Caracterização e Processamento de Expressões Temporais em Português

Caroline Hagège
Xerox Research Centre Europe – XRCE
6 Chemin de Maupertuis – Meylan – France
Caroline.Hagege@xrce.xerox.com

Jorge Baptista
Universidade do Algarve, FCHS
L2F, INESC-ID Lisboa
Campus de Gambelas – Faro – Portugal
jbaptis@ualg.pt

Nuno Mamede
Instituto Superior Técnico
L2F, INESC-ID Lisboa
Rua Alves Redol, 9 – Lisboa – Portugal
Nuno.Mamede@inesc-id.pt

Resumo

A dimensão temporal é um elemento estruturante fundamental para a informação veiculada em textos e constitui um desafio para o processamento de língua natural, sendo igualmente importante para muitas aplicações do processamento das línguas. Este artigo constitui mais um passo para o ambicioso objectivo de tratamento da informação temporal. Para tal, apresenta-se uma proposta de classificação das expressões temporais do Português que permita esclarecer algumas incertezas relativas ao estatuto de diferentes expressões temporais e constitui uma base para a anotação destas expressões. Utilizando esta classificação, foi desenvolvida uma ferramenta de anotação automática das expressões temporais do Português, cujo desempenho foi avaliado.

1 Introdução

A descrição do tempo, assim como os processos de inferência que levam em conta a informação temporal, são assuntos que há muito tempo despertaram interesse em áreas tão diversas como a lógica, a filosofia e a linguística. Reichenbach em (Reichenbach, 1947) propõe um sistema explicativo dos tempos verbais utilizando três pontos de referência temporal: o tempo do evento (E), o tempo de referência (R) e o tempo do discurso (S). Nos anos 50, Prior em (Prior, 1957) propõe um teoria de lógica temporal, onde introduz a representação formal dos tempos usando operadores temporais.

Mais recentemente têm aparecido novos trabalhos relacionados com processos de inferência temporal. Um dos mais conhecidos em Inteligência Atificial (IA) e Processamento de Linguagem Natural (PLN) é o trabalho de Allen descrito em (Allen, 1991). Só muito recentemente, porém, apareceram os primeiros sistemas que fazem efectivamente algum tipo de processamento da informação temporal. Esta nova tendência foi impulsionada pelo facto de um tratamento adequado da componente temporal em textos permitir um melhor desempenho numa ampla gama de tarefas, tais como a resposta a perguntas, a sumarização (uni- e multidocumento) e, de um modo geral, a extracção de informação a partir de documentos. Um dos factores que ajudaram a desenvolver este renovado interesse pelo processamento de expressões temporais (ET) foi a criação do projecto TimeML (Saurí et al., 2006). Este projecto fornece um conjunto de directrizes para a anotação de expressões temporais e de eventos para o Inglês. Estas orientações foram adaptadas para o Francês (ver (Bittar, 2008)), para o Italiano e para o Romeno. Outras abordagens para a descrição e normalização de expressões temporais são apresentadas por Battistelli, Minel e Schwer em (Battistelli, Minel e Schwer, 2006). Nesta última abordagem, o tratamento temporal tem como finalidade ser usado por um sistema de navegação temporal nos textos. Para obter uma representação adequada da informação temporal, um subconjunto de expressões (designadas expressões temporais) são descritas como termos e sua composição é feita através de operadores pré-definidos.

Têm vindo a ser desenvolvidos alguns sistemas automáticos dedicados à anotação temporal e, recentemente, foi organizado um concurso para avaliar a precisão dos processadores tem-

porais automáticos para Inglês (Verhagen et al., 2007). As anotações baseiam-se nas directrizes TimeML mencionadas anteriormente e a maioria dos sistemas utiliza técnicas de aprendizagem automática e *corpora* anotados para o treino supervisionado. Infelizmente, este tipo de *corpora* com anotações temporais só estão disponíveis para o Inglês (TimeBank). O facto de um recurso como este exigir um grande esforço em termos de recursos humanos para a sua construção explica, naturalmente, a falta de recursos equivalentes noutras línguas. Mais recentemente, Parent e colegas (Parent, Gagnon e Muller, 2008) e Hagège e Tannier ((Hagège e Tannier, 2008) apresentaram sistemas baseados em regras para a anotação e normalização de expressões temporais em Francês e Inglês. Para Português, um primeiro passo para a anotação temporal foi realizado no âmbito do Segundo HAREM (Mota e Santos, 2008).

O nosso objectivo a longo prazo é o desenvolvimento de um sistema capaz de ancorar e de ordenar temporalmente os processos expressos nos textos. Para alcançar este objectivo, é necessário dar os seguintes passos:

- identificação e etiquetagem das expressões temporais que ocorrem nos textos;
- resolução das expressões temporais referenciais para que se possa proceder à sua normalização;
- identificação dos eventos associados às expressões temporais;
- caracterização das relações entre eventos e expressões temporais, o que inclui normalmente considerar o tempo, o aspecto e a modalidade;
- realização de inferência temporal.

O significado das ET referenciais não pode ser obtido directamente a partir dos elementos da expressão, requerendo algum tipo de cálculo quanto à sua referência temporal. A normalização das ET consiste, justamente, em representar esse valor de uma forma que permita esse cálculo.

Estes passos são, contudo, estreitamente interdependentes, visando um tratamento adequado da temporalidade. Este artigo aborda algumas das questões acima mencionadas. É proposto um conjunto de orientações para lidar com a identificação e etiquetagem de expressões temporais que aparecem em textos em Português. Nessa caracterização, as diferenças de estatuto referencial dessas expressões são levadas em consideração, pois levam à utilização de diferentes métodos para a normalização das expressões temporais. Desenvolvemos uma ferramenta para etiquetar automaticamente as expressões temporais de acordo com essas orientações e para realizar uma primeira etapa de normalização temporal.

O artigo começa por explicar as motivações para este trabalho, que teve um forte impulso a partir da participação na campanha de avaliação conjunta do Segundo HAREM (Mota e Santos, 2008), mostrando que uma caracterização adequada das expressões temporais não é uma tarefa trivial e que precisa de levar em consideração não apenas os elementos lexicais por que as ET são formadas mas, e de forma fundamental, também o contexto mais amplo em que se elas se encontram inseridas, por forma a que esta tarefa possa ser adequadamente executada. Serão, então, sucintamente apresentadas as directrizes para a anotação de entidades temporais e explicitaremos em que aspectos nos demarcamos das directrizes do projecto TimeML. Finalmente, apresenta-se o anotador temporal por nós desenvolvido e os resultados obtidos com o sistema naquela campanha de avaliação.

2 Motivação

A fim de motivar a dificuldade da tarefa de anotação temporal apresentam-se os seguintes exemplos:

(1) *Banana* ***de manhã*** *emagrece. Será?*
(2) *Partiu* ***esta manhã***
(3) ***A manhã*** *é um momento mágico do dia*
(4) ***Numa bela manhã****, resolveu partir*

Todas estas expressões são sintagmas nominais (SN) ou preposicionais (SP) que têm a mesma cabeça lexical (*manhã*). Mas, cada expressão tem de ser interpretada de forma diferente. Como é afirmado por Ehrmann e Hagège em (Ehrmann e Hagège, 2009), não se pode realizar de forma adequada a interpretação de uma ET sem levar em conta as suas relações com os outros constituintes da frase.

No primeiro caso, a expressão *de manhã* tem de ser interpretada como um agregado temporal (isto é, uma expressão temporal que vai ancorar o processo associado mais do que uma vez na linha do tempo). Além do mais, este agregado temporal tem um período regular (a expressão é aproximadamente equivalente a *todas as manhãs*)[1].

[1] A análise deste tipo de situação é, porém, bastante complexa, já que se trata de uma construção elíptica, correspondendo à frase: (*alguém*) *comer banana de manhã emagrece* (*alguém*); o valor genérico de *banana* é dado pela sua determinação (determinante zero ou ausência de determinante), e o valor frequentativo de *comer banana*

No segundo caso, trata-se de uma ET referencial cujo antecedente é o momento da enunciação (ou seja, partiu na manhã do dia em que a frase foi produzida). No terceiro caso, trata-se de uma expressão genérica temporal. Isto significa que a expressão não fornece qualquer ancoragem temporal para o predicado associado. Finalmente, a última expressão é uma expressão temporal vaga: existe de facto uma ancoragem do processo associado na linha temporal, mas não se pode especificar de maneira precisa onde este se situa na linha do tempo.

Estes exemplos mostram claramente que um simples esquema de emparelhamento de padrões não é suficiente para realizar uma caracterização adequada de entidades temporais. Por esta razão, parece perfeitamente defensável o ponto de vista que rejeita a inclusão da tarefa de reconhecimento de ET como uma mera sub-tarefa da tarefa mais geral de Reconhecimento de Entidades Mencionadas (como tem sido feito, por exemplo, em (Mota e Santos, 2008)).

3 Directivas para a identificação e Classificação de ET em Português

Um dos pontos-chave nas directivas que aqui se apresentam é justamente a ideia de que as ET só podem ser devidamente classificadas e anotadas quando consideradas em relação aos processos que modificam. Apesar de tal observação poder parecer óbvia, mesmo nas orientações do projecto TimeML (Saurí et al., 2006) permanece alguma incerteza quanto ao estatuto das relações entre as ET e os processos que modificam enquanto factor determinante para a sua interpretação, especialmente quando as ET são citadas, sem qualquer contexto.

3.1 A nossa proposta vs. estado da arte

O nosso trabalho inscreve-se na linha geral do projecto TimeML, embora com algumas diferenças que explicitaremos e exemplificaremos já a seguir.

O projecto TimeML constitui sem nenhuma dúvida valioso contributo e incontornável referência no quadro do processamento da informação temporal. O TimeML propõe não só uma classificação e uma normalização das ET, mas também uma proposta de anotação da informação acerca dos processos (aspecto, tempo, modalidade), informação que deve ser tomada em consideração para o tratamento adequado da temporalidade.

A nossa proposta é mais modesta, pois, neste momento, preocupámos-nos exclusivamente com expressões temporais e não propusemos ainda qualquer anotação específica para representar a informação relevante associada aos processos modificados pelas ET. Assim, por exemplo, o problema do estatuto das ET associadas a predicados modificados por diferentes modalidades não foi sequer considerado nesta altura. Por outras palavras, não tentamos dar resposta à pergunta sobre como interpretar temporalmente a frase seguinte: *É possível que venham na próxima quarta-feira*, na qual não se sabe se o processo *venham* vai ocorrer ou não. Na nossa proposta, vamos circunscrever-nos ao problema da identificação e classificação das ET, procurando desde já avançar no sentido de uma normalização da informação temporal por elas veiculada. Nesta proposta, apresentamos critérios formais operativos e reprodutíveis para identificação, segmentação e classificação das ET. Neste sentido, salientam-se desde já os aspectos em que nos distinguimos das directivas do projecto TimeML, não deixando, no entanto, de o considerar como uma referência fundamental neste domínio.

Os pontos onde nos distanciamos do TimeML são os seguintes:

- Integração sistemática da preposição que introduz uma ET;
- Proposta de critérios formais, claros e reprodutíveis, para segmentação de ET complexas;
- Clara distinção entre a anotação e os passos intermédios necessários para a realizar.

3.1.1 Integração da preposição

Consideramos que a preposição que introduz o grupo preposicional (SP) que contém uma expressão temporal deve fazer parte integrante da ET. Esta posição distingue-se da solução apresentada pelo TimeML, que anota as preposições introdutoras de ET com a categoria *SIGNAL* e as separa da expressão temporal propriamente dita. As razões desta escolha são as seguintes:
A preposição é um elemento formal que muitas vezes permite caracterizar ou classificar de forma inequívoca a expressão temporal. Assim, por exemplo, em (*a partir de/até/desde*)*segunda-feira*, o significado das ET seguintes está estreitamente ligado à escolha da preposição que introduz o nome de tempo *segunda-feira*.

está muito provavelmente relacionado com o infinitivo e a redução de um sujeito genérico (*alguém*); do mesmo modo, o valor genérico deste emprego de *emagrecer* parece resultar do uso do presente do indicativo e da redução de um complemento directo indefinido (*alguém*).

De facto, as propriedades sintácticas da construção destes adjuntos adverbiais de tempo estão directamente relacionadas com a preposição. Assim, por exemplo, enquanto que com as preposições acima o nome de tempo não obriga à presença de um artigo, se se tiver a preposição *em*, o artigo torna-se obrigatório: *na segunda-feira/*em segunda-feira*. Também a inserção de um advérbio quantificador indefinido como *aproximadamente* não se verifica com todas as preposições: (*a partir de/até/desde/*em*)*aproximadamente segunda-feira*. Em segundo lugar, a mesma ET, quando traduzida para outra língua, pode ou não ser introduzida por preposição. Por exemplo, a expressão em Português *na segunda-feira* poderá ser traduzida simplesmente em Inglês por *Monday* (sem preposição) ou por *on Monday* (com preposição). Já em Francês não se admite qualquer preposição: (*le*) *lundi*.

3.1.2 Segmentação de ET complexas

No que diz respeito à segmentação de ET complexas, a norma TimeML não fornece elementos suficientes para decidir sem ambiguidade se uma expressão complexa deve ser considerada como uma única ET ou se deverá ser segmentada várias ET independentes. Neste sentido, iremos propor, como veremos, um conjunto de critérios sintácticos e semânticos que permitem tomar esta decisão de forma clara e reprodutível.

3.1.3 Distinção entre resultado da anotação e etapas de processamento para a anotação

Finalmente, as directrizes do TimeML obrigam em certos casos a indicar as etapas intermédias de anotações (que correspondem possivelmente a diferentes etapas de processamento automático das expressões temporais). Assim, para a expressão *two days before yesterday* em *John left two days before yesterday.*, o guia de anotação TimeML preconiza a anotação de *two days* com o tipo *DURATION*, a anotação de *before yesterday* com o tipo *DATE*, e finalmente uma anotação global da expressão *two days before yesterday*. Este forma de anotar parece-nos indesejável, já que inclui etapas intermédias antes de fornecer a anotação final. Efectivamente, consideramos que as directivas para anotação não devem pressupor os meios que poderão ser utilizados para alcançar a anotação preconizada. O facto de se introduzir possíveis passos intermédios para se chegar à anotação final obriga, de certa forma, os anotadores automáticos a seguir um certo algoritmo de anotação, o que ultrapassa claramente a função de directivas.

Estando feitas estas clarificações relativamente à nossa posição perante o estado da arte, apresentamos, nas secções seguintes, a nossa proposta de identificação e classificação das ET.

3.2 Identificação

Para identificar de forma objectiva expressões temporais, apresentam-se vários critérios. Uma expressão é uma ET quando satisfaz simultaneamente os critérios 1 e 2 ou, então, é considerada uma ET genérica, definida pelo critério 3:

1. **Critério 1** - uma expressão temporal em contexto pode responder adequadamente a uma das interrogativas *quando?*, *quanto tempo?*, eventualmente precedido de uma preposição, ou *com que frequência?*;
2. **Critério 2** - uma expressão temporal contém pelo menos uma unidade lexical que corresponda a um dos seguintes tipos:
   (a) uma data numérica ou alfanumérica (por exemplo, 21-Mar-2008), tanto para expressar as datas do calendário como os diferentes formatos de hora (12:30), incluindo as abreviaturas dos meses, e certas expressões adverbiais (por exemplo, *AM*, *GMT* e *a.C.*);
   (b) uma unidade de tempo (*segundo*); este conjunto inclui também unidades de tempo que não pertençam ao sistema internacional e que são de emprego "informal" como *fim-de-semana*;
   (c) os substantivos correspondentes à designação de algumas destas unidades de tempo, como os nomes dos meses (*Janeiro*), os dias da semana (*segunda-feira*) e advérbios derivados de unidades de tempo (*diariamente*);
   (d) os substantivos que designam festividades e efemérides de natureza religiosa, política, histórica ou cultural; o nome das estações do ano e o de dias festivos, que podem ou não incluir o substantivo *dia*;
   (e) advérbios de tempo simples e não ambíguos (por exemplo *ontem*) ou advérbios compostos (*depois de amanhã*), juntamente com advérbios tempo derivados, formados com o sufixo *-mente* (*futuramente*);
   (f) um grupo preposicional (SP), cuja cabeça é um substantivo de tempo genérico (por exemplo, *altura*, *data*, *instante*, *momento* e *vez*); estes substantivos são geralmente acompanhados

de diversos determinantes como, por exemplo, quantificadores de tipos diferentes (*por duas/várias/diversas vezes*), os pronomes demonstrativos (*nessa altura*), outros pronomes com função determinativa, inclusive pronomes possessivos (*no meu tempo*); podem também ser modificados por diferentes adjectivos e até por orações relativas (*na altura em que ela vivia em Lisboa*); n.b.: a ET não inclui a oração relativa; inclui-se ainda no conjunto dos modificadores os adjectivos (normalmente em maiúsculas), que designam um período histórico (*durante o período Barroco*); n.b. o adjectivo deverá ser considerado como estando incluído na ET;

(g) os chamados complementos determinativos envolvendo numerais e unidades de tempo que quantificam temporalmente um nome (predicativo) designativo de evento, estado ou processo (*uma viagem de 5 dias*); n.b.: a preposição *de* deve ser incluída na ET;

(h) os grupos preposicionais com unidades de tempo modificadas por vários adjectivos com valor referencial (*no ano passado*, *no próximo mês*, *durante o corrente ano* e *nos séculos vindouros*); ou em que as unidades de tempo se encontram modificadas por orações relativas envolvendo os verbos *passar*, *vir*, ou outros : *no ano que passou*, *para o mês que vem*;n.b.: estes verbos constituem um conjunto fechado;

(i) expressões com os verbos *fazer* ou *haver* e unidades de tempo: *há três anos*, *faz duas semanas*; n.b.: as expressões com *fazer* ou *ter* que indicam a idade (*O Pedro já fez/tem 18 anos*) não deverão ser classificadas como ET.

3. **Critério 3** - A expressão envolve um ou mais dos itens lexicais (ou formatos numéricos) descritos no critério 2, mas não cumpre o critério 1: *A Primavera é a mais bela estação do ano.*

3.3 Segmentação

As entidades temporais incluem a preposição, quando a ET é um grupo (ou sintagma) preposicional (SP: *no ano passado*), ou o determinante se a expressão é um grupo nominal (SN: *dois dias depois*). No caso das ET complexas, que podem eventualmente constituir sequências ambíguas, adoptam-se os critérios de segmentação definidos em (Hagège e Tannier, 2008):

Uma expressão temporal complexa deve ser dividida em unidades menores se e só se as seguintes condições forem ambas verdadeiras:

1. Cada componente da expressão é sintacticamente válida, quando combinada com o processo que modifica;
2. Cada componente da expressão é logicamente implicada na expressão complexa, ou, por outras palavras, se a expressão complexa for verdadeira, então cada expressão componente deve também ser verdadeira.

Por exemplo, na frase *Visitei o Pedro* ***dois dias nesta semana***, a expressão de tempo complexa deve ser dividida em duas ET, pois cada uma das expressões componentes se pode combinar com o evento: *Visitei o Pedro* ***dois dias*** */ Visitei o Pedro* ***nesta semana***, e cada expressão componente é tão verdadeira quanto o valor de verdade da expressão temporal complexa. Pelo contrário, na frase seguinte: *Visitei o Pedro* ***dois dias depois*** (= *dois dias mais tarde*), apenas uma ET deverá ser considerada, uma vez que, apesar de cada uma das expressões menores poder combinar-se sintacticamente com o evento: *Visitei o Pedro* ***dois dias*** */ Visitei o Pedro* ***depois***, o significado de cada combinação individual torna-se diferente do significado global da expressão complexa.

3.4 Classificação

A classificação é proposta juntamente com um conjunto de critérios. Esta classificação é inspirada em trabalhos anteriores (Saurí et al., 2006) mas também é influenciada pelo resultado da experiência de anotação temporal do Segundo HAREM (Baptista, Hagège e Mamede, 2008). Por último, está também intimamente relacionada com a classificação feita em (Ehrmann e Hagège, 2009).

O principal critério utilizado para classificar entidades temporais consiste no tipo de ancoragem (ou localização) dos processos temporais que operam. Quatro tipos principais são assim considerados:

1. DATA – a ET corresponde a uma ancoragem única do processo na linha do tempo;
2. DURAÇÃO – a ET não ancora o processo na linha do tempo, exprimindo porém uma quantificação de ordem temporal;
3. FREQUÊNCIA – a ET relaciona o processo com a linha do tempo através de várias instâncias de ancoragem;

4. GENÉRICO – a expressão não ancora qualquer processo na linha do tempo; não é realmente uma expressão temporal no sentido de que nenhuma informação temporal está associada a qualquer processo, mas mantém um significado temporal que pode ser importante para a resolução de referências temporais.

Enquanto que os três primeiros e principais tipos de expressões temporais podem constituir uma resposta adequada às interrogativas [2] com *(Prep) quando?*, *(Prep) quanto tempo?*, ou *com que frequência?*, respectivamente, o tipo genérico não pode.

A subclassificação destes tipos principais depende, sobretudo, da estrutura simples ou complexa da ET. Assim, o tipo DATA pode ser estruturado nos seguintes subtipos:

- DATAs simples, inclui não só as datas do calendário, mas também expressões temporais com horas (e.g. *20/05/2009 11:45 TMG*);
- INTERVALOs, expressões temporais envolvendo duas DATAs (*de 5 a 15 de Maio*); e
- o subtipo COMPLEXO, que corresponde a ET envolvendo expressões de DATA e de DURAÇÃO (*de hoje a quinze dias*).

Na mesma maneira, o tipo DURAÇÃO inclui um subtipo simples (por exemplo, *A reunião durará **2 horas***) e um subtipo de intervalo; o último envolve duas expressões quantificadas (*A reunião durará **entre 1 e 2 horas***).

Além disso, o tipo DATA também é classificado com base na referência temporal da ET e/ou na sua indeterminação quanto à ancoragem do processo na linha do tempo. Neste sentido, distinguem-se os seguinte subtipos:

- data ABSOLUTA, directamente computável a partir da ET (e.g. *em Maio de 2009*);
- data RELATIVA, envolvendo o cálculo de uma referência temporal; estas ET são ainda subdivididas, consoante se refiram ao momento de ENUNCIAÇÃO (e.g. *ontem*) ou a outro elemento TEXTUAL, algures no texto (e.g. *no dia seguinte*).

Uma propriedade especial, denominada *INDET*[3] em (Baptista, Mamede e Hagège, 2009) é usada em diferentes tipos de ET. Por exemplo, certas ET do tipo DATA fornecem, para o processo que modificam, um ponto de ancoragem na linha de tempo, no entanto, este ponto de ancoragem não é especificado. Assim, em: ***Numa bela manhã**, resolveu partir* o evento está ancorado no tempo, mas nada na expressão nem mesmo num contexto mais alargado permite indicar o ponto de ancoragem preciso. O mesmo tipo de indeterminação pode ser encontrado em ET dos tipos DURAÇÃO (*durante algum tempo*) e FREQUÊNCIA (*de tempos a tempos*).

Considera-se ainda outra propriedade, a que se chamou *FUZZY*, para as expressões temporais que, embora apresentem os elementos necessários para a sua normalização, se encontram modificadas por diferentes tipos de expressões que tornam *imprecisa* essa DATA(*por volta do dia 10*), DURAÇÃO (*durante cerca de 2 horas*) ou FREQUÊNCIA (*praticamente dois dias por semana*).

4 Desenvolvimento de um Sistema de Análise Temporal

Foi desenvolvido um sistema de reconhecimento de expressões temporais baseado nas directivas de identificação e de classificação acima apresentadas. Este módulo pretende ser o ponto de partida de uma cadeia de processamento das expressões temporais mais ambiciosa, isto é, que não se limite à mera identificação das ET mas que seja capaz de as classificar adequadamente, tendo como objectivo final a capacidade de ancorar temporalmente os processos expressos nos textos, assim como estabelecer relações de ordem temporal parciais entre estes mesmos processos.

Ora, como já o demonstrámos, processar a informação temporal desta maneira mais complexa obriga a ter em conta o contexto, por vezes bastante alargado, em que a ET se encontra: é necessário ter em consideração a natureza do evento, estado ou processo associado à ET; é necessário, também, que o sistema seja capaz de resolver anáforas; finalmente, é necessário ter em consideração os fenómenos de tempo e de aspecto verbal. Por estas razões, parece-nos que um sistema de reconhecimento de ET deve poder contar com informação linguística rica, a qual inclui desde a informação morfológica (para o processamento dos tempos e modos verbais) até à informação sobre cadeias anafóricas. Naturalmente, o sistema deverá poder contar com a ligação correcta entre ET e os processos modificados por estas ET. A nossa estratégia, tendo em conta estes requisitos, consistiu em integrar o processamento temporal num sistema mais geral

[2] A fim de capturar todos os tipos relevantes, outras formas interrogativas também são utilizadas, mas a sua lista completa é dada por Baptista e colegas em (Baptista, Mamede e Hagège, 2009).

[3] Este tipo de expressões é também chamado de *data indeterminada* em (Ehrmann e Hagège, 2009) e (Gosselin, 1996).

de análise linguística, considerando que o tratamento em paralelo da informação temporal e da informação sintáctica clássica deveria beneficiar tanto a análise sintáctica como o processamento temporal.

4.1 Apresentação do XIP

Uma característica importante do nosso módulo de anotação temporal é o facto de ele estar integrado numa ferramenta mais geral de processamento linguístico: O XIP-PT.

XIP (Xerox Incremental Parser) (Aït-Mokhtar, Chanod e Roux, 2002) é uma ferramenta de análise linguística cujo objectivo é a extracção de dependências sintácticas. A ferramenta oferece um formalismo de análise linguística que permite expressar um leque importante de regras, que vão da desambiguação das categorias das palavras até à construção de dependências, passando pela delimitação de sintagmas nucleares. Foram desenvolvidas gramáticas para diferente línguas no XIP. Para a gramática do Português, o sistema foi desenvolvido em conjunto no L2F, INESC-ID Lisboa e na Xerox. Este sistema é designado XIP-PT.

As várias etapas do processamento são as seguintes:

- uma fase de pré-processamento que inclui a segmentação, análise morfológica;
- a desambiguação das categorias de palavras;
- a análise sintáctica superficial;
- a análise sintáctica em dependências.

4.1.1 Pré-Processamento Linguístico

A etapa de pré-processamento inclui a segmentação e a análise morfológica das unidades textuais. O XIP-PT integra dois sistemas de pré-processamento desenvolvidos independentemente por cada uma das instituições que colaboram neste trabalho. A entrada do pré-processamento é um texto bruto ou XML. A saída do pré-processamento consiste num lista de unidades às quais é associada informação morfo-sintáctica (possivelmente ambígua). Nota-se que no XIP, uma entrada lexical é representada por um conjunto de traços (atributos:valores) que explicitam a informação linguística associada a esta entrada; trata-se, naturalmente, de informação morfos-sintáctica, mas também de natureza sintáctica e semântica.

4.1.2 Desambiguação

A desambiguação das categorias das palavras é feita de maneira híbrida: É utilizado um modelo escondido de Markov (HMM, *hidden Markov model*) em conjunto com regras construídas manualmente. Com efeito, o XIP oferece um formalismo de desambiguação que permite, considerando um contexto à esquerda e à direita de uma dada forma ambígua, escolher de entre um conjunto de categorias a categoria mais adequada ou preferencial.

4.1.3 Análise Sintáctica de Superfície

A análise sintáctica de superfície permite agrupar sequências de palavras, construindo sintagmas nucleares (sintagmas não recursivos no sentido dos *chunks* de Abney ((Abney, 1991)). Para o fazer, o XIP oferece um formalismo de regras de reescrita contextuais. É também graças a este formalismo que são elaboradas as regras do sistema de reconhecimento de entidades mencionadas (REM) integrado no XIP (Hagège, Baptista e Mamede, 2008a).

4.1.4 Análise Sintáctica em Dependências

A partir dos sintagmas nucleares delimitados na etapa anterior, e considerando a organização destes sintagmas e as propriedades lexico-sintácticas das unidades lexicais que os integram, é então possível estabelecer ligações (relações de dependência) entre os diversos elementos das frases. Estas ligações constituem relações orientadas, que são etiquetadas com o nome de uma função sintáctica. Assim, por exemplo, em Português, se um grupo nominal (NP) estiver à direita de um verbo e se um outro grupo nominal já tiver associado a esse mesmo verbo através da função sintáctica de sujeito, então, tipicamente o NP deverá desempenhar a função sintáctica de complemento directo; estas relações de dependência são construídas ligando a cabeça sintáctica dos sintagmas nucleares (chunks); no exemplo acima, a cabeça do NP estará ligada ao verbo com uma ligação de tipo complemento directo.

4.1.5 Ilustração

Para ilustrar as diferentes etapas de processamento, apresentamos a análise da sequência *Eis um exemplo que ilustra o funcionamento do XIP*.

Um excerto da saída do pré-processamento da frase inicial (apenas a cadeia *um exemplo que ilustra*) tem a seguinte forma (ver a figura 1): a sequência é inicialmente segmentada em entradas lexicais. O primeiro campo da saída corresponde à forma, o segundo ao lema da palavra e o terceiro à informação associada à palavra; os números correspondem ao *offset* da palavra; o resto da informação é apresentado sob a forma de traços booleanos. Note-se que as palavras *um*, *que* e *ilustra* são ambíguas, pelo que cada leitura corresponde a uma linha diferente.

```
um      um      +4+6+Pron+Indef+Masc+Sg+#lex+hmm_QUANTSG
um      um      +4+6+Art+Indef+Masc+Sg+#lex+hmm_QUANTSG
um      um      +4+6+Symbol+Meas+Abbr+#lex+hmm_SYM

exemplo exemplo +7+14+Noun+Masc+Sg+#lex+hmm_NSG

que     que     +15+18+Pron+Rel+MF+SP+#lex+hmm_PRONREL
que     que     +15+18+Pron+Interrog+MF+Sg+#lex+hmm_PRONSG
que     que     +15+18+Det+Interrog+MF+SP+#lex+hmm_DETINT
que     que     +15+18+Conj+#lex+hmm_CONJSUB

ilustra ilustrar +19+26+Verb+PresInd+3P+Sg+#lex+hmm_VERBF
ilustra ilustrar +19+26+Verb+Impv+2P+Sg+#lex+hmm_VERBF
```

Figura 1: A saída do pré-processamento para a sequência *um exemplo que ilustra*

À saída da fase de desambiguacão, a mesma sequência (ver a figura 2) já só apresenta para a palavra *um* uma única leitura, a de artigo indefinido.

Da mesma maneira, para a forma de entrada *que* já só subsiste a leitura de pronome relativo. No que diz respeito ao verbo, *ilustra*, na medida em que a ambiguidade se estabelece entre apenas duas formas verbais conjugadas (indicativo presente, terceira pessoa do singular e imperativo na segunda pessoa do singular) do mesmo lema *ilustrar*, não é ainda feita a sua desambiguação.

Na fase seguinte, o sistema procede a uma análise sintáctica de superfície que permite construir, a partir da frase inicial, uma sequência de *chunks*.

```
ADVP{Eis<ADV>}
NP{um<ART> exemplo<NOUN>}
SC{que<PRON>
   VF{ilustra<VERB>}
  }
NP{o<ART> funcionamento<NOUN>}
PP{de<PREP> o<ART> XIP<NOUN>}
```

Finalmente, são construídas, com base nesta primeira organização em *chunks*, uma série de relações de dependência entre os constituintes da frase:

```
MAIN(exemplo)
QUANTD(exemplo,um)
DETD(funcionamento,o)
DETD(XIP,o)
PREPD(XIP,do)
VDOMAIN(ilustra,ilustra)
MOD_POST(funcionamento,XIP)
MOD_POST(ilustra,XIP)
SUBJ(ilustra,que)
CDIR_POST(ilustra,funcionamento)
SUBORD(que,ilustra)
```

Assim, a relação `CDIR_POST` indica que o complemento directo de *ilustra* é *funcionamento*. A relação `DETD` liga a cabeça nominal *XIP* com o artigo *o*. Note-se que, nesta fase, não se desambigua a dependência do complemento preposicional *do XIP* (trata-se do conhecido problema do *PP-attachment*). Por esta razão, temos duas relações concorrentes de modificador envolvendo a palavra *XIP* e que exprime a ambiguidade de o complemento preposicional *do XIP* poder *a priori* encontrar-se ligado tanto a *funcionamento* como a *ilustra*.

4.2 Módulo XIP para a Anotação de Expressões Temporais

O desenvolvimento deste módulo foi iniciado em 2007 (Loureiro, 2007) e profundamente revisto para a campanha do HAREM em 2008 (Hagège, Baptista e Mamede, 2008b) na qual se propôs uma tarefa especial para anotação temporal.

Como se disse atrás, este módulo de processamento temporal está integrado num sistema mais geral de análise linguística, o XIP e executa as seguintes tarefas:

1. Reconhecimento e delimitação das ET nos textos;
2. Classificação destas ET;
3. Normalização (de um subconjunto) das ET;
4. Ligação entre as ET e os processos.

A realização destas tarefas é feita em paralelo às diferentes etapas de processamento do XIP descritas acima.

4.2.1 Pré-Processamento

Ao nível do pré-processamento, a implementação do módulo de análise temporal obriga à introdução de nova informação lexical. Com efeito, para o processamento temporal é necessário especificar mais a informação linguística de base associada a certos elementos lexicais (nome de

```
um       um       +4+6+Art+Indef+Masc+Sg+#lex+hmm_QUANTSG

exemplo exemplo +7+14+Noun+Masc+Sg+#lex+hmm_NSG

que      que      +15+18+Pron+Rel+MF+SP+#lex+hmm_PRONREL

ilustra ilustrar         +19+26+Verb+PresInd+3P+Sg+#lex+hmm_VERBF
ilustra ilustrar         +19+26+Verb+Impv+2P+Sg+#lex+hmm_VERBF
```

Figura 2: A saída da fase de desambiguacão para a sequência *um exemplo que ilustra*

meses, nome de dias), bem como a certas cadeias numéricas (números de 4 dígitos que possam corresponder a anos ou número de 1 a 2 dígitos potencialmente correspondentes à indicação de meses, etc.). Tecnicamente, esta especificação do léxico faz-se com introdução de novos traços, que serão depois utilizados nas gramáticas locais para reconhecimento de expressões temporais.

Por exemplo, à palavra *semana*, que, para o sistema geral de processamento do Português, é apenas considerada como um nome feminino, acrescenta-se o traço booleano `time_meas:+`, que indica tratar-se uma medida de tempo. De forma similar, ao lema nominal *primavera* acrescenta-se o traço *season:2*, que o especifica como um nome de estação do ano e o identifica com o número 2 (que será depois usado para cálculos).

4.2.2 Desambiguação de Categorias

O processamento temporal obrigou à introdução no sistema de novas regras de desambiguação. Por exemplo, o sistema de processamento do Português inicial considerava a palavra *Natal* como tendo apenas uma leitura, como nome próprio. É evidente, no entanto, que, num contexto de processamento do tempo, a distinção entre *Natal*, estado no Brasil, ou *Natal*, dia ou altura do ano, tem ser estabelecida. A regra seguinte determina que, quando a palavra *Natal* está precedida da preposição *durante*, a qual, por sua vez, pode ser seguida por um determinante e, eventualmente, por um adjectivo, esta palavra deverá ter apenas a leitura correspondente à expressão de tempo, passando, por esta razão a apresentar um traço específico `one_day` com valor `+`.

```
20> noun[maj:+,surface:Natal] %=
   |prep[lemma:durante],(art;?[dem:+]),(adj)|
   noun[one_day=+,maj=+,proper=+].
```

A primeira linha corresponde à parte esquerda da regra de desambiguação e significa que ela só será despoletada quando encontrar o nome *Natal*. A segunda linha corresponde ao padrão que deve seguir o contexto esquerdo da palavra para que a regra possa ser aplicada. Este contexto é uma expressão regular que descreve a seguinte sequência: a preposição *durante* seguida opcionalmente por um artigo ou um determinante demonstrativo, seguido ainda por um adjectivo opcional. Finalmente, a terceira linha indica os traços que devem ser acrescentadas à palavra *Natal* para que passe a ter apenas a leitura que corresponde à expressão temporal.

4.2.3 Gramáticas Locais

As gramáticas locais agrupam elementos lexicais, geralmente enriquecidos por nova informação relevante relativa ao tempo, para assim formar expressões temporais. Simultaneamente a este agrupamento, pode-se, em certos casos, proceder a uma primeira classificação de algumas expressões temporais. Por exemplo, no caso de datas completas (i.e., que incluem o número de dia, o nome de mês e o número de ano), como se trata de uma data absoluta não é necessário qualquer contexto para uma correcta classificação destas expressões. No mesmo sentido, o exemplo que se segue mostra a regra que permite construir uma ET a partir de o nome de uma estação do ano, seguido da preposição *de* e por uma sequência de dígitos correspondentes a um número que represente um ano, tal como *Primavera de 2002*.

```
18> noun[time=+,date=+,tipo_tempref=absolut]
    @=
    ?[season], prep[lemma:de],
    (?[lemma:o]), num[dig,year=+].
```

A parte esquerda da regra corresponde à expressão, constituída pelo elemento *Primavera*, que emparelha com o traço `?[season]` na parte direita; este elemento aparece então seguido pela preposição *de* e, por sua vez, por uma sequência numérica à que se vai acrescentar o traço `year:+`.

4.2.4 Dependências Sintácticas

As dependências vão permitir:

- Determinar a que predicado está ligado a ET; isto é feito graças à gramática geral

do Português, que calcula relações de modificação entre um predicado e um modificador;

- Caracterizar de forma mais pormenorizada certos tipos de ET que não podem ser classificados com um simples contexto local. Esta classificação pode ser feita graças às relações já calculadas, por exemplo entre o predicado e a ET-alvo que o modifica, mediante informação adicional obtida a partir desse predicado.

Considerem-se, por exemplo, as duas frases:

São **duas horas**
Ficou **duas horas** *em casa*

No primeiro caso, a expressão temporal constitui uma data relativa, com uma granularidade correspondente à unidade de medida hora. Trata-se, de facto, de uma expressão formular para indicar as horas, que apresenta alguma fixidez sintáctica. No segundo caso, estamos perante uma construção locativa, em que o sujeito está omisso, e que é facultativamente modificada por uma expressão de tempo do tipo DURAÇÃO e igualmente expressa em horas; nesta expressão, o verbo é tradicionalmente analisado como um verbo copulativo.

Para que um sistema automático seja capaz de fazer esta distinção, é necessário considerar o tipo de predicado expresso em cada frase e a forma como este se encontra associado à sequência *duas horas* (v.g. a construção formular de *ser*, no primeiro caso, e a construção locativa *ficar em casa*, no segundo). No léxico do sistema, está disponível a informação de que, entre outros traços, *ficar* pode ter um valor aspectual permansivo (anotado `permanency`). Com base nesta informação lexical, a regra seguinte determina que, perante um verbo copulativo com o valor permansivo, um complemento que possua o traço `time:+` deve ser classificado como uma duração.

```
// ficou 2 horas
if (^PREDSUBJ(#1[permanency],#2[time]))
        MOD[post=+](#1,#2),
        NE[tempo=+,duration=+](#2).
```

Na primeira linha, verifica-se a existência de uma relação `PREDSUBJ` entre um verbo copulativo (excluindo o verbo *ser*) e o complemento, que o verbo tem o traço `permanency`) e que o seu complemento constitui uma expressão de tempo, isto é, apresenta o traço `time`). As segundas e terceiras linhas correspondem às novas relações que são criadas se a condição expressa na primeira linha for verdadeira. Nesse caso, constrói-se uma expressão temporal `NE` do tipo DURAÇÃO e é estabelecida uma relação de modificação entre o verbo e a expressão temporal. Finalmente, é destruída a relação `PREDSUBJ` existente entre o verbo e a ET.

4.2.5 Cálculos Externos

Além destas tarefas que estão simplesmente integradas no analisador linguístico, há necessidade de realizar cálculos numéricos para proceder à normalização de expressões temporais dos tipos data absoluta, horas e durações. Assim, associam-se acções às regras para permitir realizar a normalização. Essas acções são chamadas a funções Python que podem ser executadas directamente a partir do analisador (Roux, 2006).

As expressões de subtipo DATA são normalizadas e o valor da normalização guardado no atributo `VAL_NORM` com o seguinte formato:

```
<Era><Ano><Mes><Dia>T<Hora><Minuto>
  E<ESTACAO>LM<limite\_aberto>
```

em que:

- `<Era>` corresponde a 2 caracteres: '+' ou '-', conforme a data seja depois ou antes da era de referência; e uma das seguintes letras maiúsculas, que representa a era de referência: C para a era cristã ocidental (valor por defeito), H para a era muçulmana (de Hijra, Hégira); M (anno Mundi) para o calendário judaico; P para a cronologia arqueológica (Presente = 1950); etc.
- `<Milenio>` corresponde a 2 caracteres de tipo dígito que representam o valor do milénio;
- `<Seculo>` corresponde a 2 caracteres de tipo dígito que representam o valor do século;
- `<Decada>` corresponde a 2 caracteres de tipo dígito que representam o valor da década;
- `<Ano>` corresponde a 4 dígitos que representam o valor do ano;
- `<Mes>` corresponde a 2 dígitos que representam o valor do mês;
- `<Dia>` corresponde a 2 dígitos que representam o valor do dia;
- `<Hora>` corresponde a 2 dígitos que representam o valor da hora;
- `<Minuto>` corresponde a 2 dígitos que representam o valor dos minutos;
- `<Segundo>` corresponde a 2 dígitos que representam o valor dos segundos;
- `<Milissegundo>` corresponde a 2 dígitos que representam o valor dos milissegundos;

- `<ESTACAO>` corresponde a 2 letras capitalizadas correspondente às estações do ano: `PR` para Primavera, `VE` para Verão, `OU` para Outono e `IN` para Inverno;
- `<limite_aberto>` indica se a expressão normalizada de data absoluta representa um intervalo de tempo com limite anterior ou limite posterior não determinado (em aberto). Os valores respectivos são: `A` no caso de limite anterior em aberto (neste caso a expressão temporal apresenta um limite posterior, e.g. *até 2009*); `P` no caso de limite posterior em aberto (neste caso, a expressão temporal tem um limite anterior, e.g. *desde 2009*); e, finalmente, `-`, quando a data absoluta corresponde a um intervalo sem limites abertos, e.g. *em 2009*.

No caso da data absoluta não ser expressa em termos de algum destes campos, o campo omiso é substituído por um ou mais "-". Por exemplo, a expressão *a 3 de Janeiro de 1986* é normalizada através de "`VAL_NORM=+19860103T----E--LM-`", a expressão *na Primavera de 1996* através de "`VAL_NORM=+1996----T----EPRLM-`"e a a expressão *antes das 3:00 da tarde* através de "`VAL_NORM="+--------T15--E--LMA`".

Para as expressões de tipo DURAÇÃO ou as DATAs relativas, o valor da normalização também é registada no atributo `VAL_DELTA` usando os seguintes campos:

```
A<digitos>D<digitos>H<digitos>
  M<digitos>S<digitos>M<digitos>
```

onde:

- as letras `A`, `D`, `H`, `M`, `S`, `M` são constantes que devem aparecer nesta ordem e que correspondem, respectivamente, ao valores de Anos, Dias, Horas, Minutos, Segundos e Milissegundos;
- os `<digitos>` à direita das letras correspondem ao valor dos campos respectivos; no caso das expressões de DURAÇÃO, este são simplesmente o valor temporal do intervalo de tempo; no caso das DATAs relativas, estes valores correspondem ao intervalo de tempo que se deve adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Usa-se a tabela 1 para converter as unidades de tempo durante a normalização de durações. Como princípio geral, os valores de VAL_DELTA devem ser convertidos para indicar de forma precisa a duração temporal referida. Para essa consideram-se três intervalos de valores:

- valores superiores a 1 ano;
- valores inferiores a 1 ano e superiores a 1 dia;
- valores inferiores a 1 dia;

Para efectuar a conversão usam-se as seguintes regras:

- todos valores de VAL_DELTA superiores ao ano são convertidos em anos;
- todos valores de VAL_DELTA inferiores a 1 ano e superiores a 1 dia são convertidos em dias;
- para valores de VAL_DELTA inferiores a 1 dia utilizam-se as unidades imediatamente inferiores (horas, minutos, segundos e milissegundos)

As expressões com valores inteiros inferiores a 1 dia não sofrem qualquer conversão: as expressões temporais são normalizadas pela transposição dos valores referidos nas correspondentes unidades temporais (horas, minutos, segundos e milissegundos).

No caso das expressões fraccionárias:

- fracções das unidades temporais são convertidas para a unidade imediatamente inferior;
- se a conversão não corresponder a um inteiro, arredonda-se para o inteiro mais próximo;
- para durações que combinam várias unidades temporais, a conversão faz-se para cada um das quantidades individuais.

Por exemplo, as seguintes expressões devem ser normalizadas como indicado:

- *durante um ano e dez dias* ("`VAL_DELTA=A1D10H0M0S0M0`")
- *durante meio ano* ("`VAL_DELTA=A0D183H0M0S0M0`")
- *2/3 da semana* ("`VAL_DELTA=A0D5H0M0S0M0`")
- *meio dia* ("`VAL_DELTA=A0D0H12M0S0M0`")
- *por hora e meia* ("`VAL_DELTA=A0D0H1M30S0M0`")

Das conversões apresentadas há uma que merece uma chamada de atenção por ser uma aproximação, já que aceitando que um ano tem 365 dias (ignorando os anos bissextos), um mês tem na realidade 30,41(6) dias e não 30 dias.

Por outro lado, as unidades de tempo que efectivamente ocorrem na ET são registadas noutro

Unidade 1	Unidade 2
1 milénio	1000 anos
1 século	100 anos
1 década	10 anos
1 ano	365 dias
1 mês	30 dias
1 quinzena	14 dias
1 semana	7 dias
1 dia	24 horas
1 hora	60 minutos
1 minuto	60 segundos

Tabela 1: Conversão entre unidades para o cálculo do VAL_DELTA.

campo, *UMED*. Por exemplo, a expressão temporal na frase *Fiquei* **dois meses** *em Lisboa* é normalizada através de "`VAL_NORM=A0D60H0M0S0MO UMED=meses`".

A tarefa de normalização é obtida através da análise dos pares atributo:valor associados aos elementos constituintes de cada entidade temporal normalizável. Para simplificar a tarefa de normalização, atribuem-se alguns traços específicos aos diferentes elementos que constituem a expressão temporal, nomeadamente aos dígitos que podem representar anos, meses, dias, horas, minutos e segundos, assim como as sequências alfabéticas para os meses e respectivas abreviaturas e os nomes das estações do ano. Assim, na expressão *25/Dez/2009*, o número *25* é associado à propriedade `day:+`, *Dez* à propriedade `month:+` e *2009* recebe a propriedade `year:+`.

A normalização "resume-se", então, a percorrer todos os nós das entidades temporais e a converter para o formato adequado todos os nós que contiverem um dos traços relevantes para a normalização. Contudo, é ainda necessário tratar de forma especial todas as ET que:

- contém elementos em numeração romana (*século XVI*);
- envolvem unidades de tempo não representadas na normalização final (*fim-de-semana*, *semana*, *quinzena* ou *semestre*);
- exprimem fracções de unidades de tempo(*meio ano*, *um mês e meio*);
- constituem maneiras informais de indicação das horas, como por exemplo *meia-noite*, *3 horas da tarde*, *2 menos um quarto* e *5 para as 3*;
- incluem expressões não numéricas referentes a durações (*uma quinzena de dias*) ou advérbios de tempo (*amanhã*, *anteontem* e *antes de ontem*);
- incluem diferentes tipos de modificador com valor referencial particular (*no dia seguinte*, *na semana que vem*, *no mês passado* e *no ano que há-de vir*).

4.3 Resultados

A avaliação do Módulo de Anotação Temporal teve lugar na campanha do Segundo HAREM (tarefa de anotação de expressões temporais). Sete sistemas participaram nesta tarefa, embora nem todos pretendessem tratar as ET ao mesmo nível de granularidade. Ainda assim, esta forte participação mostra o interesse da comunidade de processamento computacional do Português por este tema. Apenas um sistema se apresentou com o objectivo de realizar todas as dimensões da tarefa, (incluindo a de normalização).

Os resultados obtidos pelo sistema apresentado neste artigo são bastante animadores. Considerando-se as tarefas de identificação e de classificação de expressões temporais[4], o sistema atingiu uma precisão de 0,85 e uma abrangência de 0,76. Alguns erros ficaram a dever-se ao facto de, por enquanto, a tarefa de identificação e o processo de classificação terem sido feitos apenas ao nível local e, por essa razão, a semântica particular do processo associado às ET não ter podido ainda ser levado em linha de conta. Outros erros deveram-se a uma ainda incompleta codificação no léxico dos elementos que funcionam como índices (*triggers*) lexicais temporais. A normalização das datas absolutas e normalização parcial de datas referenciais também produziu resultados promissores, tendo o sistema conseguido um resultado de 0,74 de medida-f. No entanto, é opinião dos autores de que apenas a consideração do contexto mais alargado que envolve as expressões temporais poderá vir a melhorar de forma significativa estes resultados.

5 Conclusão

O processamento temporal é uma tarefa ambiciosa mas importante no domínio mais amplo de extracção de informação a partir de textos. Esta linha de pesquisa tem vindo a ser desenvolvida há já algum tempo e para diversos idiomas. Para Português, no entanto, a investigação neste domínio está ainda no seu início.

Em primeiro lugar, uma das dificuldades, consiste justamente em caracterizar de forma adequada o que se entende por expressões temporais,

[4] As directrizes para a classificação das ET propostas no Segundo HAREM são ligeiramente diferente das que aqui apresentámos, no entanto, elas são compatíveis.

tendo em conta as suas propriedades referenciais e sem perder de vista o objetivo principal da tarefa, isto é, a ordenação parcial dos processos, estados ou eventos expressos nos textos ao longo do eixo temporal. Têm sido desenvolvidas diferentes orientações e critérios para identificar, segmentar e classificar expressões temporais. Este artigo apresenta um conjunto de directrizes, inspiradas nas normas das campanhas de avaliação internacionais, com o objectivo de dar, assim, um passo firme mas significativo no sentido de um processamento temporal eficiente de textos em Português.

Foi igualmente desenvolvido um módulo temporal para reconhecer e classificar automaticamente as expressões temporais que aparecem nos textos de acordo com essas orientações. Nesta fase, o módulo temporal só opera ao nível sintáctico, mas os resultados obtidos são já bastante encorajadores. É, no entanto, bastante óbvio que o contexto imediato da frase não é suficiente para ancorar temporalmente os eventos, efectuar uma ordenação (parcial) temporal precisa dos eventos ou resolver todas as relações de referência temporal. De facto, um cálculo adequado da dimensão temporal veiculada nos textos tem de ter em conta muito mais elementos, como o tempo e o aspecto verbal, e diferentes processos de modelização do discurso, que alteram as condições de ancoragem temporal dos eventos. Também deverá ser necessário ter em consideração fenómenos que relevam da organização do discurso como, por exemplo, os referidos por Lascarides e Asher ((Lascarides e Asher, 1993)).

Acreditamos que, perante a quantidade e a diversidade de parâmetros que devem ser considerados para o tratamento da dimensão temporal e dado o facto de a anotação manual da informação temporal ser um trabalho extremamente difícil e custoso, uma abordagem baseada em regras e que explora informação linguística construída manualmente é a estratégia mais adaptada para esta tarefa. O nosso trabalho constitui apenas um primeiro passo nessa direcção. Esperamos poder avançar a pouco e pouco neste caminho, integrando progressivamente na nossa ferramenta de processamento do Português uma caracterização cada vez mais precisa dos diferentes tipos de eventos, estados e processos e desenvolvendo um módulo de cálculo referencial.

Referncias

Abney, S. P. 1991. Parsing by chunks. Em R. C. Berwick, S. P. Abney, e C. Tenny, editores, *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer, Dordrecht, pp. 257–278.

Aït-Mokhtar, Salah, Jean-Pierre Chanod, e Claude Roux. 2002. Robustness beyond shallowness: Incremental deep parsing. Em *Natural Language Engineering, 8*. Cambridge University Press, New York, NY, USA, pp. 121–144.

Allen, James F. 1991. Time and time again: The many ways to represent time. Wiley and Sons, pp. 341–355.

Baptista, Jorge, Caroline Hagège, e Nuno Mamede. 2008. Capítulo 2: Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo HAREM e o futuro. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Baptista, Jorge, Nuno Mamede, e Caroline Hagège, 2009. *Time Expressions in Portuguese. Guidelines for Indentification, Classification and Normalization (Internal Report L2F–INESC-ID)*, May, 2009.

Battistelli, Delphine, Jean-Luc Minel, e Sylviane Schwer. 2006. Représentation des expressions calendaires dans les textes: vers une application à la lecture assistée de biographies. *Traitement Automatique des Langues*, pp. 11–37.

Bittar, André. 2008. Annotation des informations temporelles dans des textes en français. Em *Actes de RECITAL 2008*.

Ehrmann, Maud e Caroline Hagège. 2009. Proposition de caractérisation et de typage des expressions temporelles en contexte. Em *Actes de TALN 2009*, Senlis, France.

Gosselin, Laurent. 1996. *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect*. Duculot.

Hagège, Caroline e Xavier Tannier. 2008. Xtm: A robust temporal text processor. Em *Proceedings of CICLing 2008*, Haïfa, Israël.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2008a. Capítulo 15: Reconhecimento de entidades mencionadas com o xip: Uma colaboração entre o INESC-L2F e a Xerox. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede, 2008b. *Proposta de anotação*

e normalização de expressões temporais da categoria TEMPO para o HAREM-MII. `http://www.linguateca.pt/aval_conjunta/HAREM/2008_04_13_Tempo.pdf`.

Lascarides, Alex e Nicholas Asher. 1993. Temporal interpretation, discourse relations, and commonsense entailment. Springer, http://www.springerlink.com, pp. 437–493.

Loureiro, João Miguel. 2007. Reconhecimento de Entidades Mencionadas (Obra, Valor, Relações de Parentesco e Tempo) e Normalização de Expressões Temporais. Tese de Mestrado, Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisboa, Portugal, November, 2007.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.* Linguateca, Aveiro, Portugal.

Parent, Gabriel, Michel Gagnon, e Philippe Muller. 2008. Annotation d'expressions temporelles et d'événements en français. Em *Actes de TALN 2008*, Avignon, France.

Prior, Arthur N. 1957. *Time and Modality.* Oxford University Press.

Reichenbach, Hans. 1947. *Elements of Symbolic Logic.* Reprinted, 1980, Dover Publications, New York.

Roux, Claude. 2006. Coupling a linguistic formalism and a script language. Em *Proceedings of CSLP-06 - Coling-ACL*, Sydney, Australia.

Saurí, Roser, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, e James Pustejovsky, 2006. *TimeML Annotation Guidelines Version 1.2.1*, January, 2006. `http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf`.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, e J. Pustejovsky. 2007. Xrce-t: Xip temporal module. Em *SemEval-2007 - Task 15 TempEval Temporal Relation Identification.*