton, Wolverhampton, 2011.

MARTÍNEZ, H. L. **A Semantic Approach to Temporal Information Processing**. 2011. University of Alicante - Departamento de Lenguajes y Sistemas Informáticos, Alicante, 2011.

MAZUR, P.; DALE, R. WikiWars: A New Corpus for Research on Temporal Expressions. In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics**, p. 913–922, MIT, Massachusetts, USA, 9-11 October, 2010.

MELO, R. S. et al. **Dados semi-estruturados**. Material de um tutorial para o SBBD2000, 2000

NEGRI, M.; MARSEGLIA, L. **Recognition and Normalization of Time Expressions**: ITC-irst ar TERN 2004, 2004.

OLIVEIRA, F. A. D. DE. **Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua**

portuguesa. Universidade Federal do Rio Grande do Sul - Instituto de Informática. Programa de Pós-graduação em Computação, (2012). Disponível em: <<http://www.inf.ufrgs.br/gppd/disc/cmp135/trabs/992/Parser/parser.html>>. Acesso em: 10/11/2012.

ROYSTON, P. Remark AS R94: A remark on Algorithm AS 181: The W test for normality. *Applied Statistics*, 44, 547–551. 1995.

ROMÃO, L. C. da S. **Reconhecimento de entidades mencionadas em língua portuguesa: locais, pessoas, organizações e acontecimentos**. 2007. (Dissertação de Mestrado). Instituto Superior Técnico - Universidade Técnica de Lisboa, Lisboa, 2007.

SALTON, G. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

SAQUETE, E. ID 392:TERSEO + T2T3 Transducer. A System for Recognizing and Normalizing TIMEX3. In: **Proceedings of the 5th International Workshop on Semantic Evaluation**, 2010.

SCHIEL, U. **Aspectos temporais em Sistemas de Informação**. Relatório Técnico DSC-001/96, Universidade Federal da Paraíba - UFCG, 1996.

STROTGEN, J.; GERTZ, M. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In: **Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics - ACL 2010**, p. 321–324, Uppsala, Sweden, 15-16 July 2010. FERRO, L. et al. TIDES: 2005 Standard for the Annotation of Temporal Expressions, 2005.

TIMEML WORKING GROUP. **Guidelines for Temporal Expression Annotation for English for TempEval 2010**, 2009.

TIMEML. (2012). Disponível em: <<http://www.timeml.org/site/index.html>>. Acesso em: 20 out, 2012.

VERHAGEN, M. **Temporal closure in an annotation environment**. *Language Resources and Evaluation*, n. 39, p. 211–241, May, 2005.

VERHAGEN, M. et al. Automating temporal annotation with TARSQI. In: **Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics**, Ann Arbor, USA, 2005.

ANEXO A – Exemplo de documento da WikiWars

O texto seguinte foi extraído do documento “21_NigerianCivilWar.sgm” presente na corpora WikiWars.

```
<DOC>
<DOCID>21_NigerianCivilWar</DOCID>
<DOCTYPE SOURCE="21_wikipedia"> HISTORY ARTICLE </DOCTYPE>
<DATETIME> 2009-12-19T17:00:00 </DATETIME>
<TEXT>
```

Breakaway

The military governor of the Igbo-dominated southeast, Colonel Odumegwu Ojukwu, citing the northern massacres and electoral fraud, proclaimed with southern parliament the secession of the south-eastern region from Nigeria as the Republic of Biafra, an independent nation on 30 May 1967. Although there was much sympathy in Europe and elsewhere, only five countries recognized the new republic.

Several peace accords especially the one held at Aburi, Ghana (the Aburi Accord) collapsed and the shooting war followed. Ojukwu managed at Aburi to get agreement to a confederation for Nigeria, rather than a federation. He was warned by his advisers that this reflected a failure of Gowon to understand the difference and, that being the case, predicted that it would be reneged upon. When this happened, Ojukwu regarded it as both a failure by Gowon to keep to the spirit of the Aburi agreement, and lack of integrity on the side of Nigeria Military Government in the negotiations toward a united Nigeria. Gowon's advisers, to the contrary, felt that he had enacted as much as was politically feasible in fulfilment of the spirit of Aburi.

Civil War

The Nigerian government launched a "police action" to retake the secessionist territory. The war began on 6 July 1967 when Nigerian Federal troops advanced in two columns into Biafra. The Nigerian army offensive was through the north of Biafra led by Colonel Shuwa and the local military units were formed as the 1st Infantry Division. The division was led mostly by northern officers. The right-hand Nigerian column advanced on the town of Nsukka which fell on 14 July, while the left-hand column made for Garkem, which was captured on 12 July. At this stage of the war,

the other regions of Nigeria (the West and Mid-West) still considered the war as a confrontation between the north (mainly Hausas) against the east (mainly Igbos). But the Biafrans responded with an offensive of their own when, on 9 July, the Biafran forces moved west into the Mid-Western Nigerian region across the Niger river, passing through Benin City, till they were stopped at Ore (in present day Ondo State) just over the state boundary on 21 August, just 130 miles east of the Nigerian capital of Lagos. The Biafran attack was led by Lt. Col. Banjo with the Biafran rank of brigadier. The attack met little resistance and the Mid-West was easily taken over. This was due to the pre-secession arrangement that all soldiers should return to their regions to stop the spate of killings, in which Igbo soldiers had been major victims. The Nigerian soldiers that were supposed to defend the Mid-West state were mostly Mid-West Igbo and were in touch with their eastern counterpart. General Gowon responded by asking Colonel Murtala Mohammed to form another division (the 2nd Infantry Division) to expel the Biafrans from the Mid-West, as well as defend the West side and attack Biafra from the West as well. Colonel Murtala later became military head of state. As Nigerian forces retook the Mid-West, the Biafran military administrator declared the Republic of Benin on 19 September.

Although Benin City was retaken by the Nigerians on 22 September, the Biafrans succeeded in their primary objective by tying down as many Nigerian Federal troops as much as they could. Gen. Gowon also launched an offensive into Biafra south from the Niger Delta to the riverine area using the bulk of the Lagos Garrison command under Colonel Benjamin Adekunle (called the Black Scorpion) to form the 3rd Infantry Division (which was later renamed as the 3rd Marine Commando which was made up of the Nigerian marines). As the war continued, the Nigerian Army recruited amongst a wider area, including the Yoruba and Edo. Four battalions of the Nigerian 2nd Infantry Division were needed to drive the Biafrans back and eliminate their territorial gains made during the offensive. The Nigerians were repulsed three times as they attempted to cross the River Niger during October, resulting in the loss of thousands of troops, dozens of tanks and equipment. The first attempt by the 2nd Infantry Division on 12 October to cross the Niger from the town of Asaba to the Biafran city of Onitsha cost the Nigerian Federal Army over 5,000 soldiers killed, wounded, captured or missing.

Anexo B – Arquivo de configuração do RISO-TT

O RISO-TT utiliza um arquivo de configuração que mapeia todos os padrões e regras que serão utilizados no processamento dos documentos. Abaixo, está disponível a arquitetura desse arquivo.

```
<?xml version="1.0"?>
<configuracao>
  <padroes>
    <padrao arquivo="padroes/estrutura_pre_temporal.xml"/>
  <padrao arquivo="padroes/unidade_temporal.xml"/>
    <padrao arquivo="padroes/unidade_numeral.xml"/>
    <padrao arquivo="padroes/hora.xml"/>
    <padrao arquivo="padroes/dia.xml"/>
    <padrao arquivo="padroes/data_numeral.xml"/>
    <padrao arquivo="padroes/mes.xml"/>
    <padrao arquivo="padroes/ano.xml"/>
    <padrao arquivo="padroes/datas_especiais.xml"/>
    <padrao arquivo="padroes/partes_do_ano.xml"/>
    <padrao arquivo="padroes/estacao_do_ano.xml"/>
    <padrao arquivo="padroes/parte_do_dia.xml"/>
    <padrao arquivo="padroes/parte_da_semana.xml"/>
    <padrao arquivo="padroes/expressoes.xml"/>
    <padrao arquivo="padroes/preposicoes.xml"/>
    <padrao arquivo="padroes/adverbio.xml"/>
  <padrao arquivo="padroes/curtas_expressoes.xml"/>
  </padroes>
  <regras>
    <regra arquivo="regras/expressao_temporal.xml"/>
  </regras>
</configuracao>
```

ANEXO C – Exemplos de padrões mapeados para o RISO-TT

Conforme mencionado ao longo dessa dissertação, o RISO-TT utiliza de padrões para as composições das regras. As informações para obter a lista completa dos padrões mapeados no RISO-TT estão disponíveis na sessão de Empacotamento do Modelo, presente na conclusão desse trabalho.

Um padrão é formado pela estrutura:

```
<simbolo nome="NOME DO PADRÃO">
<expressao>EXPRESSÃO</expressao>
</simbolo>
```

Uma expressão pode ser formada por associações de padrões, espaço e caracteres. Abaixo, seguem alguns exemplos de padrões.

```
<simbolo nome="estrutura_basica_temporal"; sigla="EBT" >
  <expressao>mes dia "," ano</expressao>
    <expressao>mes dia "," ano</expressao>
  <expressao>mes dia ano</expressao>
    <expressao>mes dia "of" ano</expressao>
    <expressao>mes dia "-" mes dia ano"s"</expressao>
    <expressao>mes dia "-"mes dia ano"s"</expressao>
    <expressao>mes dia "-" mes dia ano</expressao>
    <expressao>mes dia "-"mes dia ano</expressao>
    <expressao>mes dia "-" mes dia</expressao>
    <expressao>mes dia "-"mes dia</expressao>
    <expressao>dia mes "," ano</expressao>
    <expressao>dia mes ano</expressao>
  <expressao>mes "," ano</expressao>
    <expressao>mes "," ano</expressao>
    <expressao>dia mes "-" dia mes ano"s"</expressao>
    <expressao>dia mes "-"dia mes ano"s"</expressao>
    <expressao>dia mes "-" dia mes ano</expressao>
```

<expressao>dia mes "-" dia mes ano</expressao>
 <expressao>dia mes "-" dia mes</expressao>
 <expressao>dia mes "-" dia mes</expressao>
 <expressao>dia "and" dia mes</expressao>
 <expressao>dia mes</expressao>
 <expressao>mes "of" ano</expressao>
 <expressao>mes "and" mes ano</expressao>
 <expressao>dia "of" ano</expressao>
 <expressao>mes "/" mes ano</expressao>
 <expressao>mes ano</expressao>
 <expressao>mes dia "-" dia</expressao>
 <expressao>mes dia</expressao>
 <expressao>dia "-" dia mes</expressao>
 <expressao>mes</expressao>
 </simbolo>

<simbolo nome="estrutura_minima_temporal"; sigla="EMT">
 <expressao>ano "-" ano</expressao>
 <expressao>ano "and" ano</expressao>
 <expressao>ano "-" dia</expressao>
 <expressao>ano "s"</expressao>
 <expressao>ano</expressao>
 </simbolo>

<simbolo nome="numeros_de_0_a_999"; sigla="??">
 <expressao>"from [0-9]{3} to [0-9]{3} BC"</expressao>
 <expressao>"[0-9]{3}-[0-9]{3} BC"</expressao>
 <expressao>"[0-9]{3} and [0-9]{3} BC"</expressao>
 <expressao>preposicoes "[0-9]{3} BC"</expressao>
 <expressao>mes "[0-9]{3} BC"</expressao>
 <expressao>"the end of the [0-9]{3} BC"</expressao>
 <expressao>"[0-9]{3} BC"</expressao>
 </simbolo>

<simbolo nome="intervalos"; sigla="??">

<expressao>unidade_temporal "from" estrutura_basica_temporal "to end of" estrutura_basica_temporal</expressao>

<expressao>unidade_temporal "from" unidade_numeral unidade_temporal "to end of" unidade_numeral unidade_temporal</expressao>

<expressao>unidade_temporal "from" unidade_temporal "to" unidade_temporal</expressao>

<expressao>unidade_temporal "from" estrutura_basica_temporal "to end of" unidade_numeral unidade_temporal</expressao>

<expressao>unidade_temporal "from" estrutura_basica_temporal "to end of" unidade_temporal</expressao>

<expressao>unidade_temporal "from" unidade_numeral unidade_temporal "to end of" estrutura_basica_temporal</expressao>

<expressao>unidade_temporal "from" unidade_numeral unidade_temporal "to end of" unidade_temporal</expressao>

<expressao>unidade_temporal "from" unidade_numeral unidade_temporal "to end of" unidade_numeral unidade_temporal</expressao>

<expressao>unidade_temporal "from" unidade_temporal "to" unidade_temporal</expressao>

<expressao>unidade_temporal "from" unidade_temporal "to" unidade_numeral unidade_temporal</expressao>

<expressao>unidade_temporal "from" unidade_temporal "to" estrutura_basica_temporal</expressao>

<expressao>"from" estrutura_basica_temporal "to" estrutura_basica_temporal"," ano</expressao>

<expressao>"from" estrutura_basica_temporal "to" estrutura_basica_temporal</expressao>

<expressao>"from" estrutura_minima_temporal "to" estrutura_minima_temporal"," ano</expressao>

<expressao>"from" estrutura_minima_temporal "to" estrutura_minima_temporal</expressao>

<expressao>"from" estrutura_basica_temporal "until"


```

estrutura_basica_temporal</expressao>
    <expressao>"from"          estrutura_minima_temporal          "until"
estrutura_minima_temporal</expressao>
    <expressao>"from" mes "to" mes</expressao>
    <expressao>"from" ano "to" ano</expressao>
    <expressao>"from" ano "through" ano</expressao>
    <expressao>preposicoes    mes    "and    again"    preposicoes
mes</expressao>
    <expressao>preposicoes mes "of that" unidade_temporal</expressao>
    <expressao>"a          few"          unidade_temporal          "and"
unidade_temporal</expressao>
    <expressao>preposicoes estacao_do_ano "of" ano</expressao>
</simbolo>

```

ANEXO D – Arquivo de regras do RISO-TT

Uma regra é formada por um conjunto de associações de padrões e ordenada de acordo com a prioridade, conforme visto na definição de regras. A estrutura das regras é formada por:

```
<regras>
  <simbolo nome="EXPRESSÃO TEMPORAL">
    <expressao tipo="TIPO DA EXPRESSÃO"> EXPRESSÃO
  </expressao>
</simbolo>
</regras>
```

Onde:

EXPRESSÃO TEMPORAL é o nome da expressão temporal;

TIPO DA EXPRESSÃO é o nome dado à associação entre os padrões;

EXPRESSÃO é o valor da expressão formada pela associação dos padrões.

Abaixo, segue o arquivo de regras utilizadas no experimento dessa dissertação.

```
<?xml version="1.0"?>
<regras>
  <simbolo nome="expressao_temporal">
    <expressao tipo="I">intervalos</expressao>
    <expressao          tipo="EPT-EBT-Adv">estrutura_pre_temporal
estrutura_basica_temporal adverbio</expressao>
    <expressao          tipo="EPT-EBT-UT">estrutura_pre_temporal
estrutura_basica_temporal unidade_temporal</expressao>
    <expressao          tipo="EPT-UT-Adv">estrutura_pre_temporal
unidade_temporal adverbio</expressao>
```

```

      <expressao                                tipo="Pre-EBT-Adv">preposicoes
estrutura_basica_temporal adverbio</expressao>
      <expressao                                tipo="Pre-EBT-UT">preposicoes
estrutura_basica_temporal unidade_temporal</expressao>
      <expressao                                tipo="EPT-EMT-Adv">estrutura_pre_temporal
estrutura_minima_temporal adverbio</expressao>
      <expressao                                tipo="EPT-EMT-UT">estrutura_pre_temporal
estrutura_minima_temporal unidade_temporal</expressao>
      <expressao      tipo="Pre-EA-A">preposicoes      estacao_do_ano
ano</expressao>
      <expressao                                tipo="EPT-EBT">estrutura_pre_temporal
estrutura_basica_temporal</expressao>
      <expressao                                tipo="Pre-EBT">preposicoes
estrutura_basica_temporal</expressao>
      <expressao                                tipo="Pre-EMT">preposicoes
estrutura_minima_temporal</expressao>
      <expressao                                tipo="EPT-UT">estrutura_pre_temporal
unidade_temporal</expressao>
      <expressao                                tipo="EPT-EMT">estrutura_pre_temporal
estrutura_minima_temporal</expressao>
      <expressao tipo="Pre-A-Adv">preposicoes ano adverbio</expressao>
      <expressao                                tipo="EBT-UT">estrutura_basica_temporal
unidade_temporal</expressao>
      <expressao                                tipo="EBT-Adv">estrutura_basica_temporal
adverbio</expressao>
      <expressao      tipo="D-EMT-Adv">dia      unidade_temporal
adverbio</expressao>
      <expressao tipo="Pre-A-Adv">preposicoes ano adverbio</expressao>
      <expressao                                tipo="EPT-EN">estrutura_pre_temporal
estacao_do_ano</expressao>
      <expressao tipo="Pre-A">preposicoes ano</expressao>
      <expressao      tipo="D-EMT-A">dia      unidade_temporal      "of"
ano</expressao>

```

```
<expressao tipo="D-EMT-A">dia unidade_temporal</expressao>  
<expressao tipo="EBT">estrutura_basica_temporal</expressao>  
<expressao tipo="EMT">estrutura_minima_temporal</expressao>  
<expressao tipo="EBT-H">preposicoes_hora</expressao>  
<expressao tipo="EBT-N">numeros_de_0_a_999</expressao>  
<expressao tipo="DE">datas_especiais</expressao>  
<expressao tipo="CE">curtas_expressoes</expressao>  
</simbolo>  
</regras>
```

ANEXO E – Exemplos de arquivos processados pelo RISO-TT

Fragmento do documento "21_NigerianCivilWar.sgm" da coleção WikiWars

The Nigerian federal forces launched their final offensive against the Biafrans on 23 December 1969 with a major thrust by the 3rd Marine Commando Division (the division was commanded by Col. Obasanjo, who later became president twice) which succeeded in splitting the Biafran enclave into two by the end of the year. The final Nigerian offensive, named "Operation Tail-Wind", was launched on 7 January 1970 with the 3rd Marine Commando Division attacking, and supported by the 1st Infantry division to the north and the 2nd Infantry division to the south. The Biafran town of Owerri fell on 9 January, and Uli fell on 11 January. The war finally ended with the final surrender of the Biafran forces in the last Biafra-held town of Amichi on 13 January 1970. Only a few days earlier, Ojukwu fled into exile by flying by plane to the republic of Côte d'Ivoire, leaving his deputy Philip Effiong to handle the details of the surrender to Yakubu Gowon of the federal army.

Saída produzida pela função TAG

The Nigerian federal forces launched their final offensive against the Biafrans <RISOTime type=Pre-EBT>on 23 December 1969</RISOTime> with a major thrust by the 3rd Marine Commando Division (the division was commanded by Col. Obasanjo, who later became president twice) which succeeded in splitting the Biafran enclave into two <RISOTime type=EPT-UT>by the end of the year</RISOTime>. The final Nigerian offensive, named "Operation Tail-Wind", was launched <RISOTime type=Pre-EBT>on 7 January 1970</RISOTime> with the 3rd Marine Commando Division attacking, and supported by the 1st Infantry division to the north and the 2nd Infantry division to the south. The Biafran town of Owerri fell <RISOTime type=Pre-EBT>on 9 January</RISOTime>, and Uli fell <RISOTime type=Pre-EBT>on 11 January</RISOTime>. The war finally ended with the final

surrender of the Biafran forces in the last Biafra-held town of Amichi <RISOTime type=Pre-EBT>on 13 January 1970</RISOTime>. Only <RISOTime type=EPT-UT-Adv>a few days earlier</RISOTime>, Ojukwu fled into exile by flying by plane to the republic of Côte d'Ivoire, leaving his deputy Philip Effiong to handle the details of the surrender to Yakubu Gowon of the federal army.

Saída produzida pela função LIST

Pre-EBT -> on 23 December 1969

EPT-UT -> by the end of the year

Pre-EBT -> on 7 January 1970

Pre-EBT -> on 9 January

Pre-EBT -> on 11 January

Pre-EBT -> on 13 January 1970

EPT-UT-Adv -> a few days earlier

Saída produzida pela função NORM

on 23 December 1969 <--> 23-12-1969

by the end of the year <--> Padrão EPT-UT ainda não foi mapeado

on 7 January 1970 <--> 07-01-1970

on 9 January <--> 9-01-XXXX

on 11 January <--> 11-01-XXXX

on 13 January 1970 <--> 13-01-1970

a few days earlier <--> Padrão EPT-UT-Adv ainda não foi mapeado.