DANIEL COUTO VALE

Interpretador Automático para Textos Escritos em Língua Informal

Projeto Orientado em Computação (POC) apresentado à Universidade Federal de Minas Gerais (UFMG) – Campus Pampulha, sob orientação da Doutora Mariza Andrades da Silva Bigonha, sob co-orientação do Doutor Marcello Peixoto Bax e sob consultoria da Doutora Adriana Silvina Pagano como parte das exigências da disciplina Projeto Orientado em Computação do curso de bacharelado em Ciência da Computação (7° período).

Belo Horizonte
2005

VALE, Daniel Couto.
Interpretador Automático para Textos Escritos em Língua Informal.
Daniel Couto Vale
Belo Horizonte: [s.n.], 2005.

Projeto Orientado em Computação
sob orientação da Doutora Mariza Andrades da Silva Bigonha,
sob co-orientação do Doutor Marcello Peixoto Bax.

_____

Mariza Andrades da Silva Bigonha
(orientadora)

_____

Doutor Marcello Peixoto Bax
(co-orientador)

À Doutora Adriana Silvina Pagano, por ter orientado meus primeiros trabalhos acadêmicos com extrema dedicação e confiança; ao Doutor Newton José Vieira, pelas leituras indicadas; ao Doutor Marcello Peixoto Bax, por ter me orientado com bastante liberdade na continuação do meu projeto; à Doutora Mariza Andrades da Silva Bigonha, por sua orientação e sua dedicação em corrigir prontamente as várias versões deste trabalho; à Tânia Liparini, por suas argumentações e sugestões de leitura; a todo o grupo NET, pelo suporte e acolhimento, eu dedico meus mais sinceros agradecimentos.

Abstract: Many reliable syntactic parsers have already been implemented for most natural languages. Despite this effort, Natural Language Processing softwares available in the market cannot yet extract the information transported by a text as carried out by readers. In the field of Artificial Intelligence, John F. Sowa (2000) developed a Knowledge Representation theory which describes the memory and the human language through Conceptual Graphs, a theory based upon Peirce's semiotic models. Drawing on syntactic parses and Sowa's theory, this paper discusses for taking into account concept definitions (SOWA, 2000), declarative and procedural knowledge and reception expectations in order to create algorithms that map syntax onto text interpretation.

Resumo: Atualmente, existem analisadores sintáticos automáticos confiáveis para várias línguas naturais; contudo, os aplicativos de Processamento de Língua Natural implementados até o presente momento não conseguem extrair a informação veiculada por um texto de modo similar a uma pessoa. No campo da Inteligência Artificial, John F. Sowa (2000) propôs uma teoria dedicada à Representação do Conhecimento que descreve a memória e a linguagem humana representando-as em Grafos Conceituais, uma teoria baseada nos modelos semióticos de Charles S. Peirce. Assim, com base nos analisadores sintáticos automáticos e na teoria de John F. Sowa, este trabalho demonstra a necessidade de considerar as definições dos conceitos (SOWA, 2000), os conhecimentos declarativos, procedimentais e as expectativas de recepção para criar algoritmos que realizem o mapeamento entre a sintaxe e a informação.

Sumário

1. Introdução 7
    1.1. Objetivos 8
2. Trabalhos Relacionados 9
    2.1. Grupo de Pesquisa em NLP da Microsoft 9
    2.2. Grupo de Pesquisa em NLP da Universidade de Nova Iorque 9
    2.3. Grupo de Pesquisa em NLP da Universidade de Sheffield 10
    2.4. Análise dos Projetos 10
3. Histórico 12
4. Metodologia de Desenvolvimento 13
5. Elementos Semânticos 15
    5.1. Tempo 15
    5.2. Aspecto 19
    5.3. Participantes 21
6. Implementação 29
    6.1. Camada 1: Analisador Atômico 30
    6.2. Camada 2: Analisador Léxico 30
    6.3. Camada 3: Analisador Sintático 31
    6.4. Camada 4: Analisador Semântico 31
    6.5. Banco de Conhecimento 32
7. Resultados 33
8. Discussão dos Resultados 35
9. Conclusão 36
10. Referências 37

1. Introdução

A Lingüística e a Ciência da Computação se desenvolveram no último século a ponto de ser possível a implementação de analisadores sintáticos automáticos eficientes para línguas informais[1]. Entretanto, o Formalismo, o Funcionalismo e o Cognitivismo no campo da Lingüística, a Inteligência Artificial na Ciência da Computação e a Semiótica não se dedicaram a formalizar a interpretação de textos informais de modo explícito e algorítmico[2], i. e. a semântica, enquanto ligação entre a sintaxe e a informação, das línguas informais ainda é pouco descrita pelas gramáticas. Devido a essa ausência de uma teoria que proponha um bom algoritmo de interpretação, os aplicativos de interpretação de texto implementados até o presente momento não conseguem extrair o conhecimento contido em um texto de modo similar a uma pessoa.

Descrever a semântica de uma língua qualquer é descrever a relação entre os signos e os conceitos de tal língua[3]. Para que isso seja possível, é necessário não somente interpretar um texto da língua, o que configuraria o uso da semântica, mas também explicitar como essa interpretação é feita. Segundo essa visão, precisamos formalizar, primeiro, os mecanismos que nos permitam ler um texto e extrair sua informação, para depois descrever a semântica.

Contudo, o principal problema que se encontra na descrição da semântica de uma língua qualquer é a necessidade de se usar uma representação para o conhecimento extraído de um texto (SOWA, 2000), pois o próprio texto é uma representação dos conceitos representados por ele mesmo. Portanto, a mais perfeita representação dos conceitos representados por um texto é o próprio texto[4].

---

[1] Textos Informais são textos escritos em Línguas Informais, que, por sua vez, formam o conjunto das línguas que não permitem a construção de um validador que confirme a corretude de um texto segundo regras morfossintáticas e semânticas preditivas. Todas as línguas naturais pertencem a essa categoria. Uma discussão mais profunda sobre as línguas informais se encontra no artigo Interpretação de Textos Informais (a ser publicado) do mesmo autor.

[2] Apesar de a semiótica se dedicar exatamente à formalização das relações entre signo, conceito e objeto, os principais teóricos semióticos, Ferdinand de Sausurre, Louis Hjemslev, Charles S. Peirce e Umberto Eco não sistematizam em algoritmos os mecanismos mentais que permitem associar o signo, o conceito e o objeto.

[3] Mesmo que a semântica descreva a ligação entre os signos e os conceitos, são válidas as contribuições da semiótica em que se propõe a tricotomia {Signo, Conceito, Objeto}: primeiro, sob a forma linear proposta por Platão sobre a arte {Imitação, Idéia, Objeto} (SOCRATES, República) em que se nega a ligação direta entre o signo e o conceito; segundo, sob a forma triangular proposta por Peirce {Signo, Interpretante, Objeto} (NETTO, 2003) em que existem ligações diretas entre os três pontos da tricotomia.

[4] Quando as línguas sob interpretação são línguas formais, podemos contar com uma precisão tamanha na morfossintaxe e na semântica que é possível traduzir um texto de uma língua formal para outra sem perda de conteúdo nem adições de informações. Já quando as línguas em questão são informais, não existe um único conhecimento passível de ser interpretado em cada texto e as traduções de uma língua para outra sempre implicam em perdas e adições intrínsecas à interpretação.

Apesar de não ser possível executar uma tradução perfeita, a descrição da semântica depende da existência de uma representação do conhecimento para que seja viável explicitar os mecanismos que mapeiam a sintaxe à informação. Portanto, neste trabalho é adotada a abordagem teórica de representação do conhecimento por Grafos Conceituais, proposta por John F. Sowa e baseada no modelo teórico de Charles S. Peirce (SOWA, 2000).

Com base nas teorias computacionais consolidadas em morfossintaxe, na representação de conhecimento de John F. Sowa e em modelos lingüísticos de interpretação da semiótica, este trabalho determina alguns dos elementos necessários para a criação de algoritmos dedicados a realizar o mapeamento entre a sintaxe e a informação[5]. Algoritmos que, neste trabalho, determinam a informação de textos informais e o representam em Grafos Conceituais.

Esta pesquisa promove aprimoramentos na interpretação automática de textos informais e, conseqüentemente, proporciona melhorias na interação entre usuários e computadores. Os resultados desta pesquisa também podem ser aplicados na classificação automática de textos informais, na alimentação de bancos de dados e nas buscas computadorizadas em *corpora* textuais de grande porte.

1.1. Objetivos

O trabalho descrito neste texto tem como objetivo geral explicitar alguns dos elementos necessários para criar algoritmos que realizem o mapeamento entre a sintaxe e a informação, implementando algoritmos que automatizam a interpretação para uma dada descrição da semântica.

Como objetivo específico, este trabalho se propõe a explicitar a necessidade de um banco de conhecimento que contenha a definição dos conceitos, conhecimentos declarativos, procedimentais e expectativas de recebimento para que seja possível criar algoritmos que realizem o mapeamento entre a sintaxe e a informação. Como ilustração dessa necessidade, foi proposto um problema semântico de difícil solução e implementado um algoritmo que, provido de um banco de conhecimento, consegue apresentar uma solução próxima à interpretação humana.

[5] Os esforços em direção à sistematização da semântica é uma fronteira comum a vários campos de pesquisa. Entre eles se encontram a Lingüística (com suas correntes formalistas, funcionais e cognitivas), a Semiótica, a Filosofia e a Inteligência Artificial. Nesses esforços, interessa à Lingüística, à Semiótica e à Filosofia a pura descrição da semântica, enquanto à Inteligência Artificial interessa a simulação da interpretação.

2. Trabalhos Relacionados

Nesta seção, a presente pesquisa será contrastada com os trabalhos de três centros estrangeiros de relevância internacional que promovem pesquisas na área de Processamento de Linguagem Natural (NLP): Microsoft, NYU e The University of Sheffied.

2.1. Projetos do grupo de pesquisa em NLP da Microsoft

Atualmente, seis projetos são desenvolvidos pela equipe de Processamento de Linguagem Natural da Microsoft: Machine Translation, Paraphrase, Japanese NLP, Amalgam, MindNet, IntelliShrink (MICROSOFT, 2005).

O projeto Machine Translation é o principal objetivo do grupo atualmente. Em contraposição aos sistemas tradução automática (MT systems) existentes no mercado, o grupo afirma ter obtido uma abordagem orientada por dados em que todo conhecimento de tradução é aprendido a partir de textos bilíngües existentes (*corpora* paralelos). O projeto Paraphrase oferece um imenso *corpus* etiquetado de paráfrases. As etiquetas determinam o tanto que os pares textuais são sentidos como veículos do mesmo conteúdo. O projeto Japanese NLP propõe uma "sintaxe neutra" teoricamente válida para todas as línguas. O projeto Amalgam visa à geração de textos fluentes em língua natural a partir de uma representação semântica. O projeto MindNet produz automaticamente redes semânticas a partir de textos e definições dos conceitos. Por último, o projeto IntelliShrink utiliza semântica e colocações para diminuir o tamanho de mensagens de texto enviadas por entidades computacionais móveis.

2.2. Projeto do grupo de pesquisa em NLP da Universidade de Nova Iorque

O projeto Natural Language Processing and The Representation of Clinical Data é o principal trabalho correlato desenvolvido recentemente pela equipe de Processamento de Linguagem Natural da NYU coordenado pelos seguintes membros: Phd Naomi Sager, Md Margaret Lyman, Md Christine Bucknall, Phd Ngo Nhan, Phd Leo J. Tick (SAGER, 2005).

O projeto visa à coleta automática de dados a partir de receitas médicas escritas em língua natural por meio de uma sintaxe exaustiva e um módulo semântico mínimo. A sintaxe

distingue os substantivos em várias classes segundo classificação semântica de modo a permitir a criação de regras sintáticas específicas para cada tipo de substantivo, elevando, assim, a taxa de acerto.

2.3. Projeto do grupo de pesquisa em NLP da Universidade de Sheffield

O projeto Information Extraction do grupo de pesquisa da Universidade de Sheffield (SHEFFIELD, 2005) produziu uma linha de programas de anotação de páginas virtuais. A anotação é feita por comparação léxica entre os átomos textuais e um catálogo de nomes importantes contidos nas ontologias e uma posterior análise sintática. A saída dos programas consiste em uma página de Internet anotada.

2.4. Análise dos Projetos

Entre todos os projetos analisados, o que mais se assemelha à proposta deste trabalho é o projeto MindNet. Em sua descrição, os pesquisadores do grupo de pesquisa em NLP da Microsoft afirmam utilizar algoritmos sintáticos e definições de conceitos para extrair redes de conhecimento de textos em língua natural.

O projeto Natural Language Porcessing and The Representation of Clinical Data, por outro lado, dedica-se com maior peso à análise sintática. A classe dos substantivos é dividida em dez subgrupos (semanticamente motivados) para possibilitar que todos os fenômenos sejam tratados pela sintaxe. A taxa de acerto atingida é bastante alta, mas o algoritmo demanda um enorme esforço computacional.

O projeto Information Extraction consiste no desenvolvimento de anotadores de página. Sua abordagem é bastante semelhante à adotada pelo grupo de pesquisa da NYU, com a distinção de que delegam exclusivamente à sintaxe a função de prover os dados necessários para o preenchimento de ontologias pré-produzidas.

A presente pesquisa se distingue das demais por sua ênfase na descrição semântica. A novidade da presente abordagem consiste em condicionar a interpretação não apenas à sintaxe (como fazem os grupos de pesquisa da Sheffield University e da NYU) ou às definições de conceitos (como faz o grupo de pesquisa da Microsoft no projeto MindNet), mas também aos

outros tipos de conhecimento declarativo, procedimentais e às expectativas de recepção. Com essa nova abordagem, torna-se evidente que vários problemas enfrentados por Naomi Sager (SAGER, 2005) advêm da tentativa de tratar todos os fenômenos lingüísticos pela sintaxe. Uma solução bem mais próxima da interpretação humana para problemas similares pode ser vista na Seção 5.3.

3. Histórico

O presente trabalho é resultado de um projeto que teve início em 2001 na Faculdade de Letras da UFMG em que Daniel Vale era bolsista de iniciação científica do CNPq sob orientação da Dra. Adriana Silvina Pagano. Durante esse período, o bolsista se dedicou ao estudo de padrões retóricos recorrentes em corpora paralelos de romances policiais. Em 2002, ainda como bolsista de iniciação científica do CNPq sob orientação da Dra. Adriana Pagano, dedicou-se à automatização da localização dos padrões retóricos encontrados no primeiro ano de pesquisa, o que resultou em um programa batizado de Intelligentī Pauca. Em 2003, sob consultoria do Dr. Newton José Vieira, o programa foi generalizado para receber gramáticas como entrada e foi renomeado Intelligentī Pauca v2. Em 2004, com outra bolsa de iniciação científica do CNPq sob orientação do Dr. Marcello Bax e consultoria da Dra. Adriana Silvina Pagano e do Dr. Newton José Vieira, foi desenvolvido o programa Intelligentī Pauca v3 que contém uma base de conhecimento declarativo, procedimentais e expectativas de recepção em grafos conceituais para auxílio na interpretação semântica de textos escritos em línguas informais.

4. Metodologia de Desenvolvimento

O programa Intelligentī Pauca, desenvolvido nesta pesquisa, possui mais de quinze mil linhas de código JAVA, distribuídas em três grandes módulos que serão descritos nas próximas seções. Como se trata de um projeto de tamanho considerável, foi necessária a aplicação de orientação por objetos, padrões de desenho e um processo pessoal de desenvolvimento de programas, apoiado no modelo de ciclo de vida Prototipagem Evolutiva.

Esse modelo de ciclo de vida foi escolhido para esta pesquisa por permitir construir programas em prazos curtos e adicionar novas características e recursos à medida que a experiência provar necessário. Nesse modelo, é desenvolvida uma série de versões provisórias denominadas *protótipos*, que cobrem um conjunto maior dos requisitos a cada nova iteração. O processo continua em ciclos até que o programa desejado seja concluído (FILHO, 2001).

Esse modelo de processo é vantajoso para a presente pesquisa por suprir a necessidade de definir os requisitos do programa à medida que novas hipóteses lingüísticas são formuladas. Ele permite uma grande flexibilidade de programação e produz ao fim de cada ciclo uma nova versão que pode ser submetida a testes, comprovando ou não a hipótese lingüística implementada.

Os problemas lingüísticos foram abordados por duas vias: de um lado, propunha-se uma descrição para um fenômeno lingüístico, de outro, testava-se a descrição ao simulá-la por meio de algoritmos computacionais. Aos poucos, as descrições foram se tornando mais complexas e, após um esforço de generalização, foi proposto um algoritmo que, para uma dada descrição gramatical externa ao programa, analisa textos informais e gera grafos conceituais que contêm a informação dos textos.

A adoção do modelo de processo Prototipagem Evolutiva permitiu a produção de uma série de gramáticas parciais[6], uma para cada protótipo do programa, nas quais se descreviam um número cada vez maior de fenômenos lingüísticos. Assim, a interpretação do programa Intelligentī Pauca pôde ser aproximada gradativamente à interpretação mais comum dos falantes.

As iterações são subdivididas em quatro passos. Primeiramente, define-se formalmente o fenômeno lingüístico a ser tratado; em seguida, cria-se ou altera-se uma gramática parcial de modo a abranger o novo fenômeno. O terceiro passo consiste em implementar a nova gramática. Por último, testa-se a nova gramática. Esta estratégia é, então,

[6] Gramática Parcial é o nome dado a uma gramática que visa a descrever um conjunto limitado de frases de uma língua informal.

repetida até o momento em que todos os fenômenos que se visa a tratar sejam abrangidos pela gramática parcial.

5. Elementos Semânticos

A gramática portuguesa parcial implementada pelo programa Intelligentī Pauca foi produzida para interpretar orações simples cujo verbo expressa uma ação ou mutação[7] sob os mesmos princípios que embasam a descrição dos papéis verbais proposta por Mário A. Perini em sua Gramática Descritiva do Português (PERINI, 1998).[8]

A representação do conhecimento usada nesta seção utiliza os Grafos Conceituais propostos por John F. Sowa (SOWA, 2005). Esses grafos conceituais são grafos que representam conhecimentos humanos e que podem ser processados por computador. O atual projeto visa a demonstrar que tal tradução é possível ao implementar um algoritmo que, em caráter experimental, traduz textos informais para grafos conceituais, seguindo uma gramática externa ao programa.

Esta seção reproduz um segmento do documento Cognição Artificial (do mesmo autor, a ser publicado) que declara e define conceitos importantes para a representação e processamento de conhecimentos por um computador. Os trechos selecionados tratam os tempos diretamente relacionados com o Agora e tratam os aspectos de um evento. Por último, é feita uma breve explicitação de um fenômeno que envolve a definição de participantes de um evento.

5.1 Tempo

Um evento ou um estado, por mais curtos que sejam, demoram um certo tempo para acontecerem. Qualquer proposição, quando entendida de um modo racional, gera uma imagem mental que é associada a uma duração, que pode ser medida em segundos, em nanossegundos ou em milênios de acordo com sua ordem de grandeza. Contudo, apesar de sabermos medir o tempo com precisão nunca antes vista pela humanidade, as línguas humanas ainda não vislumbram desta precisão científica em suas estruturas antigas como desinências verbais, conjunções, pronomes, advérbios, preposições e partículas. Normalmente o tempo preciso é expresso em fórmulas sintáticas genéricas por categorias abertas como substantivos, o que diminui sua prioridade no estudo lingüístico. Assim, para interpretar as frases e suas estruturas, basta sabermos que toda cena, todo objeto e toda relação existem por um recorte finito do tempo, i. e. têm uma duração.

Na matemática, uma duração é representada por uma diferença de tempo entre o fim e o começo representada pela letra 'Δ' antes de uma variável. Importarei a representação matemática em minha descrição e aplicarei todas as suas regras ao tempo lingüístico, a começar pela definição de *mora* como um recorte de tempo Δt, começado em t_1 e terminado em t_2. Sua definição possui cinco relações implícitas: *Igne, Fine, Ante, Post e Pons*. Igne e Fine representam o início e o fim da mora. Ante e Post significam respectivamente o antes e o depois da mora, enquanto Pons é a diferença de tempo entre o fim e o início. Sua representação em grafos conceituais pode ser vista abaixo:

---

[7] Verbos de Mutação são aqueles que indicam uma transição entre dois estados duradouros de um objeto. Por exemplo, uma batata *crua*, quando *assada*, se torna uma batata *assada*.

[8] Mais detalhes sobre o tema no sítio Cognição Artificial (VALE, 2005).

```
[Mora: λx] {
  [Δ: x]->(t1)->[Mora: Igne]<-(t2)<-[Mora: Ante]
        ->(t2)->[Mora: Fine]<-(t1)<-[Mora: Post]
        ->(t2minust1)->[Grandeza: Pons]
}
```

De todas as moras que podem ser referidas no discurso humano, duas se destacam pela freqüência em que são usadas em todas as línguas: a mora *agora* e a mora *princípio*. A primeira se refere à duração da comunicação entre o enunciador e o enunciatário, a qual se estabelece por meio de metalinguagem retórica. A segunda se refere ao começo dos tempos, ao marco antes do qual nada se pode falar ou imaginar. No atual estado da ciência, este marco zero seria o Big Bang (: Big Bang : grande explosão), enquanto para várias culturas como a greco-latina, a judaico-cristã, a islâmica e as nativas do Brasil o marco é menos preciso e referido genericamente como αρχη (archã : arqué : princípio), principium (:: princípio), beginning (:: princípio). Como se trata de duas moras específicas, as mesmas não constituem um tipo, mas sim exemplares do tipo Mora:

```
[Mora: #agora]
[Mora: #princípio]
```

Outras moras podem ser tidas úteis na classificação dos tempos, entre as quais podemos destacar as que determinam a grandeza da duração. Uma mora de tamanho cuja duração desprezamos leva o nome de *Instante*, enquanto uma de tamanho cuja duração é muito maior que o tempo em questão leva o nome de *Eternidade*. Uma mora cuja duração seja menor que o esperado leva o nome de *Rapidez*, enquanto outra cuja duração seja maior que o esperado leva o nome de *Demora*. Assim, pode-se definir os seguintes conceitos:

```
[Instante: λx] {
  [Mora: x]->(Pons)->[Grandeza: #infinitésimo]
}
[Eternidade: λx] {
  [Mora: x]->(Pons)->[Grandeza: #infinito]
}
[Rapidez: λx] {
  [Mora: x]->(Pons)->[Grandeza: #pequeno]
}
[Demora: λx] {
  [Mora: x]->(Pons)->[Grandeza: #grande]
}
```

Mesmo que não tenhamos definido como interpretar infinitésimos, infinitos, pequenos e grandes, é presumido que esses conceitos sejam claros e que se possa derivar deles, por exemplo, que nenhuma mora é maior que a eternidade e que nenhuma mora é menor que um instante. Faz-se aqui uma delegação de tratamento de um item por um mecanismo externo aos grafos conceituais, e esta delegação é expressa pelo caractere '#'. Com os conceitos definidos acima, podemos seguir à descrição dos tempos metalingüísticos: passado, passado distante, passado recente, futuro, futuro distante, futuro eminente. Veja que não se pretende falar sobre o *Presente* nesta seção, pois o mesmo é por essência um aspecto somado a um tempo. (BRANDÃO, 2001)

Passado

A marcação de passado sem outras informações é vista no praesens perfectum indicativum do latim, no pretérito perfeito do português, no past simple do inglês, no passé composé do francês e nas formas . e . do japonês. Define-se passado como uma mora seguida pelo momento atual. Justifica-se a utilização de três moras na definição como pode ser visto abaixo devido ao fato de as relações *post* e *ante* indicarem uma justaposição em que o fim de uma mora coincide com o início da outra enquanto precisamos indicar somente uma seqüência:

- Citō sperrectus sum.
- Acordei cedo.
- I woke up early.
- Je me suis réveillé tôt.
- .. ..

```
[Passado: λx] {
  [Mora: x]->(Post)->[Mora]->(Post)->[Mora: #agora]
}
```

Passado Remoto

A marcação de passado remoto é vista em português e em inglês, normalmente em uma construção cristalizada referente ao tempo. Em nenhuma das línguas com as quais trabalhamos, a informação de passado remoto é transmitida pelas desinências verbais. Normalmente, o que se encontra são estruturas antigas cristalizadas bastante semelhantes às estruturas construtivas.

- Assisti esse filme a muito tempo atrás.
- Faz muito tempo que assisti esse filme.
- I watched this movie a long time ago.
- It has been a long time since I watched this movie.

```
[PassadoRemoto: λx] {
  [Mora: λa]->(Post)->[Mora]->(Post)->[Mora: #agora]
                              ->(Pons)->[Grandeza: #grande]
}
```

Passado Médio

Nenhuma das sete línguas estudadas indica um passado médio com estruturas antigas. Sua representação em grafos conceituais é relevante, pois alguns estudos lingüísticos atestam que certas línguas nativas do Brasil e Swahili, cuja memória escrita é uma conquista recente, têm conjugações verbais para tal tempo:

```
[PassadoMédio: λx] {
  [Mora: x]->(Post)->[Mora]->(Post)->[Mora: #agora]
                            ->(Pons)->[Grandeza: #médio]
}
```

Passado Recente

Todas as línguas modernas desta pesquisa têm um advérbio interior ao sintagma verbal ou um verbo auxiliar em uso comum que marcam o passado recente. Em português, temos o verbo auxiliar ACABAR, em inglês, o advérbio JUST e, em japonês, o auxiliar .... Adiciona-se ainda que o advérbio AGORA em português também remete ao significado de passado recente:

- Acabei de chegar em casa.
- I just came home.
- .. ..

```
[PassadoRecente: λx] {
  [Mora: x]->(Post)->[Mora]->(Post)->[Mora: #agora]
                            ->(Pons)->[Grandeza: #pequeno]
}
```

Futuro

Enquanto o passado se associa com memórias, o futuro só existe na esfera da antecipação. Quando alguém diz algo sobre suas ações futuras, por exemplo, o enunciatário certamente entende o enunciado como uma expressão de intenções, uma agenda. Como oposição a esta interpretação, a cultura judaico-cristã renega a antecipação, afirmando que o futuro esteja determinado antes de ocorrer. Portanto, devemos ser cautelosos quanto a afirmar que todos interpretem o futuro como uma esfera essencialmente estocástica em que pouco se pode afirmar sem o risco de ser provado errado. Neste conflito entre possibilidades e pré-determinismo, a experiência prática diária nos leva a crer em uma relação direta entre intenções e a realização de nossas vontades, enquanto o misticismo alega conseguir predizer sem erro um futuro predeterminado.

Assim, mesmo que fosse possível identificar um marcador estritamente ligado ao futuro, acabaríamos por encontrar uma estrutura que alguns interpretariam como uma arrepsia, outros como uma antecipação — abordagem adotada neste trabalho — e terceiros como uma certeza. Mas há, sobretudo, uma certeza, a de que o futuro, tanto o possível quanto o predeterminado, só pode ser criado a partir de uma classificação de cenas baseada no que aprendemos como regras de causa e conseqüência. Esse julgamento pode, por exemplo, se basear na credibilidade e obstinação do enunciador. Alguém que sempre descumpre os compromissos terá pouca credibilidade quando disser: *Desta vez eu vou comparecer*. Ao mesmo tempo, outra pessoa bastante determinada dará um tom de factualidade com a frase *Vou discorrer sobre Platão na minha próxima palestra*. Algumas línguas, como o grego clássico, não possuem nenhuma marcação de futuro nos verbos. Em grego, a marcação de futuro se dá através de substantivos e de advérbios. Existem formas verbais chamadas de 'futuro', entretanto seu significado é *desiderativo* e *volitivo*, raramente factual (BRANDÃO, 2001).

Por outro lado, até mesmo o passado pode ser duvidoso em caso de esquecimento, mentira ou ignorância. O esquecimento é nítido em pessoas que sofrem de amnésia e uma frase como *Fui para João Pessoa ontem com a Fernanda* [Paciente com Amnésia] soa extremamente improvável quando dita por alguém com amnésia que se encontra em Belo Horizonte. Já um mentiroso compulsivo dirá *Ultrapassei uns oitenta carros em menos de um quilômetro* [Jovem na Savassi] e a veracidade de tal afirmativa também será descartada pela maior parte dos ouvintes. Já uma pessoa extremamente mística terá pouca credibilidade ao dizer *Fui gerada de novo dentro da barriga de Deus, e aí então eu subi! E fui parar na cabeça dEle* [Baby do Brasil em entrevista com Jô Soares]. Porém, o passado é duvidoso em casos específicos enquanto o futuro é, para muitos, inerentemente duvidoso.

Por esses motivos, os marcadores de futuro são construções instáveis nas línguas e por pouco tempo se consolidam em desinência verbal. Sendo que, uma vez formada a desinência verbal, esta logo se desfaz e é substituída por advérbios, verbos auxiliares e/ou outros artifícios semânticos que dão uma melhor precisão temporal. No português, existe a construção com o auxiliar IR que traz uma idéia de futuro. No inglês, o antigo verbo WILL, que ainda é usado em algumas construções como *May he do what he wills* e *I'm willing to give many things up for you*, atualmente funciona como um auxiliar de futuro. Em francês e em espanhol, usa-se em paralelo ao verbo conjugado formas perifrásticas com o auxiliar ALLER e IR respectivamente. Em japonês, existem várias formas perifrásticas, todas indicando o futuro com outros traços semânticos e pragmáticos agregados.

- Prandium parābō.
- Vou preparar o almoço.
- I will prepare lunch.
- Je préparerai le déjeuner.
- Voy a preparar el almuerzo.

```
[Futuro: λx] {
  [Mora: x]->(Ante)->[Mora]->(Ante)->[Mora: #agora]
}
```

Futuro Distante

Se o futuro, por si só, é uma esfera duvidosa, o futuro distante sofre de uma indefinição ainda maior para a maioria das pessoas. Seu significado se assemelha a uma convicção e se afasta de uma proposição. O futuro distante só é considerado factual em épocas bem calmas, na monotonia. Normalmente, a antecipação certa de um futuro distante é associada, na literatura e em outras expressões culturais, à impressão de que a vida já passou ou à sensação de estagnação, impotência. Em português, existe uma perífrase verbal que marca este tempo. Confira abaixo:

- Ele ainda vai acabar te notando.

```
[FuturoDistante: λx] {
  [Mora: x]->(Ante)->[Mora]->(Ante)->[Mora: #agora]
                           ->(Pons)->[Grandeza: #grande]
}
```

Futuro Médio

O futuro médio é bastante usado nas várias línguas modernas na terminologia cunhada pela teoria da administração com a expressão *a médio prazo*, em oposição a *a longo prazo*, *a curto prazo* e *de imediato*. Existe, no entanto, uma conjugação perifrástica formada pela junção do advérbio AINDA com o auxiliar IR que marca este tempo:

- Ainda vou quitar as dívidas que eu tenho no banco.

```
[FuturoMédio: λx] {
    [Mora: x]->(Ante)->[Mora]->(Ante)->[Mora: #agora]
                             ->(Pons)->[Grandeza: #médio]
}
```

Futuro Eminente

O futuro eminente é o mais certo dos futuros e várias construções que o marcam podem ser conferidas em todas as línguas com as quais trabalho. Sua marcação mais freqüente em português se dá pela conjugação perifrástica formada pela junção do advérbio JÁ com o auxiliar IR:

- Já vou te atender, espera só um pouquinho.

```
[FuturoEminente: λx] {
    [Mora: x]->(Ante)->[Mora]->(Ante)->[Mora: #agora]
                             ->(Pons)->[Grandeza: #pequeno]
}
```

5.2 Aspecto

O aspecto é um conhecimento mais complexo do que o tempo, porém é um traço semântico bem mais intuitivo, pois pode ser derivado de um raciocínio pragmático de causa e conseqüência. Isso se deve ao fato de que nós humanos, ao tentar sistematizar o mundo em narrativas, inevitavelmente elegemos uma seqüência de fatos que julgamos ser *uma cadência natural de causas e conseqüências*. (POE, 2001) É nesse sentido que devemos entender os recortes da realidade, como um esforço humano de ordenar em uma narrativa um mundo caótico e inconstante.

Como está definido, o aspecto é uma afirmação sobre o estágio de completude de um evento, o qual pode ir desde as precondições até os resultados. Perceba a relação de causa e conseqüência entre *precondição*, *evento* e *resultado*. Como trabalharemos com desinências verbais, será necessária a definição do aspecto vazio, o qual não carrega informação alguma sobre o estágio de completude dos eventos. Este é comumente conhecido como *aoristo*, um aspecto sem marcação. Juntamente, definimos um *evento* ou *estado* como aspectos marcados. Veja as seguintes definições:

```
[Aoristo: λx] {
    [T: x]->(has)->[Mora]
}
[Evento: λx] {
    [T: x]->(has)->[Mora]->(Pons)->[Grandeza: #pequeno]
}
[Estado: λx] {
    [T: x]->(has)->[Mora]->(Pons)->[Grandeza: #grande]
}
```

Existe uma impressão de mundo bastante comum devido à nossa experiência diária que nos induz a acreditar que na maior parte do tempo a disposição dos objetos no mundo permanece a mesma e que em alguns poucos instantes essa disposição é alterada. Como conseqüência disso, a grande maioria dos verbos das nossas línguas descrevem seqüências de um estado, um evento e um outro estado, no qual o evento é tido como uma transição entre dois estados. Essa seqüência instintiva é tão freqüentemente referida, que seus três elementos recebem nomes especiais: *precondição*, *transição* e *resultado*. Confiram os grafos conceituais que os definem:

```
(Seq: λx λy) {
    [T: x]->(has)->[Mora]->(Post)->[Mora]<-(has)<-[T: y]
}

[Precondição: λx] {
    [Estado: x]->(Seq)->[Evento]
}
[Resultado: λy] {
    [Estado: y]<-(Seq)<-[Evento]
}
[Transição: λx] {
    [Evento: x]<-(Seq)<-[Precondição]
               ->(Seq)->[Resultado]
}
```

O grego, o português, o espanhol e o japonês têm construções que precisam o aspecto resultativo. No entanto, somente na primeira língua o resultativo consiste em uma conjugação verbal generalizada para todos os verbos. Tanto em português e espanhol quanto em japonês, verbos ativos, quando expressos em estruturas resultativas, se tornam verbos médios, i. e. que não possuem referência a um agente. Mesmo assim, é possível um paralelo entre um verbo do português com as conjugações gregas que ressaltam bastante o traço semântico aspecto.

1.	A porta fechou.	O Pedro fechou a porta.
2.	A porta está fechada.	*Graças ao Pedro* a porta está fechada.
3.	A porta abriu.	O Pedro abriu a porta.
4.	A porta está aberta.	*Graças ao Pedro* a porta está aberta.

A cadência natural dos fatos para uma porta poderia ser entendida como uma seqüência de eventos e estados descrita por um laço eterno das proposições 1, 2, 3 e 4. É instintivo acreditar (isto pode ser contestado) que o mundo é inerte a menos que uma "ação" seja feita. Quando se diz *a porta fechou*, nossa mente tende a formular uma imagem de que algo causou o fechamento da porta, talvez o vento, talvez um motor... Não nos passa em mente a possibilidade do movimento inértico, pois a experiência de vida nos demonstra que inevitavelmente o atrito freia os objetos. O mundo é para nós essencialmente estático a menos de "ações". As proposições 2 e 4 são entendidas portanto como estados enquanto as proposições 1 e 3 são entendidas como eventos.

Avançando um pouco mais o raciocínio, vemos que as proposições 1 e 3 são entendidas como transições, pois têm um estado natural anterior e um estado natural posterior. Podemos, então, perceber que *estar fechado* é o resultado de *fechar* e que *estar aberto* é o resultado de *abrir*. Eu, Daniel Vale, acho extremamente curioso o fato de as línguas grego, português, espanhol e japonês usarem o mesmo radical verbal para designar a transição e o resultado como em **fech**ar/**fech**ado. O

mesmo não ocorre com a precondição. Talvez isso seja um indício de que acreditamos ser capazes de obter um mesmo resultado em diversas circunstâncias. Reitero que esta é uma opinião minha.

Por fim, defino o aspecto durativo como um aoristo que perdura durante todo um evento ou estado, i. e. que começa antes e termina depois de outro aoristo. O durativo é também referido nas conjugações como *contínuo*. Em inglês, denominam-se as estruturas perifrásticas formadas pela justaposição do auxiliar *be* com um verbo na forma *ing* pela alcunha de past continuous, present continuous e future continuous. Em português, podemos identificar uma marcação deste tempo no presente durativo e o pretérito imperfeito analítico.

- I was just kidding, when I said that.
- I'm singing in the rain, just singing in the rain.
- I'll be traveling the whole day.

- Estou estudando.
- Eu estava estudando quando aconteceu o acidente de ontem.

```
(Durante: λx λy){
  [T: x]->(has)->[Mora]->(Igne)->[Mora]->(Post)->[Mora]<-(Fine)<-[Mora]<-(has)<-[T: y]
  [T: x]->(has)->[Mora]->(Fine)->[Mora]->(Ante)->[Mora]<-(Inge)<-[Mora]<-(has)<-[T: y]
}

[Duração: λx]{
    [Aoristo: x]<-(durante)<-[Aoristo]
}
[PresenteDurativo: λx] {
    [Aoristo: x]<-(durante)<-[Aoristo]->(Mora)->[Mora: #agora]
}
[PassadoDurativo: λx] {
    [Aoristo: x]<-(durante)<-[Aoristo]->(Mora)->[Passado]
}
```

5.3 Participantes

Para demonstrar como é abordada a questão dos participantes, é ilustrada a criação passo a passo de uma gramática que começa pela seleção de um conjunto de frases que se pretende descrever com tal gramática. As frases são escolhidas e ordenadas de modo que a complexidade da gramática cresça gradativamente para cada conjunto de frases a serem interpretadas.

Lista de Frases

1.1. O menino picou a batata.
1.2. Meu irmão bateu a porta.
1.3. Alguém quebrou o vidro.
2.1. A batata assou.
2.2. O vidro quebrou.
2.3. A porta bateu.
3.1. O carro amassou a lataria.
3.2. Miguel cortou a mão.
3.3. Papai sujou o sapato.
4.1. Joana cortou seu próprio cabelo.
4.2. Algumas mulheres depilam as próprias pernas.
4.3. Joana pinçou a sobrancelha sozinha.
5.1. Janaína cortou o cabelo.
5.2. Janaína depilou as pernas.
5.3. Janaína pinçou a sobrancelha.
5.4. Janaína cortou as unhas.

A criação da gramática é dividida em cinco passos. Um passo consiste na descrição completa de uma gramática *naíva* parcial e provisória que contemple um subconjunto das frases listadas acima. A cada passo, esse subconjunto de frases é expandido de modo a contemplar todas as frases ao fim do quinto passo.

1)

As três primeiras frases da lista podem ser descritas por uma abordagem que estabelece primeiramente regras sintáticas e posteriormente regras semânticas. As regras sintáticas descrevem a seqüência SVO (sujeito, verbo, objeto direto), considerada a mais típica da língua portuguesa por várias gramáticas; e, posteriormente, as regras semânticas determinam que se devem interpretar os sujeitos como agentes e os objetos diretos como pacientes. As árvores sintáticas do primeiro conjunto de frases podem ser vistas na Figura 1.

Sintaxe:

```
<sub> ::= "menino" | "irmão" | "batata" | "porta" | "vidro".
<det> ::= "a" | "meu" | "o".
<sn>  ::= "alguém" | <det> " " <sub>.
<ver> ::= "bateu" | "picou" | "quebrou".
<SVO> ::= <sn> " " <verbo> " " <sn>.
```

Semântica:

1. O verbo é[9] uma ação
2. O sujeito é o agente da ação.
3. O objeto direto é o paciente da ação.

[9] O verbo 'é' é um substituto, neste contexto, para verbos maiores como 'significa' e 'simboliza'.

Primeiro Conjunto de Frases

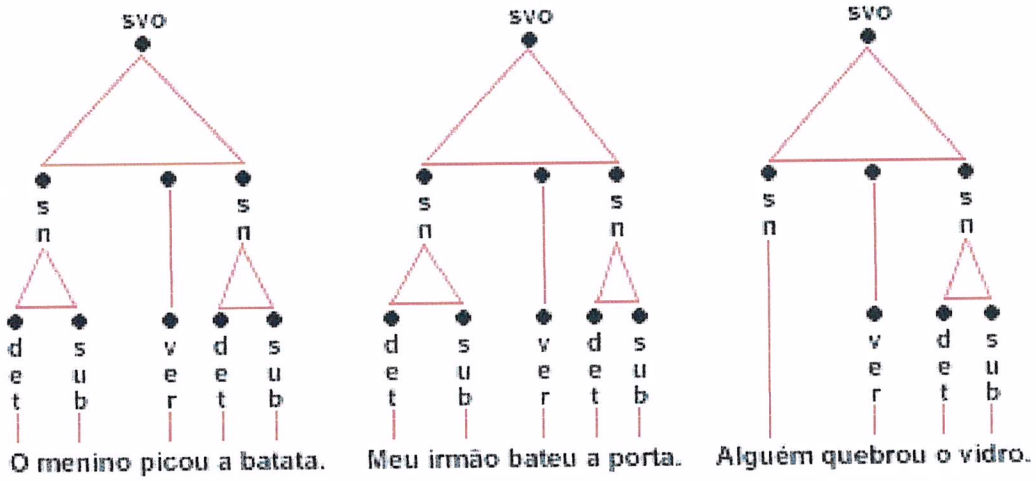

Figura 1

2)

Acrescentando mais três frases ao conjunto a ser tratado, podemos continuar com uma abordagem que estabelece que as regras sintáticas sejam tratadas antes das regras semânticas. Ao conjunto de regras sintáticas proposto pela primeira gramática parcial, é necessário adicionar a seqüência SV (sujeito, verbo). Portanto, as regras semânticas devem ser reformuladas para descrever os dois tipos de frase. As árvores sintáticas do segundo conjunto de frases podem ser vistas na Figura 2.

Sintaxe:

```
<sub> ::= "menino" | "irmão" | "batata" | "porta" | "vidro".
<det> ::= "a" | "meu" | "o".
<sn>  ::= "alguém" | <det> " " <sub>.
<ver> ::= "assou" | "bateu" | "picou" | "quebrou".
<SVO> ::= <sn> " " <ver> " " <sn>.
<SV>  ::= <sn> " " <ver>.
```

Semântica:

Se a estrutura for SVO

1) O verbo é uma ação.
2) O sujeito é o agente da ação.
3) O objeto é o paciente da ação.

Se a estrutura for SV

1) O verbo é um acontecimento.
2) O sujeito é o paciente do acontecimento.

Segundo Conjunto de Frases

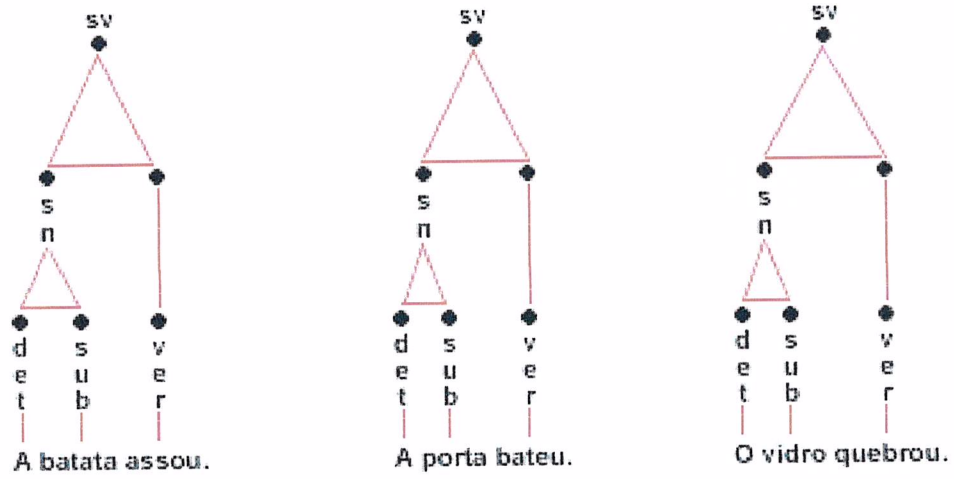

Figura 2

3)

Ao acrescentar as próximas três frases, não é mais possível determinar a análise semântica apenas a partir das estruturas sintáticas. As orações em questão possuem uma estrutura SVO, mas significam 'acontecimento' como se fossem orações SV, enquanto, segundo a presente descrição, elas deveriam ter sentido de 'ação'. É improvável que a frase "O carro amassou a lataria." seja entendida por alguém como se o carro fosse o agente da ação de amassar.

Portanto, faz-se necessário encontrar um elemento distintor (qualquer que este seja) para separar as orações SVO em dois grupos de modo que um contenha todas as orações com sentido de ação e o outro, todas as orações com sentido de acontecimento. Um possível distintor seria o fato de que todos os objetos diretos dessas orações[10] se referem a partes das entidades referidas pelos sujeitos. Esse distintor é baseado em semântica declarativa e prescinde do conhecimento prévio dos conceitos e entidades referidos. Mas, apesar de conhecimento declarativo apresentar dificuldades práticas computacionais quanto à conceitualização e à catalogação, como esse é o distintor mais genérico que se pode encontrar para a implementação da terceira gramática parcial, sua escolha é justificável. As árvores sintáticas do terceiro conjunto de frases podem ser vistas na Figura 3.

Sintaxe:

```
<sub> ::= "menino" | "irmão" | "batata" | "porta" | "vidro" | "lataria" |
          "mão" | "sapato".
<det> ::= "a" | "meu" | "o".
<sn>  ::= "alguém" | "Miguel" | "Papai" | <det> " " <sub>.
<ver> ::= "assou" | "bateu" | "picou" | "quebrou".
<SVO> ::= <sn> " " <ver> " " <sn>.
<SV>  ::= <sn> " " <ver>.
```

Semântica:

Se a estrutura for SV

1) O verbo é um acontecimento.
2) O sujeito é o paciente do acontecimento.

Se a estrutura for SVO

Se o objeto direto ***for parte*** do sujeito

1) O verbo é um acontecimento.
2) O objeto direto é o paciente do acontecimento.

[10] Orações SVO com sentido de acontecimento

3) O sujeito é o todo do paciente.

Se o objeto direto ***não for parte*** do sujeito.

1) O verbo é uma ação.
2) O sujeito é o agente da ação.
3) O objeto direto é o paciente da ação.

Terceiro Conjunto de Frases

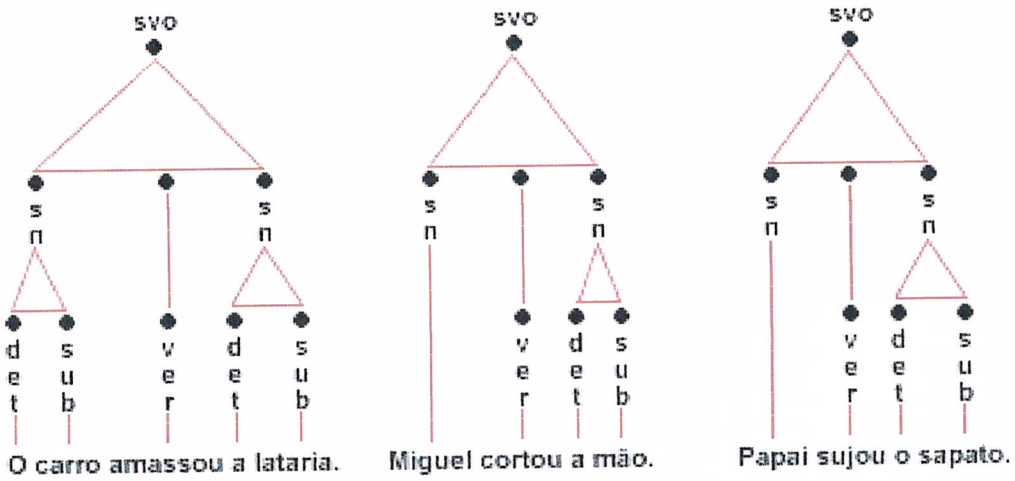

Figura 3

4)

Acrescentando mais três frases ao conjunto a ser tratado pela próxima gramática provisória, a composição dos sintagmas nominais terá de ser reformulada de modo a abarcar todos os casos. A semântica também terá de ser reformulada de modo que a sintaxe passa a determinar em paralelo com o conhecimento declarativo se uma oração se refere a um acontecimento ou a uma ação. O grande impulsionador da reformulação da semântica proposta até aqui é o fato de que, apesar de apresentarem estrutura SVO e semanticamente suas entidades referidas se relacionarem em parte e todo, as novas frases significam 'ação', ao passo que a gramática provisória proposta prevê o significado de acontecimento para essas frases.

As palavras "próprio", "próprias e "sozinha" são os novos distintores e podem ser consideradas tanto como elementos sintáticos sistematizados que determinam a semântica, quanto como elementos sintáticos de classe aberta catalogados pelo conhecimento declarativo. Uma abordagem mais genérica as classificaria como elementos desuniformemente sistematizados pelo código lingüísticos porque, provavelmente, falantes com diferentes graus de domínio da língua passarão por mecanismos distintos na compreensão. A gramática provisória que se segue, utiliza esses elementos em um passo posterior à sintaxe e à semântica, apesar de seu devido lugar ser entre a pura sintaxe (primeiro distintor) e a pura semântica (segundo distintor). As árvores sintáticas do quarto conjunto de frases podem ser vistas na Figura 4.

Sintaxe:

```
<sub> ::= "menino" | "irmão" | "batata" | "porta" | "vidro" | "lataria" |
          "mão" | "sapato" | "cabelo" | "mulheres" | "pernas" |
          "sobrancelha".
<det> ::= "a" | "algumas" | "as" | "meu" | "o" | "seu".
<pro> ::= "próprio" | "próprias".
<sn>  ::= "alguém" | "Miguel" | "Papai" | "Joana" | <det> " " <sub> |
          <det> " " <pro> " " <sub>.
<ver> ::= "assou" | "bateu" | "picou" | "quebrou" | "depilam" | "pinçou".
<SVO> ::= <sn> " " <ver> " " <sn> |
          <sn> " " <ver> " " <sn> "sozinha".
<SV>  ::= <sn> " " <ver>.
```

Semântica:

Se a estrutura for SV

1) O verbo é um acontecimento.
2) O sujeito é o paciente do acontecimento.

Se a estrutura for $SVO_{sozinha}$

1) O verbo é uma ação.
2) O objeto direto é o paciente da ação.
3) O sujeito é o todo do paciente e o único agente da ação.

Se a estrutura for SVO

Se o objeto direto ***não for parte*** do sujeito.

1) O verbo é uma ação.
2) O sujeito é o agente da ação.
3) O objeto direto é o paciente da ação.

Se o objeto direto ***for parte*** do sujeito

Se o determinante do objeto direto for "seu próprio" ou "as próprias"

1) O verbo é uma ação.
2) O objeto direto é o paciente da ação.
3) O sujeito é o todo do paciente e o agente da ação.

Se o determinante do objeto direto não for "seu próprio" ou "as próprias"

1) O verbo é um acontecimento.
2) O objeto direto é o paciente do acontecimento.
3) O sujeito é o todo do paciente.

Quarto Conjunto de Frases

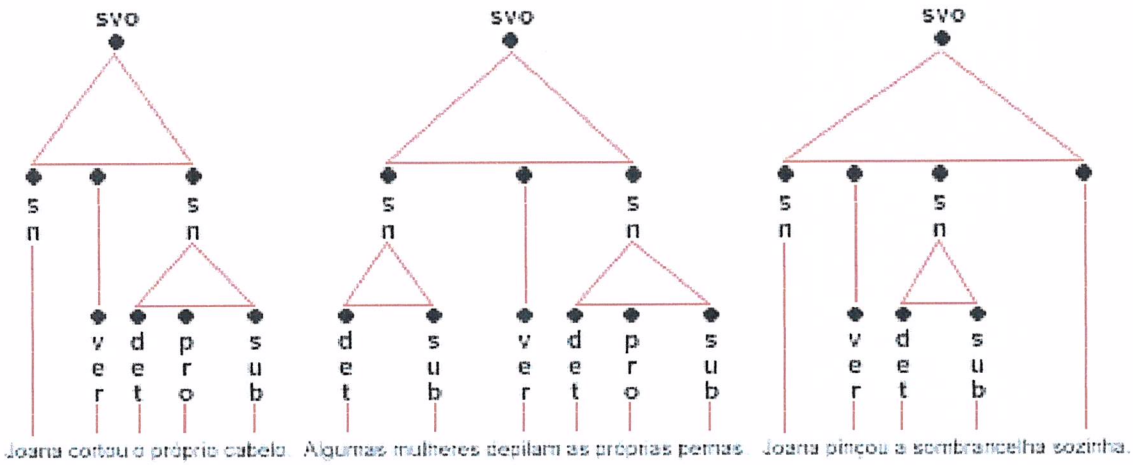

Figura 4

5)

Acrescentando as últimas quatro frases da lista ao conjunto a ser tratado pela última gramática provisória, observa-se que a sintaxe não precisa ser reformulada, pois as novas frases seguem a estrutura SVO padrão. Contudo, a semântica de todas as quatro novas frases apresenta um desafio de tratamento computacional extremamente mais complexo que o visto até o presente momento.

A quarta gramática provisória interpreta as frases como acontecimentos ocorridos com partes do corpo de Janaína. Esta interpretação é bastante menos refinada que o entendimento humano, pois, como sabemos que as frases se referem a procedimentos estéticos, os espertos na

língua, que tenham algum conhecimento de estética feminina brasileira, entenderão que tais acontecimentos são 'ações'.

Há ainda uma outra característica dessas frases que aparentemente dificultam mais ainda a extração de informações por computadores e esta jaz na atribuição de quem seria o agente dos procedimentos estéticos. Levantemos algumas hipóteses: 1) se soubermos que Janaína é o nome de uma jovem de classe média alta, a interpretação mais esperada para estas frases é a de que a jovem tenha ido a um centro de estética para se submeter aos procedimentos. Nesse caso, a informação das frases é de que outras pessoas omitidas sejam agentes da ação. 2) Outra interpretação possível seria a de que a própria Janaína tenha executado os procedimentos estéticos sobre seu próprio corpo. Isso seria esperado caso conhecêssemos Janaína e soubéssemos seus costumes, ou caso as mulheres com quem convivamos tenham o costume de se produzirem sozinhas. 3) Todo esse cenário se subverteria se soubéssemos que Janaína é uma *coifeuse* esteticista e que ela estava em seu salão. Nesse caso, a interpretação esperada seria a de que ela seja o agente dos procedimentos que, por sua vez, seriam aplicados sobre outra pessoa.

Assim, a semântica proposta pela quarta gramática provisória deve ser alterada de modo a utilizar as expectativas e incertezas do receptor como parâmetros para decidir sobre que interpretação deve ser produzida. Estas novas adições à semântica entram no campo da incerteza, uma interface bastante estreita entre a pura semântica e a pura pragmática. Esses distintores não são mais baseados apenas no conhecimento declarativo dos conceitos e entidades, mas se suportam em expectativas e incertezas provindas de contextos e experiência com esses contextos. As árvores sintáticas do quinto conjunto de frases podem ser vistas na Figura 5.

Sintaxe:

```
<sub> ::= "menino" | "irmão" | "batata" | "porta" | "vidro" | "lataria" |
          "mão" | "sapato" | "cabelo" | "mulheres" | "pernas" |
          "sobrancelha" | "unhas".
<det> ::= "a" | "algumas" | "as" | "meu" | "o" | "seu".
<pro> ::= "próprio" | "próprias".
<sn>  ::= "alguém" | "Miguel" | "Papai" | "Joana" | "Janaína" |
          <det> " " <sub> | <det> " " <pro> " " <sub>.
<ver> ::= "assou" | "bateu" | "picou" | "quebrou" | "depilam" |
          "pinçou" | "cortou" | "depilou".
<SVO> ::= <sn> " " <ver> " " <sn> |
          <sn> " " <ver> " " <sn> "sozinha".
<SV>  ::= <sn> " " <ver>.
```

Semântica:

Se a estrutura for SV

1) O verbo é um acontecimento.
2) O sujeito é o paciente do acontecimento.

Se a estrutura for $SVO_{sozinha}$

1) O verbo é uma ação.
2) O objeto direto é o paciente da ação.
3) O sujeito é o todo do paciente e o único agente da ação.

Se a estrutura for SVO

Se o objeto direto ***não for parte*** do sujeito.

1) O verbo é uma ação.
2) O sujeito é o agente da ação.
3) O objeto direto é o paciente da ação.

Se o objeto direto ***for parte*** do sujeito

Se o determinante do objeto direto for "seu próprio" ou "as próprias"

1) O verbo é uma ação.
2) O objeto direto é o paciente da ação.
3) O sujeito é o todo do paciente e o agente da ação.

Se o determinante do objeto direto não for "seu próprio" ou "as próprias"

Caso **não haja expectativas**

1) O verbo é um acontecimento.
2) O objeto direto é o paciente do acontecimento.
3) O sujeito é o todo do paciente.

Caso o sujeito **seja esperado** delegar a ação

1) O verbo é uma ação.
2) O objeto direto é o paciente.
3) O sujeito é o todo do paciente
4) O agente existe e é omitido.

Caso o sujeito **seja esperado** executar a ação sobre si

1) O verbo é uma ação.
2) O objeto direto é o paciente.
3) O sujeito é o todo do paciente e o agente da ação.

Caso o sujeito **seja esperado** executar a ação sobre outrem

1) O verbo é uma ação.
2) O objeto direto é o paciente.
3) O sujeito é o agente da ação.
4) O todo do paciente existe e é omitido. **Metonímia**

Quinto Conjunto de Frases

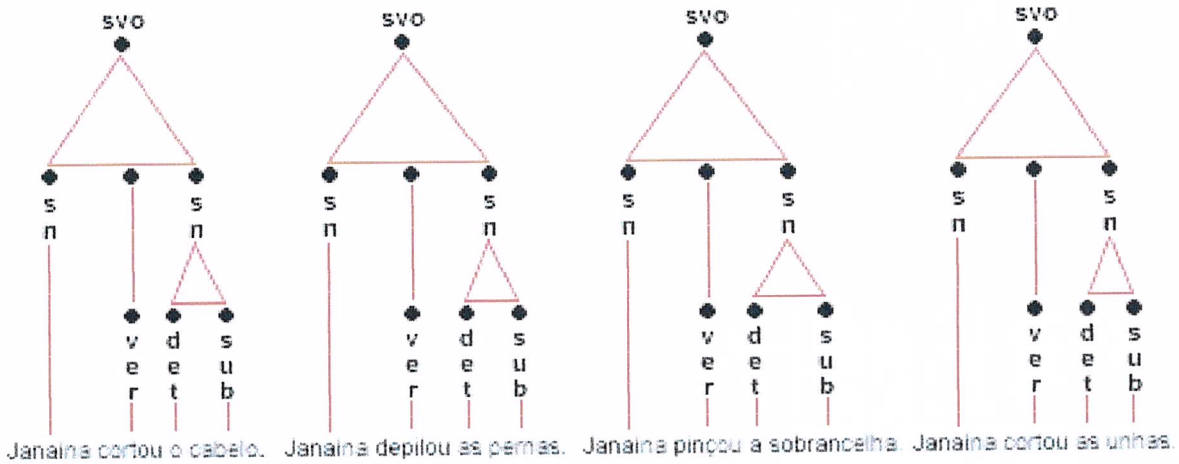

Figura 5

6. Implementação

O programa Intelligentī Pauca implementa um interpretador de textos informais na linguagem de programação Java. Seus parâmetros são uma gramática de uma língua qualquer, um banco de conhecimentos e um texto informal. Sua saída são grafos conceituais, uma representação similar a bancos de entidade e relacionamento, que descrevem a informação do texto informal.

A gramática é implementada em quatro camadas e todas elas podem entrar em contato com o banco de conhecimentos. As quatro camadas são Athomizon (:: analisador atômico), Legon (:: analisador léxico), Tasson (:: analisador sintático), Semainon (:: analisador semântico). A Figura 6 exemplifica a passagem dos elementos pelas várias camadas da gramática. Nas próximas subseções é descrita brevemente a implementação de cada uma das camadas.

Como o programa tem uma implementação bastante extensa (15.000 linhas de código), não é possível abordar todos os detalhes neste documento. Uma documentação mais detalhada pode ser encontrada no sítio virtual Documentação: Intelligenti Pauca v.3 (VALE, 2005), cujo endereço será mantido o mesmo até o fim do ano 2005.

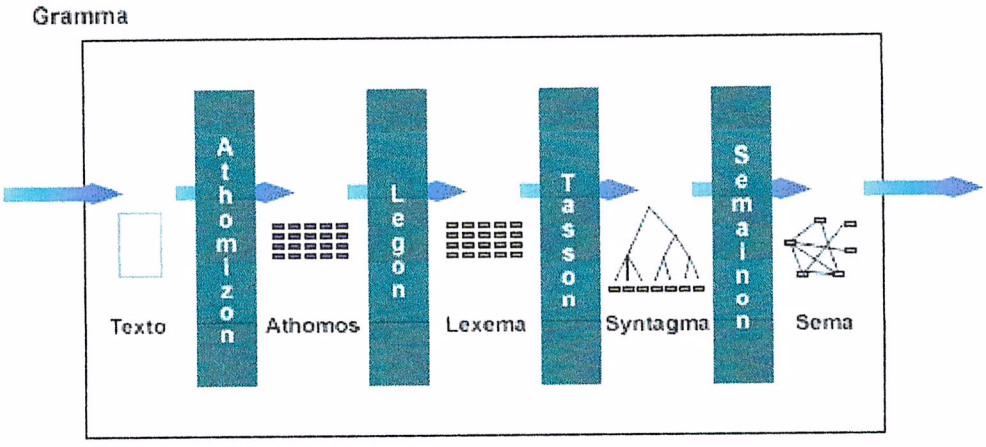

Figura 6

Recebido este catálogo de fixos, o programa Intelligentī Pauca monta um autômato finito para identificar rapidamente a que lexema um átomo deve ser associado. Para o caso em existam mais de um lexema possível para um mesmo átomo, foi implementado um mecanismo que permite propagar a ambigüidade para as próximas camadas.

6.3. Camada 3: Analisador Sintático

Uma instância do analisador sintático recebe uma lista de lexemas e produz uma árvore sintática. Como parâmetro de construção da instância do analisador sintático, é recebida uma lista de regras gramaticais que descrevem como o programa deve montar a árvore. O algoritmo implementado contorna problemas provocados por gramáticas ambíguas ao aplicar todas as regras a todos os átomos e construções sem aceitar repetição de uma mesma regra em um mesmo escopo de texto. O algoritmo também faz restrições quanto à seqüência da aplicação das regras, que é mantida linear. Ao fim de sua execução, são produzidas várias árvores sintáticas de alturas e escopos distintos.

Foi decidido que o analisador sintático não propagaria as ambigüidades para o analisador semântico devido ao alto custo da computação da camada semântica. Por isso, foi implementado um algoritmo de caminho mínimo que escolhe as melhores árvores, considerando a altura das árvores como peso e o número de árvores entre o início e o fim do texto como a distância. Assim, o modulo de análise sintática escolhe o menor número de árvores baixas que descrevem a sintaxe entre o início e o fim do texto.

6.4. Camada 4: Analisador Semântico

Uma instância do analisador semântico recebe uma seqüência de árvores sintáticas e produz um grafo conceitual que representa o entendimento daquilo que foi dito no texto. Esse analisador recebe uma descrição do procedimento de análise semântica adotado pela língua aos moldes da descrição da Seção 5. Com o auxílio do banco de conhecimento com seus

procedimentos de análise semântica produzidos para português abrangem os campos semânticos descritos neste documento.

6.5. Banco de Conhecimento

Foi implementado um banco de conhecimento que lê arquivos no formato KB (KnowledgeBase), desenvolvido durante esta pesquisa. O programa Intelligentī Pauca utiliza um único arquivo do tipo KB chamado memoria.kb. Esse arquivo contém um grafo conceitual que descreve um banco de conhecimento com todo o seu conteúdo: conceitos, relações, instâncias, definições entre outros. É provido ao programa uma classe KnowledgeBase cuja única instância carrega o arquivo memoria.kb e constrói um grafo em objetos Java. Esses objetos são então percorridos pela única instância da classe KnowledgeBase para que a mesma possa responder perguntas às várias camadas de análise implementadas pelas gramáticas. A classe KnowledgeBase implementa uma fachada de manipulação dos grafos de modo a não permitir que seu conteúdo seja indevidamente alterado.

7. Resultados

Durante a pesquisa, foi produzida uma gramática parcial que interpreta frases simples do português de modo a extrair informações sobre tempo, aspecto e os papeis dos participantes. A gramática não possui mecanismos para tratamento de termos desconhecidos a incluir nomes próprios e vocabulário não incluído. O vocabulário abrangido consiste em aproximadamente 200 palavras dos seguintes campos semânticos: pessoas (dona de casa, padeiro, chefe de cozinha), partes do corpo (braço, mão, rosto), alimentos (batata, arroz, feijão) e ações encontradas em livros de culinária (assar, bater, flambar), todas relacionadas ao contexto 'cozinha'.

Abaixo são listados os conhecimentos extraídos pelo programa Intelligentī Pauca de quatro conjuntos de três frases. Eles demonstram, respectivamente, a interpretação dos tempos, dos aspectos, dos participantes e das vozes verbais pelo programa. Os testes aqui apresentados não são, de nenhum modo, exaustivos, mas ilustram o funcionamento da presente abordagem.

Conjunto 1

- A dona de casa picou a batata.
- A dona de casa acabou de picar a batata.
- A dona de casa ainda vai picar a batata.

```
[Picar]-
   <-(Mudador)<-[Dono_de_Casa]
   <-(Mudado)<-[Batata]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

```
[Picar]-
   <-(Mudador)<-[Dono_de_Casa]
   <-(Mudado)<-[Batata]
   <-(Tempo)<-[PassadoRecente]
   <-(Aspecto)<-[Aoristo]
```

```
[Picar]-
   <-(Mudador)<-[Dono_de_Casa]
   <-(Mudado)<-[Batata]
   <-(Tempo)<-[FuturoMédio]
   <-(Aspecto)<-[Aoristo]
```

Conjunto 2

- O menino fechou a porta.
- A porta está fechada.
- A porta está fechada graças ao menino.

```
[Fechar]-
   <-(Mudador)<-[Menino]
   <-(Mudado)<-[Porta]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

```
[Fechar]-
   <-(Mudado)<-[Porta]
   <-(Tempo)<-[Presente]
   <-(Aspecto)<-[Resultativo]
```

```
[Fechar]-
   <-(Mudador)<-[Menino]
   <-(Mudado)<-[Porta]
   <-(Tempo)<-[Presente]
   <-(Aspecto)<-[Resultativo]
```

Conjunto 3

- A ajudante de cozinha cortou a mão.
- A ajudante de cozinha está com a mão cortada.
- Cortou a mão.

```
[Cortar]-
   <-(Mudado)<-[
    [Mão: λ]
      <-(Parte)
        <-[Ajudante_de_Cozinha]
   ]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

```
[Cortar]-
   <-(Mudado)<-[
    [Mão: λ]
      <-(Parte)<-
        [Ajudante_de_Cozinha]
   ]
   <-(Tempo)<-[Presente]
   <-(Aspecto)<-[Resultativo]
```

```
[Cortar]-
   <-(Mudado)<-[
    [Mão: λ]
      <-(Parte)
        <-[T: #anaphora4]
   ]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

Conjunto 4

- A batata assou.
- A batata foi assada.
- O chefe de cozinha deu uma assada nas batatas.

```
[Assar]-
   <-(Mudado)<-[Batata]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

```
[Assar]-
   <-(Mudador)
      <-[Pessoa: #anaphora]
   <-(Mudado)<-[Batata]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

```
[Assar]-
   <-(Mudador)
      <-[Chefe_de_Cozinha]
   <-(Mudado)<-[Batata]
   <-(Tempo)<-[Passado]
   <-(Aspecto)<-[Aoristo]
```

O Conjunto 1 demonstra que a alternância entre as perífrases e conjugações verbais provoca alterações na interpretação do tempo pelo programa. O Conjunto 2 demonstra que a escolha das construções perifrásticas determina a interpretação do aspecto. O Conjunto 3 demonstra que, quando o objeto direto é parte do sujeito, o programa não interpreta o sujeito como agente e entende o evento como um acontecimento, em acordo com a descrição da Quinta Gramática Parcial da Seção 5.3. O Conjunto 4 demonstra que as estruturas verbais passivas são interpretadas pelo programa de modo a recuperar o agente.

8. Discussão dos Resultados

O programa Intelligentī Pauca demonstra que é possível utilizar definições de conceitos, conhecimentos declarativos, procedimentais e expectativas de recepção para extrair a informação de textos informais de modo semelhante ao que fazem os leitores. A semântica descrita neste trabalho se suporta não apenas na sintaxe, como também em dados referentes às entidades envolvidas e ao contexto; assim, consegue simular os mecanismos humanos de interpretação.

O tratamento de diferentes estruturas sintáticas como voz ativa e voz passiva não é empecilho para a generalidade da análise semântica, porque o módulo semântico pode arquivar, no banco de conhecimento, os procedimentos que lhe permitem interpretar construções sintáticas. Na gramática parcial portuguesa implementada para o programa Intelligentī Pauca, por exemplo, a marcação temporal é um fenômeno tratado juntamente por descrição semântica e por conhecimentos procedimentais arquivados no banco de conhecimento. Essa abordagem que combina programação e catalogação dos procedimentos de interpretação se demonstrou bastante eficiente para as estruturas mais antigas da língua como as desinências e perífrases verbais.

A complexidade do algoritmo não apresentou problemas práticos para um vocabulário reduzido de aproximadamente 200 palavras e 250 conceitos. Para cada texto de 10 frases e 80 palavras, o programa Intelligentī Pauca despende em média um segundo no computador Pégaso[11]. Contudo, não é possível calcular sua ordem de complexidade, porque a camada semântica permite que seus módulos façam chamadas recursivas. Além disso, o custo computacional é bastante afetado pelos tipos de frase recebidos. Em resumo, a complexidade varia muito de acordo com a gramática parcial implementada e com o texto informal interpretado, assim impedindo que sejam feitos cálculos da ordem de complexidade.

[11] Computador Pégaso:
PC Windows XP
Processador – 2.3GHz
RAM – 512MB

9. Conclusão

A implementação do programa foi bem sucedida e o mesmo consegue identificar os agentes e pacientes das ações e acontecimentos e consegue determinar o tempo e o aspecto dos verbos ou locuções verbais. Segundo todos os testes executados, o programa apresenta respostas condizentes com a interpretação humana para textos que usam um vocabulário português de pouco mais de 200 palavras, todas relacionadas ao contexto 'cozinha'.

Assim, foi possível demonstrar que a implementação de um algoritmo que se disponha a simular a interpretação humana requer que se desenvolva um bom mecanismo de manipulação de conhecimentos declarativos, procedimentais e expectativas de recepção e um banco de conhecimentos bastante vasto, mesmo que o objeto de estudo seja de tamanho moderado.

Foi também possível demonstrar que a sintaxe, a semântica e a pragmática podem ser devidamente tratadas pela computação, se permitirmos o uso dos conhecimentos declarativos, procedimentais e pragmáticos por parte das quatro camadas do interpretador.

Por fim, foi possível determinar o tamanho do programa (15.000 linhas) na linguagem de programação Java para um escopo semântico extremamente limitado, o que nos permite dizer que esta abordagem não é viável de ser implementada para toda a língua manualmente. Para que se possa implementar um algoritmo desse porte, seria necessário automatizar a construção de várias partes do programa, como a construção das regras sintáticas, das regras semânticas e a alimentação do banco de conhecimento.

10. Referências

BRANDÃO, Jacyntho Lins; SARAIVA, Maria Olívia de Quadros. **ΕΛΛΗΝΙΚΑ**: uma introdução ao grego antigo. Vol. 1-3. Sine Loco: Sine Nomine, 2001.

DESCARTES, René. **Discours de la méthode**. Disponível em:
<http://www.gutenberg.org>
Acesso em: 23 jun 2005.

FILHO, Wilson de Pádua Paula. **Engenharia de Software**. 2. ed. Rio de Janeiro: LTC Editora, 2003. ISBN 85-216-1339-3

MICROSOFT. Microsoft NLP Research Group. Disponíevel em:
<http://research.microsoft.com/nlp/>
Acesso em: 23 jun. 2005.

NETTO, João Teixeira Coelho. **Semiótica, informação e comunicação**. São Paulo: Editora Perspactiva, 2003. ISBN 85-273-0170-9

PERINI, Mario Alberto. **Gramática descritiva do português**. 3. ed. São Paulo: Editora Ática, 1998. ISBN 85-08-05550-1

POE, Edgar Allan. **The black cat**. Disponível em:
<http://bau2.uibk.ac.at/sg/poe/Work.html>
Acesso em: 23 jun. 2005.

SAGER, Naomi . **Natural Language Processing and the representation of clinical data**. Disponível em:
<http://www.cs.nyu.edu/cs/faculty/sager/NLP-RCD.pdf>
Acesso em: 23 jun. 2005.

SHEFFIELD. Sheffield NLP Group. Disponível em:
<http://nlp.shef.ac.uk/research/areas/ie.html>
Acesso em: 23 jun. 2005.

SOWA, John F. **Knowledge representation**: Logical, Philosophical and Computational Foundations. Pacific Grove: Brooks Cole Publishing Co., 2000. ISBN 0-534-94965-7

VALE, Daniel Couto. **Cognição Artificial**. Disponível em:
<http://www2.dcc.ufmg.br/~danielcv/dcc/ic3/IC_ArtificialCognition_por/title.htm>.
Acesso em: 23 jun. 2005.

_____. **Documentação**: Intelligentī Pauca V3. Disponível em:
<http://www2.dcc.ufmg.br/~danielcv/dcc/ic3/IC_IntelligentiPaucaDocumentation_por/title.htm>.
Acesso em: 23 jun. 2005.