Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Evandro Metz Flores

MODELO DE RECONHECIMENTO DE VINCULAÇÃO TEXTUAL BASEADO EM REGRAS LINGUÍSTICAS E INFORMAÇÕES MORFOSSINTÁTICAS VOLTADO PARA AMBIENTES VIRTUAIS DE ENSINO E APRENDIZAGEM

São Leopoldo, 2014

Evandro Metz Flores

MODELO DE RECONHECIMENTO DE VINCULAÇÃO TEXTUAL BASEADO EM REGRAS LINGUÍSTICAS E INFORMAÇÕES MORFOSSINTÁTICAS VOLTADO PARA AMBIENTES VIRTUAIS DE ENSINO E APRENDIZAGEM

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa Interdisciplinar de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientador: Dr. Sandro José Rigo

São Leopoldo

2014

F634m Flores, Evandro Metz.
Modelo de reconhecimento de vinculação textual baseado em regras linguísticas e informações morfossintáticas voltado para ambientes virtuais de ensino e aprendizagem / Evandro Metz Flores. – 2014.
125 f. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa Interdisciplinar de Pós-Graduação em Computação Aplicada, 2014.
"Orientador: Dr. Sandro José Rigo."

1. Reconhecimento de vinculação textual. 2 Regras linguísticas. 3. Perguntas e respostas. 4. Processamento de linguagem natural. 5. Big data. I. Título.

CDU 004

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecário: Flávio Nunes – CRB 10/1298)

Agradeço as dificuldades colocadas em meu caminho,
e as pessoas que me ajudaram a superá-las.

AGRADECIMENTOS

Agradeço a meus amigos e familiares que pacientemente ou não, entenderam a minha dedicação nesta etapa, sem mesmo ter ideia do que eu estava fazendo. Sei que estive ausente, mas enfim está acabando.

A minha namorada pela compreensão e apoio, principalmente nos finais de semana investidos em estudos.

A empresa GVdasa e ao CNPq pela bolsa, e que sem estes, nada presente neste documento existiria.

Aos meus colegas da GVwise pelo apoio e incentivo em vários momentos.

Ao meu orientador que esteve muito presente ao logo de toda a minha caminhada, me incentivando e cobrando, desde a etapa final da minha graduação.

RESUMO

A rápida evolução das tecnologias de informação e comunicação tem possibilitado o desenvolvimento de modalidades de ensino e educação, tais como a Educação a Distância, capazes de alcançar pessoas anteriormente impossibilitadas de frequentar o ensino superior. Um aspecto importante destas modalidades é o amplo uso de recursos de mediação digital, sendo que estes podem gerar um grande volume de dados o qual, por vezes, não é viável para utilização proveitosa de forma manual pelos professores envolvidos nesta interação. Este contexto gera a necessidade e oportunidade de definição de ferramentas que possam atuar para automatizar parte deste trabalho. Uma destas possibilidades é a verificação de correção de respostas textuais, onde o objetivo é identificar vinculações entre amostras textuais que podem ser, por exemplo, diferentes respostas textuais a uma pergunta.

Embora sejam utilizadas com bons resultados, as técnicas atualmente aplicadas a este problema apresentam deficiências ou características que diminuem sua precisão ou adequação em diversos contextos. Poucos trabalhos são capazes de realizar a vinculação textual caso seja alterada a flexão verbal, outros não são capazes de identificar informações importantes ou em que posição na frase as informações se encontram. Além disso, poucos trabalhos são adaptados para a língua portuguesa.

Este trabalho propõe um modelo de reconhecimento de vinculação textual baseado em regras linguísticas e informações morfossintáticas voltado para ambientes virtuais de ensino e aprendizagem, que busca contornar estes problemas apresentando uma nova abordagem através do uso combinado da análise sintática, morfológica, regras linguísticas, detecção da flexão de voz, tratamento de negação e do uso de sinônimos. O trabalho também apresenta um protótipo desenvolvido para avaliar o modelo proposto. Ao final são apresentados os resultados obtidos, que até o momento são promissores, permitindo a identificação da vinculação textual de diferentes amostras textuais com precisão e flexibilidade relevantes.

Palavras-Chave: Reconhecimento de Vinculação Textual, Regras Linguísticas, Perguntas e Respostas, Processamento de Linguagem Natural, *Big Data*.

ABSTRACT

The fast evolution of information and communication technologies has enabled the development of modalities of teaching and learning, such as distance education, that allow to reach people previously unable to attend higher education. An important aspect of these modalities is the extensive use of digital mediation resources. These resources can generate a large volume of data that sometimes is not feasible for beneficial manual use by the teachers involved in this interaction. In this context there is a necessity and opportunity for defining tools and approaches that can act to automate part of this work. One of these possibilities is the verification of textual responses correctness, where the goal is to identify linkages between textual samples, which can be, for example, different textual answer to a question.

Although presenting good results, techniques currently applied to this problem have deficiencies or characteristics that decrease their accuracy or suitability in several contexts. Few studies are able to perform textual entailment in case the verbal inflection was changed; others are not able to identify important information or position in the sentence where the information is found. Moreover, few works are adapted to Portuguese.

This work proposes a model to recognition of textual entailment based on linguistic rules, which seeks to overcome these problems by presenting a new approach through the combined use of syntactic analysis, morphology, linguistic rules, detection of the bending voice, treatment of denial and the use of synonyms. This work also presents a prototype developed to evaluate the model proposed herein. The end results, which are promising, allow the identification of textual linking of different textual samples accurately and with flexibility.

Keywords: Recognising Textual Entailment, Linguistic Rules, Question and Answering, Natural Language Processing, *Big Data*.

LISTA DE SIGLAS

EaD	Educação a Distância
AVEAs	Ambientes Virtuais de Ensino e Aprendizagem
SEED	Secretaria Especial de Educação a Distância
MOOCs	*Massive Online Open Courses*
RVT	Reconhecimento de Vinculação Textual
PLN	Processamento de Língua Natural
BNF	*Backus-Naur Form*
CG	*Constraint Grammar*
QA	*Question Answering*
NILC	Núcleo Internacional de Linguística Computacional
XML	*Extensible Markup Language*
CSV	*Comma-Separated Values*
P	Precisão
AC	Avaliações Corretas
AI	Avaliações Incorretas
SBIE	Simpósio Brasileiro de Informática na Educação
CLEI	Centro Latino-americano de Estudos em Informática

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo é apresentado o contexto geral deste trabalho e o principal problema abordado, além da questão de pesquisa, objetivos e metodologia proposta.

1.1 Contexto geral e problema abordado

A rápida evolução das tecnologias de informação e comunicação tem possibilitado o desenvolvimento de modalidades de ensino capazes de alcançar pessoas anteriormente impossibilitadas de frequentar qualquer tipo de ensino superior por motivos econômicos, por falta de tempo ou até mesmo por dificuldades relacionadas à sua localização. A Educação a Distância (EaD) é um exemplo deste avanço. Segundo o CensoEAD.BR (2011), no Brasil, de 2009 à 2011 o número de matriculados na EaD aumentou quase seiscentos por cento, passando de 528.320 matriculados em 2009 para 3.589.373 em 2011, enquanto que o número de instituições que ministram estes cursos EaD, subiram quarenta por cento, representando o número total de 128 instituições em 2009 e chegando a 181 em 2011. Estas informações indicam claramente o crescimento que ocorre no setor do EaD.

Outro aspecto a destacar é o aumento do volume de dados gerados em ações de mediação digital no ensino. O grande volume de dados produzidos atualmente não é exclusividade da modalidade de Educação a Distância, já que muitos cursos presencias e semipresenciais utilizam os recursos presentes nos Ambientes Virtuais de Ensino e Aprendizagem (AVEAs) para complementar o conteúdo utilizado em sala de aula ou até mesmo para facilitar o acesso dos alunos aos materiais de aprendizagem (DALFOVO; JOS; DOMINGUES, 2007). Entretanto são evidenciados por diversos autores, mais recentemente, tanto os grandes volumes de dados gerados pelas interações de alunos, como também os dados em formato textual. Estes representam um formato que está associado fortemente com uma grande diversidade de atividades estimuladas nos AVEAs (BAKER; ISOTANI; CARVALHO, 2011; JOHNSON et al., 2013).

No Brasil, a Educação a Distância é regulamentada pelo Ministério de Educação e Cultura (MEC), através de Secretaria Especial de Educação a Distância (SEED). Esta secretaria é responsável pela definição de normativas que estabelecem parâmetros adequados

de funcionamento para esta modalidade de educação. Um deles é o número de alunos por turma, bem como o papel dos tutores e professores. Porém mesmo com o arcabouço normativo, observa-se um grande crescimento na quantidade de turmas, sendo que em boa parte das instituições adotando esta modalidade de ensino, o limite máximo de alunos por professor ou tutor é constatado frequentemente (ABED, 2013). Portanto, percebe-se uma carga de trabalho grande para os professores e tutores, gerando dificuldades de realização de um acompanhamento detalhado de todo o conteúdo produzido pelos alunos. Esta sobrecarga de trabalho faz com que muitos professores acabem não conseguindo acompanhar de forma adequada todos os alunos, não sanando suas dúvidas corretamente ou em tempo hábil, com reflexos no baixo desempenho ou na evasão de alguns alunos (ADACHI, 2009; BAKER; ISOTANI; CARVALHO, 2011).

O novo contexto tecnológico também permite que alguns aspectos metodológicos antes inviáveis de serem amplamente adotados possam ser avaliados como alternativas concretas de aplicação. O *Mastery Learning* (BLOOM, 1968) é um ramo da psicologia da educação que presume que todas as pessoas são capazes de atingir o domínio completo de um determinado assunto, se lhe forem fornecidas as condições de aprendizagem apropriadas. O cerne deste conceito está no respeito ao ritmo individual dos estudantes, a partir do qual todos poderiam atingir o mesmo nível de compreensão sobre assuntos estudados, mesmo que em tempos diferentes. Em um trabalho precursor, Bloom (1984) apresentou evidências e discutiu a possibilidade de resultados significativamente melhores com o uso do *Mastery Learning* do que os obtidos em metodologias convencionais. Em outro experimento nesta mesma época (CLARK; GUSKEY; BENNINGA, 1983) foi demonstrada a obtenção de resultados também favoráveis com o uso desta abordagem.

Uma dificuldade para adoção deste conceito do *Mastery Learning* na prática pode ser observada na necessidade de acompanhamento individual dos estudantes, o que pode ser inviável em determinados contextos. Entretanto, parte deste acompanhamento individual pode ser realizada de forma automatizada. Isso pode ocorrer a partir da análise das respostas dos alunos a determinadas atividades, e da geração das próximas atividades de acordo com o resultado proveniente da análise da resposta anterior. A realização desta análise pode ser qualitativa. Neste caso, o resultado adquirido apresentará mais detalhes sobre o conhecimento do aluno, ao invés de simplesmente detectar que a resposta está certa ou errada (CUNNINGHAM JR., 1991). Evidências neste sentido começam a ser observadas com a adoção de cursos na modalidade aberta, tais como os MOOCs (*Massive Online Open*

Courses), que, segundo alguns autores, geram atualmente a oportunidade para que os estudantes possam utilizar recursos de *Mastery Learning* de modo proveitoso (CHUONG et al., 2013).

Uma das ferramentas mais frequentemente utilizadas pelos alunos e professores para a interação e apresentação de dúvidas e necessidades é o fórum de discussões, no qual podem ser publicadas mensagens em formato textual. O acompanhamento de dados textuais volumosos disponíveis nos fóruns de diversas turmas e cursos seria facilitado com o desenvolvimento de ferramentas automatizadas que avaliem qualitativamente a participação dos alunos entre si e junto ao professor tutor. Existem iniciativas com objetivo de auxiliar os docentes na extração de informações relevantes ao acompanhamento dos alunos. Uma destas é a MineraFórum, ferramenta que se utiliza de grafos para representar os textos a serem analisados (FABRÍCIO, 2011). Outros trabalhos buscam monitorar o comportamento dos alunos junto aos fóruns de discussão, tais como o trabalho de Kim et al. (2007), que utilizando técnicas de aprendizado de máquina classifica postagens e apresenta ao docente informações úteis sobre a colaboração de cada aluno para a solução de um problema coletivo. Ainda em Macedo et al. (2009) são destacadas as possibilidades de uso de processamento de textos para apoiar professores na avaliação de fóruns de discussão. Outros trabalhos abordam o problema da identificação de eventos expressos textualmente a partir de análises de expressões textuais, ferramentas de classificação textual ou de mineração textual (OLIVEIRA; ESMIN, 2012; SILVA et al., 2012; FABRÌCIO, 2011). Como relatado em Lesmo et al. (2013), as abordagens para identificação de eventos ou informações com base apenas em aspectos textuais podem ser melhoradas nos seus aspectos de precisão e de flexibilidade quando acrescidas de camadas adicionais de informações, tais como as geradas por elementos linguísticos ou semânticos (DEDEK; VOJTAS, 2011; WIMALASURIYA; DOU, 2010).

Portanto, percebe-se que a utilização de recursos de mediação digital na Educação a Distância e na Educação presencial, bem como algumas modalidades mais recentes, como os MOOCS, podem gerar um grande volume de dados que por vezes, devido à sua dimensão, não é viável para utilização proveitosa de forma manual pelos professores envolvidos nesta interação. Deste modo, surge a necessidade e oportunidade de definição de ferramentas e abordagens que possam atuar para automatizar parte deste trabalho.

A verificação de correção de respostas textuais é um dos focos de atuação possíveis para estas ferramentas. Neste caso específico, o objetivo está associado com o processo de

identificar vinculações entre amostras textuais, que podem ser, por exemplo, diversas respostas textuais a uma pergunta, ou então podem representar diversos comentários sobre um determinado assunto discutido em um fórum de discussões. A área de Reconhecimento de Vinculação Textual (RVT), do inglês *Recognizing Textual Entailement*, desenvolve uma série de ferramentas que podem apoiar em atividades de análise de textos em educação (DAGAN et al., 2013). Diversos níveis de representação podem ser utilizados para estas ferramentas manipularem o texto, tais como o nível léxico, sintático ou semântico. Além disso, o tratamento e análise dos textos pode ser realizado com base em diversas abordagens, envolvendo recursos como a representação e identificação de sinônimos entre termos utilizados nas respostas ou nos comentários, ou a identificação de significados para os elementos textuais e o seu relacionamento com bases de conhecimento.

Dentre as diversas abordagens para identificar a vinculação entre amostras textuais (ANDROUTSOPOULOS; MALAKASIOTIS, 2010; DAGAN et al., 2009; DAGAN et al., 2013), uma das mais comumente utilizada é a estatística, observada em um número relevante de trabalhos (FABRÍCIO; BEHAR; REATEGUI, 2011; KIM et al., 2007; LIN; HSIEH; CHUANG, 2009). Neste contexto entende-se por abordagem estatística a identificação dos termos mais importantes de uma amostra textual através do número de vezes que este termo está presente. Após identificados, estes termos são então verificados em outro *corpus* visando a sua qualificação.

Outra forma de se realizar a vinculação entre amostras textuais ocorre através da análise sintática de textos (DAGAN et. al, 2013), sendo que neste processo a classificação gramatical de cada palavra e sua posição nas sentenças é identificada e utilizada para montar uma árvore com a estrutura existente em cada *corpus*, também conhecida como árvore de dependência, por sua vez utilizada para análises posteriores, com finalidade de identificar a existência ou não de vinculação entre amostras textuais (CHEN; LIU, 2009; LAMJIRI; KOSSEIM; RADHAKRISHNAN, 2007; SHARIATMADARI; MAMAT, 2009).

A literatura apresenta ainda o reconhecimento aproximado baseado em sinônimos, que de acordo com Androutsopoulos e Malakasiotis (2010) apresenta bons resultados e está sendo amplamente utilizado, como pode ser observado em diversos trabalhos (BARBUR; BLAGA; GROZA, 2011; LIU; ZHAO; YU, 2006; RAY; SINGH; JOSHI, 2010). Nesta técnica, o objetivo é avaliar a aproximação entre um *corpus* e outro considerando, entre outras possibilidades, a consulta em um léxico de sinônimos para cada um dos elementos do *corpus*.

Embora sejam utilizadas com bons resultados, as técnicas expostas apresentam deficiências ou características que diminuem sua precisão ou adequação em diversos contextos. Grande parte dos trabalhos que se utilizam da análise estatística, por exemplo, não consideram afirmativas negativas, pois isso não altera a presença do termo na sentença. A análise léxica, no entanto, se aplicada isoladamente, possibilita apenas verificar se determinados elementos de um *corpus* estão ou não contidos dentro de outro *corpus* (ANDROUTSOPOULOS; MALAKASIOTIS, 2010, DAGAN et. al, 2013).

Conforme observado nos trabalhos consultados, as diversas abordagens utilizadas apresentam pontos específicos que representam desvantagens em determinadas situações, sendo que não são observados trabalhos que busquem implementar recursos de análise qualitativa entre trechos textuais através do uso combinado da análise sintática, morfológica, de regras linguísticas e do uso de sinônimos. Por outro lado, a crescente utilização da modalidade de Educação a Distância, dos cursos abertos (MOOCs) e a ampla adoção de AVEAs como apoio em atividades de ensino e aprendizagem configuram um cenário que pode se beneficiar da utilização de ferramentas que permitam o tratamento automático de trechos textuais, o que pode ser um elemento de apoio para os professores e para os alunos.

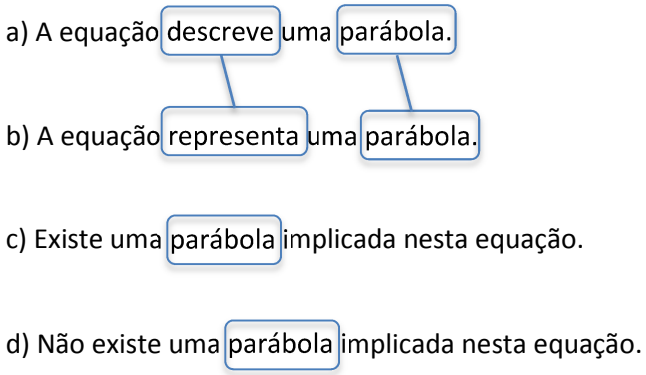
Nesta dissertação é proposta a experimentação de combinações mais representativas de informações para realizar a avaliação de vinculação entre mensagens textuais, buscando a identificação de modelos de representação que sejam qualitativamente relevantes e ao mesmo tempo possuam custo computacional baixo, de modo a permitir sua implementação em larga escala, junto de ambientes como os AVEAs, utilizados atualmente na educação presencial e na educação à distância.

A linguagem natural possui uma natureza complexa e ampla, sendo bastante variadas as possibilidades de descrição de frases e resposta de perguntas. Apesar deste contexto, parte-se do pressuposto de que em ambientes controlados, tais como os AVEAs, e quando tratando-se de temas ou domínio de conhecimento específicos, torna-se viável o tratamento de mensagens textuais em língua natural, para a finalidade de avaliação da vinculação entre frases curtas.

Um exemplo de avaliação da vinculação textual entre frases curtas e suas diversas implicações pode ser visto na Figura 1. Nesta figura observam-se quatro frases curtas contendo a resposta para uma pergunta hipotética. Considera-se que a primeira frase (no item “a“) é a frase considerada como texto ótimo, ou seja, a frase contendo a resposta correta. As

demais frases (nos itens “b“, “c“ e “d“) são os textos candidatos, ou seja, as respostas providenciadas pelos alunos e que precisam ser avaliadas. O processo de identificação da vinculação textual pode ser realizado, por exemplo, com a comparação léxica entre os elementos das frases. Neste caso, mesmo estando correta e possuindo estrutura textual semelhante ao texto ótimo, o texto candidato “b” seria penalizado na análise, por empregar um sinônimo para o verbo utilizado no texto ótimo (no caso, a palavra “representa“ em vez da palavra ”descreve”). O texto candidato do item “c“ seria penalizado neste contexto por estar estruturado em outra forma de escrita. Já o último texto candidato, item “d”, seria penalizado por estes mesmos pontos e também poderia incorporar um erro caso a negação que a palavra “não” implica não ser corretamente tratada.

Figura 1: Exemplos de abordagens para reconhecimento de vinculação textual.



Fonte: Elaborada pelo autor.

Com este breve exemplo, pretende-se indicar as possibilidades de melhorias neste processo, que são obtidas com a incorporação de maior quantidade de recursos a serem utilizados no processo de vinculação textual.

1.2 Questão de pesquisa

A partir do exposto, a seguinte questão de pesquisa foi proposta: o reconhecimento de vinculação textual com base no conjunto de informações sintáticas, informações morfológicas, regras linguísticas e descrição de sinônimos possibilita a identificação correta e flexível de respostas breves para perguntas sobre autoria e composição?

Como delimitações do estudo, considera-se como correta a resposta que está em concordância com uma resposta identificada previamente pelo autor de um conjunto de questões. A identificação flexível é considerada como sendo aquela que permite a identificação de respostas corretas mesmo com uso de elementos tais como sinônimos, elementos complementares e apostos, ou ainda com diferentes estilos de escrita, tais como voz ativa ou passiva. Por fim, o tratamento de mensagens textuais breves, contendo frases curtas com não mais de 20 palavras é um aspecto adotado neste trabalho para inicialmente proporcionar uma experiência que seja ao mesmo tempo viável computacionalmente e também relevante para alguns contextos educacionais.

1.3 Objetivos

O objetivo geral do trabalho é desenvolver e descrever um modelo que possibilite realizar vários tipos de análises que identifiquem a vinculação entre uma mensagem curta textual contendo a resposta correta para uma pergunta e outras mensagens textuais contendo as respostas de alunos para esta mesma pergunta.

Para que este objetivo seja alcançado, foram definidos os seguintes objetivos específicos:

I. Desenvolver um modelo para análise de vinculação textual, com base no estudo de trabalhos relacionados e abordagens conhecidas para esta questão.

II. Definir e desenvolver conjuntos de recursos adicionais para uso no modelo definido, tais como regras da língua portuguesa ou léxicos descrevendo relações de sinonímia.

III. Implementar um protótipo de um sistema computacional que operacionalize o modelo proposto.

IV. Aplicar e avaliar a precisão do protótipo executando suas métricas sobre um *corpora* previamente avaliado por um especialista em linguística.

Para que os objetivos traçados nesta proposta sejam concluídos, na seção seguinte será descrita a metodologia a ser utilizada.

1.4 Metodologia adotada

Para atingir os objetivos propostos e responder à questão de pesquisa foram adotadas as seguintes etapas metodológicas:

- Realização de revisão bibliográfica das áreas de conhecimento envolvidas;

- Estudo de trabalhos e aplicações relacionadas;

- Entrevistas com profissionais da área da linguística, para avaliação de possibilidades e necessidades para geração de regras linguísticas e análise textual;

- Definição do modelo geral a ser adotado;

- Implementação de um protótipo para validação de aspectos do modelo;

- Delimitação de estudos de caso;

- Realização de estudos de caso;

- Avaliação de resultados.

Com este estudo, pretende-se demonstrar o quanto a abordagem proposta favorece a construção e avaliação de alternativas para a análise textual.

1.5 Organização do texto

Este trabalho apresenta no capítulo 2 o embasamento teórico associado com esta dissertação. O capítulo 3 relata alguns trabalhos relevantes na área de reconhecimento de vinculação textual. O capítulo 4 trata da descrição do modelo proposto. O capítulo 5 descreve a implementação do protótipo construído para avaliar o modelo descrito nesta dissertação. O capitulo 6 apresenta os testes e estudos de caso realizados. O capítulo 7 traz as conclusões e as atividades futuras previstas, junto com as contribuições desta dissertação. Por fim, são descritas as referências bibliográficas utilizadas.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos que embasam este trabalho. Na seção 2.1 são apresentadas informações acerca da linguística e como a mesma contribuirá para o trabalho proposto. Na seção 2.2 é apresentada a área de Processamento de Língua Natural e suas características, seguida pela descrição do analisador morfossintático PALAVRAS na seção 2.3. Já na seção 2.4 é abordada a área de Reconhecimento de Vinculação Textual, relacionada com o campo no qual esta proposta atua. Na seção 2.5 são destacadas informações referentes à WordNet e a forma como a mesma está estruturada.

2.1 Linguística

É a área de estudo cientifico da linguagem. A pesquisa linguística é realizada por filósofos e cientistas da linguagem que se preocupam em investigar quais são as nuances e desdobramentos envolvidos na linguagem humana.

2.1.1 Divisões da linguística

Para serem estudadas de forma independente, linguistas dividem o estudo da linguagem em áreas. As mais comuns são:

a) Filologia: estudo dos textos e das linguagens antigas;

b) Fonética: estudo dos diferentes sons empregados na linguagem;

c) Fonologia: estudo dos padrões dons sons básicos de uma língua;

d) Morfologia: estudo da estrutura interna das palavras;

e) Sintaxe: estudo de como a linguagem combina palavras para formar frases gramaticais;

f) Semântica: estudo dos sentidos das frases e das palavras que a integram;

g) Lexicologia: estudo do conjunto das palavras de um idioma, área que contribui na elaboração de enciclopédias e dicionários;

h) Terminologia: estudo focado ao conhecimento e análise dos léxicos próprios das ciências e das técnicas;

i) Estilística: estudo do estilo da linguagem;

j) Pragmática: estudo do modo como as orações são empregadas;

2.1.2 Contribuição da linguística ao trabalho proposto

O cerne do trabalho proposto está em analisar respostas construídas com diferentes estruturas textuais, e poder devolver ao usuário uma correlação textual independente das estruturas gramaticais existentes nas frases. Essa funcionalidade, no entanto, só se torna viável caso sejam separadas as informações pertinentes apenas para a construção da frase, daquelas que comportam os dados realmente de interesse a representação do conhecimento. Para se realizar esse processo são adotadas regras morfológicas ou sintáticas, que são padrões de informação com o objetivo de prever o maior número de repostas possíveis, para assim identificar e, em alguns momentos, destacar palavras de interesse.

Para a detecção destes padrões de informação sintática e morfológica, se faz necessário o trabalho em conjunto com um especialista em linguística. Este profissional possui como atribuição analisar frases utilizadas para responder questões em um AVEA e representar de forma clara o que está presente nelas, em termos de padrões que relacionam informações morfológicas e sintáticas. Como exemplo pode ser citada o seguinte padrão: “Verbo” + “Informações” + “Conjunção” + “Informação”. Este padrão pode ser extraído da resposta: “São: ciano, magenta, amarelo e preto”, e, no entanto pode estar presente em qualquer resposta cuja pergunta apresente o pronome “quais”, como nas respostas: “São: branco e preto”, “São: liberdade, fraternidade e igualdade”, “São: Rio de Janeiro, São Paulo e Minas Gerais” e “São: zinco e bário”.

As possibilidades de se construir uma resposta são grandes, dada a flexibilidade da língua natural. Por isso se faz necessário utilizar um número de padrões que alcancem o máximo possível de informações presentes em uma frase. Com o padrão do último exemplo citado (“Verbo” + “Informações” + “Conjunção” + “Informação”) pode se identificar informações presentes em respostas com estruturas semelhantes, tal como nas respostas: “Elas são liberdade, fraternidade e igualdade” e “Os estados são Rio de Janeiro, São Paulo e Minas

Gerais". Porém para alguns modelos de resposta, este mesmo padrão não identifica todas as informações necessárias, tal como no caso da resposta: "Os estados que merecem destaque são Rio de Janeiro, São Paulo e Minas Gerais". Neste caso seria necessário um padrão com estrutura mais próxima da apresentada na resposta, como por exemplo, o padrão: "Substantivo" + "Informações" + "Verbo" + "Informações" + "Conjunção" + "Informação", que neste caso identificaria o trecho "...que merecem destaque..." não destacado pelo padrão anterior.

Para representar o grande número de possibilidades existentes para a construção das frases respostas, este trabalho contou com o apoio de um especialista em linguística, que analisando amostras textuais em situações reais, foi capaz de identificar estruturas morfológicas genéricas, e acrescentá-las ao banco de padrões. Para realizar a ampliação dos padrões possíveis e permitir aos usuários representarem todo o tipo de informação presente nas respostas, o modelo proposto conta com uma interface para a adição de regras linguísticas.

2.2 Processamento de Língua Natural

O conceito de Processamento de Língua Natural (PLN) teve seu inicio em torno de 1950 como uma área comum entre a linguística e a inteligência artificial. O PLN foi originalmente diferenciado da extração de informação textual, que emprega um alto percentual de técnicas baseadas em estatística para procurar e indexar eficientemente grandes volumes de textos (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Os primeiros estudos de PLN desenvolveram uma máquina de tradução palavra-por-palavra Russo-Inglês, que acabou sendo derrotado pela homografia (palavras com a mesma grafia, mas com significados diferentes) e metáforas, como no exemplo da passagem bíblica "the spirit is willing, but the flesh is weak" que foi traduzido para "the vodka is agreeable, but the meat is spoiled" (HUTCHINS, 2004).

A análise teórica de Chomsky (1956) da gramática da língua providenciou uma estimativa sobre a dificuldade deste problema, influenciando, em 1963, a criação da notação *Backus-Naur Form* (BNF), usada para especificar a "gramática livre de contexto", e é comumente utilizada para representar a sintaxe da linguagem de programação. Chomsky

também identificou gramáticas mais restritivas, criando a base das expressões regulares empregadas para especificar padrões de busca em texto (CHOMSKY, 1956).

Já na década de 70, geradores de analisadores léxicos (léxico) e geradores de análise, assim como a combinação *lex & yacc* utilizavam a gramática Levine (1992). Neste trabalho um léxico transformava um texto em um *token*, e o analisador por sua vez, validava a sequência de *tokens.* Geradores de léxico/análise simplificaram a implementação da linguagem de programação em grande parte por usarem especificações das expressões regulares e da BNF, respectivamente, como entrada, e gerar código e tabelas de pesquisa que determinam decisões léxicas/análise.

Embora a gramática livre de contexto seja, teoricamente, inadequada para a língua natural, ela é frequentemente empregada para a prática do PLN.

2.2.1 Léxico

O léxico pode ser visto como um componente das línguas que tem por função produzir, armazenar, processar e transmitir símbolos que os usuários de uma língua utilizam como matéria-prima na elaboração de raciocínios. A gramática compila as regras, as restrições e as condições que presidem ao funcionamento, aos níveis fonológico, morfológico, sintático e semântico das unidades simbólicas de texto (REU-DEBOVE; MORAIS, 1984).

No léxico há palavras invariáveis e variáveis (de acordo com sua estrutura interna), cuja configuração morfológica é afetável por variações sintaticamente determinadas. Existem também palavras funcionais, como as preposições, conjunções e conectores em geral, e as chamadas unidades lexicais, a saber: nomes, adjetivos e verbos (ALVES, 1984).

Um léxico pode ser definido como todo o universo de palavras que os usuários de uma determinada língua têm a sua disposição, e as utilizam para se expressar por escrito ou oralmente. O léxico é caracterizado pela sua mutabilidade, já que com o passar do tempo algumas palavras se tornam antiquadas, outras mudam de sentido, outras são adicionadas e isso ocorre de forma gradual gerando uma constante evolução. O usuário de uma língua utiliza o léxico para a formulação das suas expressões no intuito de efetivar o processo comunicativo. Assim, o vocabulário de um indivíduo é caracterizado pelo conhecimento que este tem do léxico e a seleção e emprego dos termos. Quanto maior for o vocabulário do

usuário, mais opções este têm para a elaboração do seu processo comunicativo (BIDERMAN, 1996).

2.2.2 Sintaxe

É a parte da gramática que trata desde a derivação de palavras mediante combinação de afixos, até o estudo das regras que presidem à combinação de palavras para a construção de uma frase.

Em uma construção, a análise compreenderá o conjunto completo até os elementos indivisíveis, a saber: mensagem ou texto (escrito ou oral), frase, sintagma ou locução, palavra e morfema. Em mensagens escritas e livros, por exemplo, ainda há a divisão de volume, capítulo e parágrafo.

Cada unidade dessas representa uma classe cujos elementos são formados pela unidade imediatamente inferior, até a classe morfema que é indivisível. Essa composição pode ser observada na Tabela 1.

Tabela 1: Elementos que compõem a comunicação

Comunicação	
Classes	**Elementos**
Texto	Frase(s)
Frase	Sintagma(s)
Sintagma	Palavra(s)
Palavra	Morfema(s)
Morfema	

Fonte: Elaborada pelo autor.

Se observa, na Tabela 1, que a maneira como são combinadas as palavras, acabam formando os sintagmas, esses que por sua vez combinados, constroem frases. A importância de cada palavra, dentro dos sintagmas, é dado o nome de sintaxe.

Por exemplo, na frase “Maria comprou um carro” de acordo com a análise sintática encontramos:

Maria: sujeito;

comprou: núcleo do predicado verbal (comprou um carro);

um: adjunto adnominal;

carro: núcleo do objeto direto (um carro).

2.2.3 Morfologia

A morfologia é a parte da gramática que estuda as palavras de acordo com a classe gramatical a que ela pertence. Atualmente as classes gramaticais estão divididas em dez: substantivos, artigos, pronomes, verbos, adjetivos, conjunções, interjeições, preposições, advérbios e numerais.

Se aplicarmos a análise morfológica sobre a mesma frase utilizada no capítulo anterior obteremos:

Maria: substantivo;

comprou: verbo;

um: artigo;

carro: substantivo.

2.3 O analisador morfossintático PALAVRAS

O sistema de análise morfossintática PALAVRAS é o resultado de uma tese desenvolvida na Universidade de Aarhus na Dinamarca, por Eckhard Bick (BICK, 2000), entre os anos de 1994 e 1999, e é descrito como um analisador baseado em gramática e léxico para textos em língua portuguesa sem restrições. O analisador em questão pode ser utilizado para aplicações como rotulação de amostras textuais, ensino da gramática e tradução de máquinas. Suas regras gramaticais são formuladas de acordo com o formalismo *Constraint Grammar* (CG) e foca em uma desambiguação robusta, tratando vários níveis de análise da língua de uma maneira relativa.

De acordo com o autor, na avaliação de textos desconhecidos, o analisador PALAVRAS atinge um resultado superior a 99% na rotulação morfológica das palavras (parte do discurso e inflexões), e cerca de 97% para funções sintáticas, mesmo orientada a total desambiguação. Entre outras coisas, estruturas de argumentos, relações de dependência e funções de sub cláusulas são tratadas de uma forma inovadora, que permite a transformação automática da base sintática primária (CG), em uma estrutura de árvore tradicional.

As informações fornecidas pelo analisador PALAVRAS são armazenadas em um arquivo XML, como visto na Figura 2.

Figura 2: Arquivo XML gerado pelo analisador PALAVRAS.

```
<?xml version="1.0" encoding="UTF-8"?>
<xml>
<corpus>

    <body>
<s id="s1" ref="1" source="Running text" forest="1" text="A equação descreve uma parábola.">
    <graph root="s1_500">
        <terminals>
            <t id="s1_1" word="A" lemma="o" pos="art" morph="F S" extra="artd"/>
            <t id="s1_2" word="equação" lemma="equação" pos="n" morph="F S" extra="cc-r"/>
            <t id="s1_3" word="descreve" lemma="descrever" pos="v-fin" morph="PR 3S IND VFIN" extra="vH fmc mv"/>
            <t id="s1_4" word="uma" lemma="um" pos="art" morph="F S" extra="arti"/>
            <t id="s1_5" word="parábola" lemma="parábola" pos="n" morph="F S" extra="ac-cat"/>
            <t id="s1_6" word="." lemma="--" pos="pu" morph="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s1_500" cat="s">
                <edge label="STA" idref="s1_501"/>
            </nt>
            <nt id="s1_501" cat="fcl">
                <edge label="S" idref="s1_502"/>
                <edge label="P" idref="s1_3"/>
                <edge label="Od" idref="s1_503"/>
                <edge label="PU" idref="s1_6"/>
            </nt>
            <nt id="s1_502" cat="np">
                <edge label="DN" idref="s1_1"/>
                <edge label="H" idref="s1_2"/>
            </nt>
            <nt id="s1_503" cat="np">
                <edge label="DN" idref="s1_4"/>
                <edge label="H" idref="s1_5"/>
            </nt>
        </nonterminals>
    </graph>
</s>

    </body>
</corpus>
</xml>
```

Fonte: Elaborada pelo autor.

Este arquivo descreve informações referentes a estrutura de árvore, e de cada palavra presente no texto. As informações referentes à árvore, definindo a ligação entre os nodos, está presente entre as marcações "*nonterminals*" e descrevem os sintagmas existentes no texto. Já as informações referentes aos nodos estão presentes entre as marcações "*terminals*", e descrevem características morfológicas e sintáticas.

2.4 Reconhecimento de vinculação textual

A inferência é normalmente percebida como o processo no qual novas consequências são concluídas partindo de uma determinada informação. Se transportada esta informação para o campo do PLN, se pode perceber a inferência sobre o estado da língua humana. Tal inferência pode ser definida como o processo de considerar correta a informação transmitida em um texto, de acordo com uma verdadeira, presente em outro texto. Essa visão de inferência orientada a língua foi gerada pelo paradigma da vinculação textual, do inglês *textual entailment*, originalmente proposto por Dagan e Glickman (2004).

A área de Reconhecimento de Vinculação Textual (RVT) apresenta um bom potencial para aplicação no campo de PLN. Por exemplo, considere o caso relacionado área de Resposta de questões, do inglês *Question Answering* (QA), nela é apresentada a seguinte questão: "Quem pintou 'O Grito'?'". Para se obter a resposta "Edvard Munch", se baseando na resposta "Quadro mais famoso da Noruega, 'O Grito' foi pintado por Edvard Munch...", o sistema de QA precisa validar a hipótese de que a informação "Edvard Munch pintou 'O Grito'.", está presente na resposta informada (DAGAN et al., 2013).

Como discutido na literatura de vinculação textual (DAGAN et al., 2009), mecanismos de inferência têm sido desenvolvidos de formas independentes e dispersas, para satisfazer a necessidade das suas áreas de aplicação. Em oposto a isso, recursos e ferramentas genéricas existem para tratar de alguns fenômenos semânticos específicos. WordNet[1] e FrameNet[2] são exemplos de repositórios criados manualmente, portadores de informações léxicas-semânticas.

[1] http://wordnet.princeton.edu/

[2] https://framenet.icsi.berkeley.edu/fndrupal/

A tarefa de reconhecimento de vinculação textual, como descrita por Dagan, Glickman e Magnini (2006), é definida da seguinte maneira:

Vinculação textual é estabelecida como uma relação direcional entre pares de expressões textuais, chamadas de T (“Texto” que está sob avaliação) e H (“Hipótese” vinculada). É dito que T vincula H, caso houver o entendimento de leitores, de forma que ao lerem a informação do texto T suponham que a informação do texto H é provavelmente verdadeira.

Podemos usar, por exemplo a frase T: “Pedro Álvarez Cabral comandou a frota que chegou ao território onde hoje se localiza o Brasil, em 1500”, que ao ser passada como correta para um leitor, permitirá o mesmo supor que a H: “Pedro descobriu o Brasil” é provavelmente verdadeira.

Como delimitado por Dagan, Glickman e Magnini (2006), esta definição é baseada no entendimento comum que os humanos possuem da língua. Esta definição é análoga à criação de “normas de ouro” para outras aplicações de compreensão de texto, como extração de informações (IE) e sistemas de Pergunta e Resposta, onde especialistas humanos precisam julgar se a resposta ou relação avaliada pode realmente ser inferida de um texto candidato. A característica que diferencia a área de reconhecimento de vinculação textual é que ela busca capturar a associação ou vinculação textual de uma forma genérica, independente de área de aplicação.

Os critérios para julgamento do RVT não são extremamente precisos, pois em alguns casos o que um leitor normalmente iria supor ao ler os conjuntos de textos pode ser relativo. No entanto, os variados esforços para a construção de diferentes formas de se validar a vinculação textual, mostram que um número consistente de avaliações humanas pode ser obtido, permitindo a evolução da pesquisa nesse campo (DAGAN et al., 2013).

2.5 WordNet

A WordNet é uma grande base de informações morfológicas agrupadas em conjuntos de sinônimos cognitivos sem ordenação (*synsets*), cada um apresentando conceito distinto. Os *synsets* são interligados por conceitos semânticos e relações lexicais. Este repositório é livre

para cópia e consultas, e sua estrutura voltada para a linguística computacional e processamento de língua natural.

WordNet levemente se assemelha a uma enciclopédia, porém o que a distingue é a forma como as palavras são aproximadas, pois isso ocorre de acordo com os seus significados e não por ordem alfabética.

Ao longo dos anos a estrutura da WordNet desenvolvida na Universidade de Princeton, foi utilizada para o desenvolvimento de repositórios específicos para cada língua. A WordNet.Br 1.0 [3] , desenvolvida pelo Núcleo Interinstitucional de Linguística Computacional (NILC), é a que armazena as informações referentes a língua portuguesa brasileira.

De acordo com o site oficial da WordNet.Br, atualmente estão presentes na primeira versão do repositório, cerca de 3700 *synsets* repartidos por vários campos semânticos e que agrupam 5860 verbos. Já na desenvolvida pela Universidade de Princeton estão presentes cerca de 117 000 *synsets.*

O uso do repositório WordNet.Br será importante nas fases seguintes do projeto, pois suas informações a cerca dos verbos foram utilizadas para realizar consultas quanto a similaridade entre os utilizados no texto candidato e no texto ótimo.

O uso desta base de informações também poderia ser aplicado às palavras pertencentes as outras classes gramaticais, porém estas ainda não estão presentes nesta primeira versão da WordNet.Br.

[3] http://www.nilc.icmc.usp.br/wordnetbr/

3 TRABALHOS RELACIONADOS

Neste capítulo são descritos e analisados trabalhos relacionados com a temática adotada. Inicialmente uma visão geral de iniciativas na área de Reconhecimento de Vinculação Textual (RVT) é apresentada, para proporcionar ao leitor a observação de alguns aspectos históricos e alguns elementos importantes para a justificativa das principais abordagens existentes. Logo em seguida são descritos em maiores detalhes as seguintes três abordagens: híbrida, semântica e baseada em grafos. Estas abordagens foram estudadas e são aqui descritas pois representam um amplo leque de possibilidades para o tratamento do problema de RVT. Ao final do capítulo são tecidos comentários sobre aspectos destas iniciativas e a sua contribuição para a delimitação da pesquisa aqui proposta.

3.1 Visão geral de iniciativas em reconhecimento de vinculação textual

As diversas iniciativas desenvolvidas para atender aos desafios de reconhecimento da vinculação textual possuem algumas características bem delimitadas quanto a suas concepções gerais para o tratamento das informações textuais. A seguir são comentados brevemente alguns destes aspectos, tal como observados em trabalhos relacionados.

Um dos aspectos importantes é a abordagem utilizada para representação das informações, sendo que muitos trabalhos adotam a forma de representação léxica como base. Esta forma utiliza as palavras dos textos analisados e a sua grafia como a informação principal a ser tratada e comparada. O sistema de Herrera et al., (2005), por exemplo, utiliza esta abordagem e também amplia a representação possível com as relações descritas no WordNet (MILLER, et al. 1993). São incluídas as relações descrevendo sinônimos, hipônimos ou antônimos, entre outros, na descrição das informações. Já no trabalho de Montalvo-Huhn e Taylor (2008) utiliza-se a representação léxica mas neste caso são aplicadas várias métricas de comparação envolvendo as palavras originais, os sinônimos e antônimos. Cada uma das métricas apoia a métrica final utilizada para comparação da vinculação textual.

A abordagem léxica pode ser ampliada, por exemplo, com a inclusão de aspectos de contexto, como observado no trabalho de Clarke (2006). Este trabalho propõe uma avaliação de relacionamento textual com a ampliação da análise de palavras para termos compostos e

sequências de palavras. Deste modo é possível a identificação mais precisa do significado das palavras de acordo com o seu contexto, dado, neste caso, pelo conjunto de palavras vizinhas à palavra analisada.

Os trabalhos adotando a perspectiva de alguma representação de contextos podem usar aspectos mais abrangentes do que apenas o conjunto de palavras vizinhas. Por exemplo, o diferencial do trabalho de Litkowski's (2009) consiste em privilegiar a utilização de comparações entre entidades nomeadas, dentro das etapas de avaliações dos textos comparados. Já em uma abordagem relacionada, Majumdar e Bhattacharyya (2010) descrevem um sistema baseado em representação léxica, porém que faz uso de correferência entre os elementos dos textos tratados.

As métricas utilizadas para a decisão sobre a vinculação textual complementam o aspecto de representação das informações adotado em cada abordagem e também são delimitadas pelas características destas representações. Alguns trabalhos utilizam, na etapa de decisão e avaliação de similaridade, os recursos de sistemas como os classificadores. Por exemplo, no trabalho de Adams et al. (2007) uma árvore de classificação é utilizada, com base em uso de aspectos linguísticos como características para a decisão.

Com base em utilização de recursos adicionais, tais como informações sintáticas e thesaurus, o sistema MENT (VANDERWENDE et al., 2006) apresenta resultados interessantes, tais como precisão de 60.25% no teste RTE-2[4]. Outro tipo de representação utilizada com bons resultados é a representação de estruturas textuais, tal como no trabalho de Wang e Neumann (2007), que apresentam uma abordagem que descreve uma estrutura textual identificada automaticamente e a emprega para compor a decisão de similaridade. De forma relacionada, a ideia principal do trabalho de Varma et al. (2009) é identificar estruturas linguísticas, gerando padrões que permitem relacionar elementos dos dois textos sendo comparados.

Já o trabalho de Tsuchida e Ishikawa (2011) combina a utilização de uma pontuação gerada com base em nível de representação lexical e ampliada com mecanismos baseados em aprendizagem de máquina, que empregam diversas características originadas no nível léxico, como estruturas de sentenças e predicados.

[4] http://pascallin.ecs.soton.ac.uk/Challenges/RTE2/Datasets/

De uma forma geral foi verificada que os principais trabalhos pesquisados nesta dissertação, apresentaram como características estarem baseados em três áreas, a saber: híbrida, grafos ou semântica. Por este motivo, os trabalhos analisados nesta seção serão divididos de acordo com esta característica principal.

3.2 Abordagem híbrida

As abordagens híbridas são aquelas onde os autores descrevem seus trabalhos como apoiados em recursos e métricas diferentes de uma maneira a não valorizar mais a uma do que as outras. Nesta seção será apresentado o trabalho com esta característica em destaque.

O trabalho de Zhang et al. (2012) descreve um sistema desenvolvido pelo próprio autor chamado SNRTE, que combina informações léxicas, sintáticas, e 3 níveis de análises semânticas, com o auxílio de ferramentas de PLN que incluem analisadores e rotuladores de parte no discurso e WordNet.

De acordo com o autor a execução do seu sistema ocorre em quatro etapas. Na primeira há um pré-processamento onde os verbos são utilizados na sua forma sem conjugação e longos parágrafos são separados em sentenças curtas. Na segunda etapa as palavras são armazenadas individualmente e uma árvore gramatical é montada para a realização de uma primeira comparação entre as amostras. O terceiro passo realiza análises semânticas identificando sete elementos, a saber: pessoa, localização, tempo, data, dinheiro, organização e percentual. Após identificados estes elementos, são aplicadas estratégia diferentes para ponderar os resultados encontrados. Na quarta etapa 800 pares de amostras textuais (contendo textos candidatos e texto ótimo) são utilizados para treinamento gerando uma predição, esta é então utilizada para avaliar a amostra textual.

3.3 Abordagem com o foco na semântica

Semântica é o estudo do significado, e que ocorre sobre a relação entre significantes, tais como símbolos, sinais, frases ou palavras (NETO; INFANTE, 2008). Nesta seção serão apresentados os trabalhos que são fortemente baseados nos grupos de significados formados pelas palavras contidas nas suas amostras textuais analisadas.

Aplicado a área de *Question Answering*, o trabalho de Iftene (2008) está extensamente baseado na semântica, sendo que recursos para apoiar esta abordagem são obtidos em bases de conhecimento como DIRT (DEKANG E PATRICK, 2001), WordNet e Wikipedia[5]. Adicionalmente o sistema aplica complexas regras gramaticais para reformular as amostras e usa os resultados de um módulo desenvolvido pelos autores, para adquirir o conhecimento adicional necessário. O sistema apresentado pelo autor é composto por duas partes principais: a primeira realiza o pré-processamento e a segunda, que através da saída da primeira, gera uma resposta.

A etapa de pré-processamento se utiliza de recursos como o MINIPAR[6] e o LingPipe[7]. O primeiro é capaz de gerar as árvores de dependência e rotular sintaticamente texto candidato e ótimo, enquanto que a segunda trata da identificação das entidades nomeadas. A análise gerada pelo MINIPAR é realizada em formato de grafo, onde as palavras ficam localizadas nos nodos e a ligação entre elas depende da relação entre as palavras. Este conjunto de três informações é chamado pelos autores de entidade e é considerado uma relação semântica. Após mapeadas todas as entidades em ambos os textos, são utilizados os repositórios de informações para realizar pesquisas léxicas e semânticas sobre as informações em cada entidade, com o objetivo de encontrar todas as informações do texto ótimo no texto candidato.

Para o tratamento da negação os autores desenvolveram manualmente uma lista de palavras que são capazes de alterar a função de um verbo e que quando pertencentes a uma das entidades atuam sobre a relação entre os nodos. Quando as entidades são correlacionadas em ambas as amostras textuais, as pertencentes a amostra candidata recebem uma pontuação de acordo com a necessidade e número de pesquisas realizadas para alterar a entidade e deixa-la semelhante a do texto ótimo. Ou seja, entidades encontradas no texto candidato e que estão exatamente iguais no texto ótimo, recebem uma pontuação maior do que entidades que precisaram ter seu verbo alterado.

O trabalho de Rios (2012), que utiliza semântica em combinação com uma abordagem de edição de distância semântica, onde pontos são concedidos ao texto candidato de acordo com similaridade léxica e semântica entre os termos de ambas as amostras textuais. Ou seja,

[5] https://www.wikipedia.org/

[6] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

[7] http://alias-i.com/lingpipe/

será maior a pontuação associada ao texto candidato se um número menor de consultas forem necessárias para verificar a vinculação textual entre as amostras. Segundo os autores o diferencial deste trabalho está na utilização de técnicas de aprendizagem de máquina, o qual utiliza os pontos gerados pelas duas primeiras métricas para criar uma base de conhecimento suficiente para, no futuro descartar as consultas aos repositórios externos.

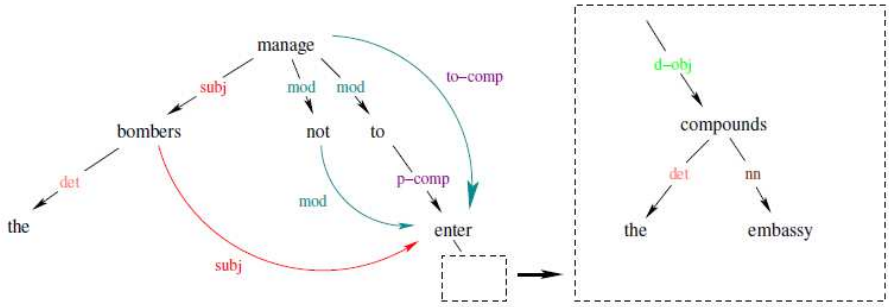
Os trabalhos de Roth e Sammons (2007), Huang, Thint e Celikyilmaz (2009) e Bar-Haim et al. (2007) também apresentam organização semelhante, inferindo notas a conjuntos de informações semânticas de acordo com a necessidade de consultas.

3.4 Abordagem baseado em grafos

Os grafos representam a dependência entre os grupos de palavras presentes nas amostras textuais. Através destes grafos de dependência é possível destacar quais grupos de palavras estão subordinados ou são complementares aos demais. Os trabalhos a seguir apresentam como fundamento características deste tipo.

No trabalho de Rus (2005) é considerado o contexto léxico-sintático tanto do texto candidato quanto ótimo e o trabalho se baseia fortemente no conceito de subordinação. O sistema desenvolvido pelos autores inicia mapeando os textos candidato e ótimo em grafos estruturados Figura 3. Nestes grafos os nodos representam os conceitos principais e as ligações entre estes são criadas de acordo com as respectivas representações sintáticas. Uma pontuação de vinculação é então computada, quantificando o grau no qual o grafo do texto ótimo se subordina ao grafo do candidato. De acordo com os autores, os resultados colhidos, quando comparados a os resultados de outros projetos que utilizam os mesmos recursos, apresentam maior acurácia na indicação de vinculação textual.

Figura 3: Exemplo de grafo para o texto: "The bombers not manage to enter the compounds embassy".

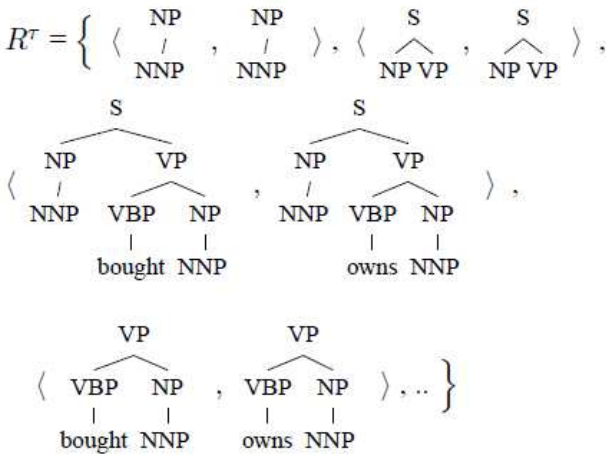


Fonte: Rus (2005).

O trabalho de Moschitti e Zanzotto (2008) demonstra a representação estatística de um aprendizado de máquina via grafos sintáticos constituídos por pares de árvores. Com isso os autores pretendem demonstrar que a forma natural de se representar a relação sintática entre texto candidato e ótimo consiste nas possibilidades de pares de fragmentos sintáticos.

No exemplo demonstrado pelos autores, estão presentes os seguintes elementos: o texto ótimo seria "Wanadoo owns KStones", enquanto que o texto candidato seria "Wanadoo bought KStones". Estes textos, após analisados pelo seu sistema, apresentam o resultado demonstrado na Figura 4.

Figura 4: Resultado da análise entre texto ótimo e candidato



Fonte: Moschitti e Zanzotto (2008).

Neste caso as combinações entre texto ótimo e candidato são agrupadas e separadas com vírgulas das demais combinações. No exemplo demonstrado na Figura 4, o sistema encontra uma totalidade de três combinações de fragmentos sintáticos na totalidade das amostras, restando apenas uma diferença léxica entre os verbos "bought" e "owns". Após a

construção dos grafos é iniciada a fase de pontuação, onde os fragmentos sintáticos são comparados gerando uma pontuação. Os pontos recebidos pelo texto candidato indicam o grau de vinculação entre as amostras textuais.

3.5 Análise dos trabalhos relacionados

Os trabalhos mencionados anteriormente apresentam aspectos relacionados as diferentes áreas e informações em que são baseadas as iniciativas de sistemas para RVT.

Os trabalhos, descritos no item 3.3, apresentam como cerne a semântica para realizar a vinculação textual, ou seja, de cada texto analisado é extraído um ou mais significados presentes. Caso seja apresentado mais do que um significado em um texto, ocorre uma tentativa de combinação entre as semânticas extraídas, com o objetivo de simplificar a comparação que será realizada entre as duas amostras textuais. Estes trabalhos se assemelham ao aqui proposto, pois também fazem uso de axiomas, porém são axiomas semânticos, enquanto que o descrito nesta dissertação faz uso de padrões morfológicos e sintáticos. Os trabalhos de Iftene (2008), Rios (2012), Huang, Thint e Celikyilmaz (2009) e Bar-Haim et al. (2007) também realizam a rotulação das palavras de forma individual.

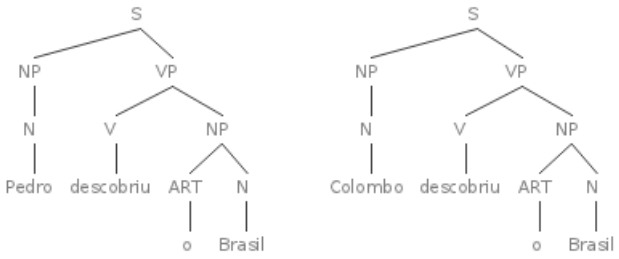
O ponto fraco dos trabalhos descritos no item 3.3, quando aplicado a língua portuguesa brasileira, está presente no uso da semântica. Para realizar a vinculação textual com base nestes recursos são necessárias, além dos axiomas semânticos, axiomas que tratem da combinação entre axiomas semânticos, elevando de forma exponencial o número de axiomas necessários para se extrair o conhecimento de interesse no texto. E este tipo de informação, como dito pelos próprios autores, é escasso, mesmo nos repositórios que buscam tratar estes assuntos. O uso das relações semânticas também dificulta a aplicação deste trabalho à língua portuguesa, justamente por estes serem quase inexistentes nos repositórios da língua. A influência da voz passiva ou ativa também não é descrita pelos autores.

Já os trabalhos descritos no item 3.4 fazem uso de informações provenientes da correspondência entre grafos. Utilizando informações de dependência ou sintáticas os trabalhos descritos por Moschitti e Zanzotto (2008) e Rus (2005) buscam a semelhança entre subconjuntos presentes tanto no texto ótimo como nos textos candidatos. Informações léxicas

provenientes de um analisador externo também são utilizadas, além de consultas ao WordNet, assim como o sistema descrito nesta dissertação.

No entanto, o uso que o trabalho faz de técnicas que realizam a correspondência entre grafos acaba por gerar uma estrutura de análise frágil. Isso ocorre na análise entre textos muito semelhantes, porém com significados distintos, o diagnóstico pode apresentar um grande grau de vinculação, como pode ser visto na Figura 5.

Figura 5: Análise de dependência entre duas amostras semelhantes.

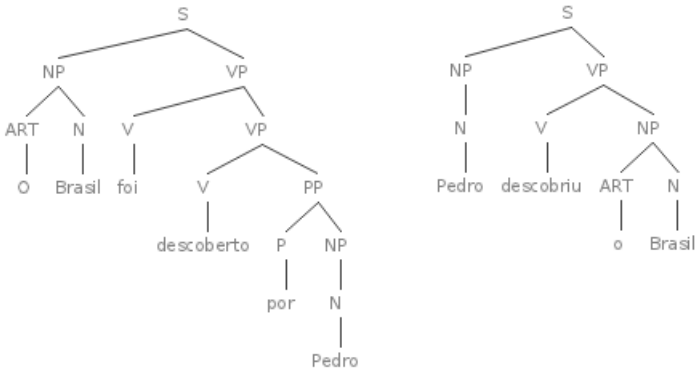


Fonte: Elaborada pelo autor.

Neste caso, a resposta correta apresenta a esquerda da Figura 5, apresenta exatamente a mesma estrutura de dependência apresentada na resposta incorreta apresentada a direita. Neste caso a indicação referente a resposta incorreta, presente no primeiro substantivo, irá ser verificada pelo apenas pelo léxico. Uma métrica simples que envolva apenas a comparação de palavras iria indicar neste caso o valor da vinculação textual igual a 75%.

Outro aspecto não abordado por trabalhos com a abordagem baseada em grafos é a vinculação textual entre textos com vozes alteradas, como pode ser observado na Figura 6. Nela tanto o grafo da esquerda quanto o grafo da direita apresentam exatamente a mesma informação a ser inferida. Porém se considerarmos o texto presente no grafo da esquerda como texto ótimo, alguns vértices e ligações que estão presentes não são representados pelo grafo da direita, e por isso uma análise de subordinação não encontrará vinculação entre as amostras.

Figura 6: Análise de dependência entre duas amostras com vozes diferentes.



Fonte: Elaborada pelo autor.

O trabalho descrito no item 3.2, no entanto, está muito próximo ao apresentado como proposta neste documento, pois se utiliza de vários recursos para realizar a análise de vinculação textual. Características apresentadas pelo autor como o uso do verbo sem conjugação, palavras analisadas individualmente e consulta a repositórios estão presentes no trabalho descrito nesta dissertação. A característica de dividir as amostras em sentenças curtas se aproxima muito da divisão das regras linguísticas em sequências presentes neste trabalho e que será apresentado no capítulo 4.3.

O uso da semântica faz com que o trabalho descrito por Zhang (2012) dificilmente possa ser adaptado a língua portuguesa brasileira, devido a ausência de um repositório para tal. No trabalho apresentado neste documento se pretende substituir o papel exercido por este recurso com o uso das regras linguísticas, que apesar de não possuírem essencialmente a mesma função, podem simplificar a necessidade de tratamento das amostras textuais viabilizando a análise da vinculação textual.

A Tabela 2 apresenta de forma resumida os principais trabalhos descritos neste capítulo, e o presente neste documento, juntamente com os recursos utilizados pelos mesmos.

Cada coluna identifica as características dos trabalhos analisados. Os itens analisados são descritos a seguir. A primeira linha trata da consulta a repositórios, ou seja, se o projeto apresentado menciona a consulta a repositórios de informação que pertençam a terceiros. A linha seguinte menciona a inferência propositiva, onde o próprio algoritmo busca mais amostras até poder avaliar as amostras textuais. A terceira linha indica os trabalhos que apresentam técnicas de aprendizado de máquina para treinar o algoritmo com a resposta correta. As linhas seguintes indicam se os trabalhos utilizam informações morfológicas,

sintáticas e léxicas, nesta ordem. A sétima linha indica os trabalhos que se baseiam nos grafos de dependência, considerando a estrutura da frase e dependência entre os segmentos da mesma. A oitava linha aponta os trabalhos baseados nas informações semânticas, que são os significados contidos nas frases. Na linha seguinte estão marcados os trabalhos baseados em regras linguísticas, que são as estruturas comumente utilizadas para a criação de respostas dissertativas. E por último estão descritos os resultados atingidos pelos trabalhos com uso de uma determinada base de dados, também especificada nesta mesma linha.

Tabela 2: Comparação entre os recursos utilizados pelos trabalhos relacionados

	Iftene (2008)	Rios (2012)	Roth e Sammons (2007)	Rus (2005)	Moschitti e Zanzotto (2008)	Zhang (2012)	Metz
Consulta repositórios	✓		✓	✓	✓		✓
Inferência Propositiva			✓				
Aprendizado de máquina	✓	✓			✓		
Morfologia							✓
Sintaxe				✓	✓	✓	✓
Léxico	✓	✓	✓	✓		✓	✓
Grafo de Dependência	✓			✓	✓		
Semântica	✓	✓	✓			✓	
Regras Linguísticas							✓
Percentual de acerto	69,13% RTE3	63,8% RTE3	65,56% RTE3		52,6% RTE2	67,5% RTE1	

Fonte: Elaborada pelo autor.

Alguns aspectos que surgem evidenciados desta análise são empregados como balizadores do modelo proposto. Como um dos objetivos desta dissertação é auxiliar os professores das modalidades de Educação a Distância a terem um melhor acompanhamento dos dados textuais construídos por seus alunos, a possibilidade de utilizar aprendizagem de máquina foi desconsiderada, tendo em vista que isso demandaria uma estrutura incompatível com o objetivo de implementação de menor custo computacional. Já a utilização de

repositórios semânticos, apesar de interessante do ponto de vista dos resultados e ganhos possíveis, foram desconsiderados no modelo proposto devido às dificuldades práticas para, atualmente, dispor deste repositório para a língua portuguesa.

O trabalho apresentado nesta dissertação tem como diferencial a combinação de dados provenientes de distintas fontes de informações e de regras linguísticas, apresentando assim uma proposta para desenvolver formas de reconhecimento da vinculação textual de modo flexível e com custo computacional adequado.

4 MODELO PROPOSTO

Neste capítulo serão apresentadas as escolhas que compõem o modelo proposto e que embasam o protótipo desenvolvido.

4.1 Aspectos introdutórios

O objetivo do modelo proposto é embasar a construção de um protótipo de sistema de informação que permita investigar o uso de técnicas para reconhecimento de vinculação textual. A abordagem adotada pretende, através do uso de informações morfológicas, sintáticas, regras linguísticas e sinônimos, realizar o reconhecimento de vinculação textual de respostas candidatas quando comparadas a uma resposta ótima. Neste caso, se considera que uma resposta candidata seria uma resposta fornecida por um aluno e a resposta ótima seria a resposta fornecida pelo professor. Para a geração desta comparação são utilizadas métricas específicas que podem variar de acordo com os recursos de representação utilizados.

Existem duas informações relevantes para o modelo proposto. A primeira informação se refere às regras linguísticas, estrutura existente em qualquer construção de raciocínio escrito pertinente à língua portuguesa. A segunda informação se refere à resposta ótima indicada e construída na maioria das vezes pelo professor, portadora de todas as informações fundamentais com a representação correta e satisfatória da resposta. A primeira informação possui característica dinâmica e poderá ser usada em todas as análises futuras, já a segunda informação é específica para cada questão ou tema a ser abordado.

A avaliação ou reconhecimento de vinculação textual ocorre quando se compara uma amostra textual que contém a hipótese de uma resposta correta, que pode ter sido construída por um aluno, com uma amostra que contém o conteúdo representante de todo o conhecimento necessário para se identificar determinada resposta, texto este normalmente desenvolvido por um professor. Estas amostras serão chamadas de texto candidato e texto ótimo respectivamente.

Esta análise pode ser realizada de diversas formas, como visto anteriormente. E o somatório de algumas técnicas de Processamento de Língua Natural (PLN) podem auxiliar na melhor avaliação da vinculação entre textos, pois através destas, é possível destacar as

informações importantes em um texto ótimo, e usá-las como informações a serem buscadas nos textos candidatos a que serão analisados.

Cabe destacar que o objetivo deste trabalho é atuar no contexto de perguntas e respostas textuais que possuem certa restrição quanto ao seu formato e sua abrangência. Esta delimitação é necessária dado que a linguagem natural apresenta uma grande complexidade para o seu tratamento, em virtude da flexibilidade existente na organização das sentenças, no uso de termos polissêmicos ou no emprego de figuras de linguagem, entre outros aspectos. Por estes motivos somente serão tratadas neste trabalho perguntas do tipo Qual/Quais e Quem, como por exemplo “Quais são os estados que compõem a região sul?” e “Quem foi o primeiro homem a pisar na lua” e que apresentam respostas curtas como “É composta por três estados: Rio Grande do Sul, Santa Catarina e Paraná” e “O primeiro homem a pisar na lua foi Neil Armstrong”.

4.2 Visão geral do modelo proposto

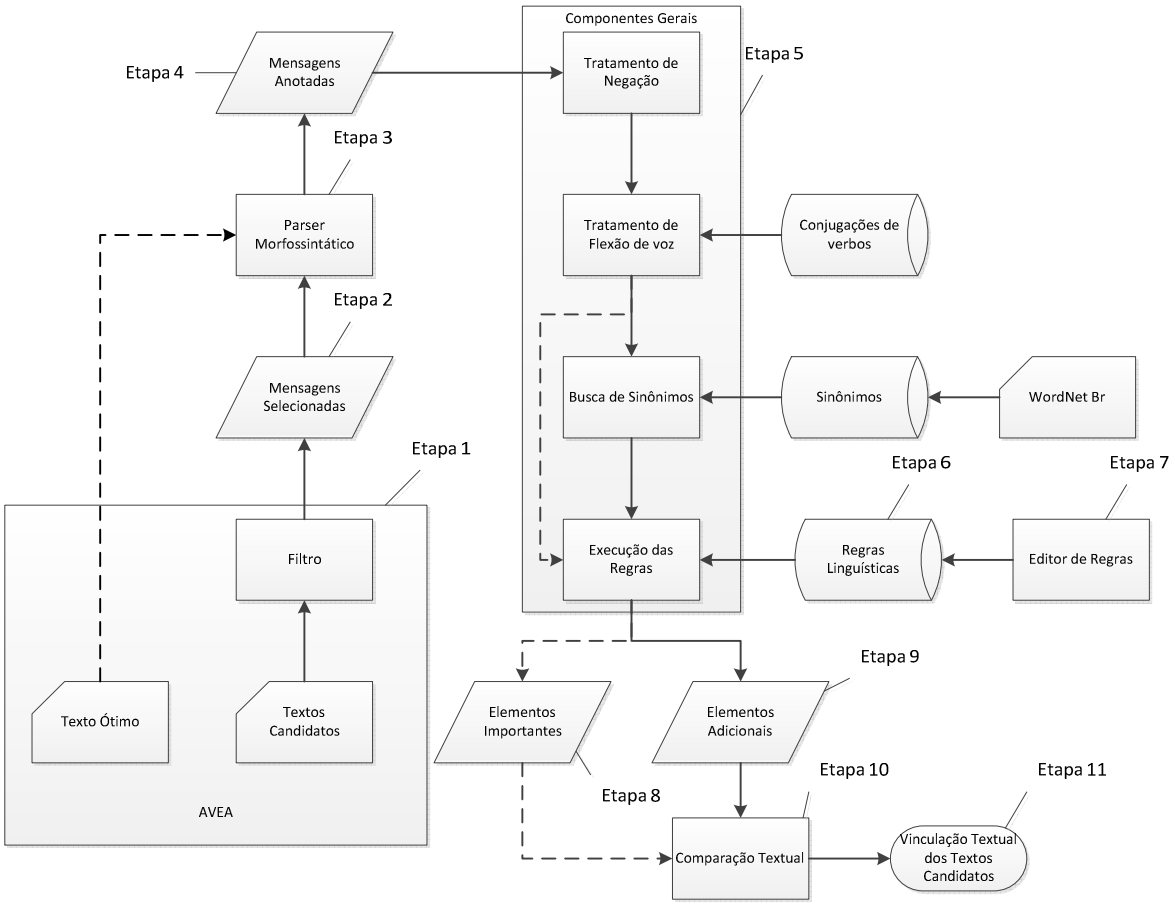
A visão geral do modelo proposto pode ser vista na Figura 7. A dinâmica de funcionamento do modelo e o detalhamento dos seus componentes está comentada a seguir.

O modelo considera a integração com um AVEA, sendo que a etapa 1 descrita na Figura 7 considera justamente o tratamento de amostras textuais presentes em um AVEA. Nesta etapa inicial o usuário do protótipo construído com base no modelo proposto, normalmente o professor de uma disciplina, deve indicar qual é a pergunta e qual é o texto que contém as informações importantes e que devem estar contidas nas respostas dos alunos, sendo este texto, conforme já mencionado, chamado de texto ótimo. O usuário também deve indicar quais as respostas dos alunos que serão avaliadas, gerando assim as mensagens selecionadas presentes na etapa 2. Estas operações estão representadas na Figura 7 através dos componentes “Texto ótimo“, “Texto Candidato“ e “Filtro“. No modelo são considerados mecanismos para automatizar estas atividades, com base na associação desta coleta e seleção de dados com os recursos existentes nos AVEAs, de forma flexível. O objetivo é permitir que o modelo seja utilizado com um conjunto extenso de AVEAs.

Após selecionadas, as mensagens contento tanto o texto ótimo como os textos candidatos são anotadas pelo analisador morfossintático escolhido para uso no modelo. No

caso da implementação inicial deste modelo, foi empregado o analisador Palavras (BICK, 2000), devido a sua grande precisão. Entretanto o modelo permite que outros recursos sejam utilizados nesta etapa, sem prejuízo de seu funcionamento. A atividade de anotação dos textos está indicada na Figura 7 como sendo a etapa 3, que deve gerar como resultado as respectivas mensagens anotadas em um formato adequado para manipulação, na etapa 4. Habitualmente os sistemas de anotação morfossintática proporcionam como resultado arquivos textuais estruturados, normalmente em formato XML (*eXtensible Markup Language*), contendo as informações morfossintáticas necessárias para a avaliação da vinculação textual.

Figura 7: Visão geral do modelo proposto.



Fonte: Elaborada pelo autor.

As mensagens textuais anotadas são manipuladas na etapa 5, através de diversas operações que definem o modelo e aspectos de sua flexibilidade. Inicialmente é realizado o tratamento da negação nas frases, com o objetivo de armazenar qualquer informação sobre negação que pode estar presente no texto, bem como para qual ação ou estado está associada. Na segunda operação é verificada a questão ligada à flexão de voz, para garantir que

diferenças entre o uso de voz ativa ou voz passiva nas frases não gere interferências na avaliação final das respostas. Para tal, se faz necessária a alteração da estrutura da frase e troca de tempo verbal do verbo principal quando existe esta diferença no uso de tipos de voz no texto. Após a alteração da voz, o modelo prevê uma operação onde os textos candidatos têm seus verbos comparados aos do texto ótimo. Estes verbos são comparados entre si e nesta comparação se utiliza o auxílio de uma base de informações extraída do repositório WordNet Br[8], que proporciona o acesso à sinônimos para os verbos encontrados nas frases.

A última operação da etapa 5 trata da execução de regras linguísticas sobre os textos ótimo e candidato. O objetivo desta operação é identificar qual regra linguística abrange maior parte do texto sendo analisado. Estas regras linguísticas são provenientes de um repositório de regras linguísticas (etapa 6) que está associado a um editor de regras linguísticas (etapa 7), onde estas podem ser adicionadas ou removidas por especialistas em linguística.

Na etapa 8 está descrita uma operação que utiliza a regra linguística que foi selecionada como a de melhor aderência para com o texto ótimo com a finalidade de destacar no texto a localização das palavras chaves importantes para a resposta. Ao final desta operação estas palavras são armazenadas e identificadas como elementos importantes para análise da resposta.

Na etapa 9 é realizada uma operação semelhante à operação da etapa 8. A etapa 9 tem como objetivo aplicar a regra linguística selecionada para cada um dos textos candidatos com o objetivo de destacar no texto a localização das palavras utilizadas pelo aluno que são candidatas a estarem vinculadas as do professor. Estas palavras são armazenadas e identificadas como elementos candidatos.

A identificação dos termos considerados como elementos importantes (no texto ótimo) e como elementos candidatos (no texto candidato) é realizada a partir da descrição das regras linguísticas, processo que será detalhado adiante neste texto, no item 4.3.

Na etapa 10 e na etapa 11 são realizadas as comparações entre os elementos importantes e elementos candidatos de cada conjunto de textos. A este processo é adicionada a execução de uma métrica definida no modelo para possibilitar a identificação de uma categorização entre estas frases. Desta forma, a quantidade de palavras encontradas de forma

[8] http://www.nilc.icmc.usp.br/wordnetbr/

coincidente nos dois conjuntos indica o percentual de vinculação textual entre texto ótimo e candidato.

Dentro deste contexto geral do modelo proposto, são destacados a seguir maiores detalhes de seus componentes.

4.2.1 Seleção de mensagens textuais

Este item está associado com o componente destacado como etapa 1 na Figura 7. Um dos objetivos deste modelo é caracterizar este componente como flexível, a ponto de permitir que diferentes AVEAs ou diferentes ferramentas de interação digital possam ser utilizadas para a seleção e coleta de conjuntos de mensagens contendo a pergunta, o texto ótimo e os textos candidatos.

As mensagens textuais consideradas como o alvo do material a ser tratado neste modelo, no contexto da implementação do protótipo de validação, são aquelas apresentadas em disciplinas onde o seu conteúdo está restrito a área das Ciências Exatas. Estas mensagens seriam provenientes de ferramentas de interação online do tipo questionários, onde o professor/tutor apresenta a questão aos alunos e obtém a respostas dos mesmos através do mesmo meio, viabilizando assim que o professor apresente para o modelo a resposta ótima, que será utilizada para avaliar as respostas que serão fornecidas pelo corpo docente.

4.2.2 Anotação com *Parser* Morfossintático

Esta atividade está associada com as etapas 3 e 4 da Figura 7 e descreve a anotação das frases com o *parser* morfossintático. Esta operação irá rotular todas as mensagens a serem analisadas, apresentando ao final de seu processo um arquivo textual, geralmente em formato XML, com todas as informações morfossintáticas necessárias para o andamento da análise de vinculação textual. Uma das necessidades que determinou a inclusão desta atividade está relacionada com a utilização de regras linguísticas em etapas posteriores.

No protótipo descrito neste texto, criado a partir deste modelo proposto, foi realizada a escolha de um *parser* morfossintático específico, o *parser* palavras (BICK, 2000), mas a proposta do modelo inclui a possibilidade de alternar esta escolha, de acordo com

conveniências de acesso ou disponibilidade das ferramentas nesta área, ou então de acordo com a necessidade do conjunto de atividades a serem utilizadas na própria atividade geral de reconhecimento de vinculação textual.

4.2.3 Geração de regras linguísticas

No modelo geral, temos os elementos 6 e 7 relacionados com a geração e manutenção de regras linguísticas. Fundamentais a construção do modelo proposto, as regras linguísticas têm como objetivo principal extrair e rotular cada informação existente tanto na resposta ótima quanto nas respostas candidato. Estas regras também possibilitam a identificação de voz passiva e ativa presente tanto no texto ótimo quanto no candidato. As regras são geradas manualmente por especialistas em linguística, com uso de um editor desenvolvido junto do escopo do modelo e no protótipo implementado.

Este recurso de regras linguísticas possui como objetivo flexibilizar a análise das frases escritas em diferentes formatos. Por exemplo, na apresentação da pergunta “Quais cores apresenta o padrão CMYK?” podem ser obtidas, entre outras, as seguintes respostas, todas corretas, porém escritas em formatos diferenciados:

“As cores do padrão CMYK são: ciano, magenta, amarelo e preto”.

“São: ciano, magenta, amarelo e preto”.

“Ciano, magenta, amarelo e preto”.

“As cores são: ciano, magenta, amarelo e preto”.

“O padrão CMYK apresenta as cores ciano, magenta, amarelo e preto”

“Ciano, magenta, amarelo e preto, são as cores”.

“Ciano, magenta, amarelo e preto, são as cores do padrão CMYK”.

“Ciano, magenta, amarelo e preto, segundo fulano de tal, são as cores do padrão CMYK”.

O número de respostas que podem ser fornecidas a uma pergunta pode ser muito grande, porém algumas informações são fundamentais para a representação do conhecimento requisitado pelo professor. Como é improvável que alunos e professor construam a resposta

da questão exatamente da mesma forma, se faz necessário o uso das regras linguísticas para distinguir as informações pertinentes a construção da resposta daquelas que são realmente importantes na avaliação do conhecimento presente no texto. As regras linguísticas são compostas por uma sequência de elementos textuais, de acordo com aspectos de morfologia, junto com elementos que indicam os elementos a serem comparados no texto, denominados anteriormente de elementos importantes ou elementos candidatos.

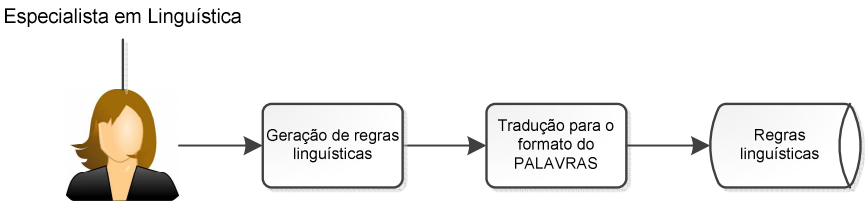
Por exemplo, se o professor apresentar a resposta: “São: ciano, magenta, amarelo e preto” como a resposta ótima, a seguinte regra linguística poderá ser utilizada para descrever o padrão de informações morfossintáticas observadas nesta frase: “Verbo” + “Informações Importantes” + “Conjunção” + “Informação Importante”.

Já no caso do aluno afirmar a seguinte resposta: “As cores são: ciano, magenta, amarelo e preto”, pode ser utilizada a regra linguística: “Artigo” + “Substantivo” + “Verbo” + “Informações Candidatas” + “Conjunção” + “Informação Candidata”. Neste momento embora apresentando respostas diferentes, tanto professor quanto aluno respondem a esta pergunta mencionando as mesmas cores, ou seja, a resposta do aluno, embora apresente uma regra diferente da do professor, está correta. A comparação para determinar a correção da resposta será feita com base nos elementos identificados como “Informação Candidata” e “Informação Importante” descritas nas regras.

O uso destas regras linguísticas, tal como visto acima, é fundamental para a identificação correta da resposta, porém se faz necessário um grande número de regras para se alcançar o tratamento correto do maior número de respostas possíveis. No entanto, estas regras são comuns a qualquer tipo de análise qualitativa textual a ser realizada futuramente, ou seja, depois de construído o repositório de regras, a estas poderão ser realizadas apenas algumas adições, mas sem demandar um maior envolvimento de um especialista.

Por este motivo, no protótipo construído para validar o modelo proposto, foi construída uma interface, com o objetivo de facilitar a ampliação do banco de regras linguísticas. Esta interface será utilizada por especialistas na área que poderão, através de uma mediação amigável, construir a representação de frases através de suas informações morfológicas. Estas informações serão então adequadas para o formato encontrado nas marcações utilizadas pelo analisador morfológico escolhido e só assim armazenadas. As etapas principais deste processo podem ser vistas na Figura 8.

Figura 8: Geração de regras Linguísticas.



Fonte: Elaborada pelo autor.

4.2.4 Tratamento de Negação

O tratamento da negação está ligado a identificação dos advérbios de negação, que são palavras que pertencem a uma subclasse dos advérbios, e que podem ser modificadores do grupo verbal ou de constituintes do grupo verbal. Podem ser considerados advérbios de negação: não, tampouco, nem, nunca e jamais (NETO; INFANETE, 1998).

Existem também as locuções de palavras que funcionam como um advérbio de negação, também chamados de locuções adverbiais de negação, a saber: de modo algum, de jeito nenhum e de forma nenhuma (NETO; INFANETE, 1998).

Em circunstâncias onde ocorrem construções de negação frásica, a distribuição do advérbio é bastante restrita. Nestes casos, ocorrem sempre em posição de adjacência à esquerda do verbo, mesmo em construções interrogativas que envolvem a inversão do sujeito com o verbo. Como por exemplo: “Carlos (não) pulou o muro.” e “Quem (não) pulou o muro?” (NETO; INFANETE, 1998).

Já nos casos de construções de negação do grupo verbal, o advérbio pode modificar qualquer constituinte do grupo verbal. Neste caso, o advérbio de negação encontra-se adjacente ao constituinte modificado, como os exemplos abaixo:

- Modificando todo o grupo verbal

  O Carlos (não comprou um carro à Bruna ontem).

- Modificando o objeto direto

  O Carlos comprou à Bruna ontem (não um carro), mas flores.

- Modificando o advérbio de tempo

O Carlos comprou um carro à Bruna (não ontem), mas hoje.

- Modificando o objeto indireto

O Carlos comprou um carro ontem (não à Bruna), mas à Juliana.

No modelo descrito nesta dissertação, o tratamento da negação está ligado a identificação do advérbio de negação e qual a situação de modificação ocorre no texto. Identificada estas duas características no texto ótimo e candidato, suas palavras e informações morfossintáticas são armazenadas para posterior comparação. Esta comparação identifica se ambos os textos mencionam a mesma informação de forma negativa ou positiva.

4.2.5 Tratamento da Flexão de Voz Ativa/Passiva

A voz verbal indica se o ser a que o verbo se refere é paciente ou agente do processo verbal (NETO; INFANETE, 1998). Há três situações possíveis:

a) Voz ativa – o ser a que o verbo se refere é o agente do processo verbal. Em “Pedro descobriu o Brasil”, a forma verbal “descobriu” está na voz ativa porque “o Brasil” é o agente do processo verbal.

b) Voz passiva – o ser a que o verbo se refere é o paciente do processo verbal. Em “O Brasil foi descoberto por Pedro”, a locução verbal “foi descoberto” está na voz passiva porque “O Brasil” é o paciente da ação verbal.

c) Voz reflexiva – o ser a que o verbo se refere é, ao mesmo tempo, agente e paciente do processo verbal, pois age sobre si mesmo. Como no exemplo “O funcionário demitiu-se”, onde a forma verbal “demitiu-se” está na voz reflexiva, pois o funcionário ao mesmo tempo é agente e paciente: demitiu a si mesmo.

O trabalho proposto aqui focará na identificação da voz ativa e passiva, por possuírem uma estrutura possível de se identificar com as informações geradas pelo analisador morfossintático utilizado.

Para a constatação da voz verbal passiva se faz necessária apenas a identificação do agente da passiva, pois o mesmo só existe nesta flexão de voz. Um agente da passiva está localizado após uma sequência de verbo principal (ser), verbo no particípio e preposição (por,

pelo(s), pela(s), de). Salvo quando o sujeito da voz ativa for indeterminado, neste caso não haverá o agente da passiva (NETO; INFANETE, 1998).

Alguns outros elementos também podem ser identificados visando a alteração na flexão de voz, a saber:

a) Sujeito - substantivo antecedido ou não por um artigo;
b) Verbo Transitivo Direto - verbo que não é seguido por uma preposição;
c) Verbo Transitivo Indireto - verbo seguido por uma preposição;
d) Objeto Direto - localizado após o verbo transitivo direto e que pode ser antecedido ou não por um artigo;
e) Objeto Indireto - localizado após o verbo transitivo indireto e que é antecedido por uma preposição;
f) Locução Verbal - verbo no particípio antecedido por um verbo;

Segundo Neto e Infante (1998) o processo de alteração de um texto em voz ativa para a voz passiva pode ser iniciado pelo objeto direto que passa para o início da frase. Na sequência está o verbo transitivo direto, que passa para a sua forma no particípio e recebe um verbo auxiliar (ser/estar) o antecedendo, no tempo verbal do verbo transitivo direto, formando assim a locução verbal. Para finalizar, o sujeito da ativa passa para depois da locução verbal e é acrescentada entre os dois uma preposição de acordo com o verbo que o antecede.

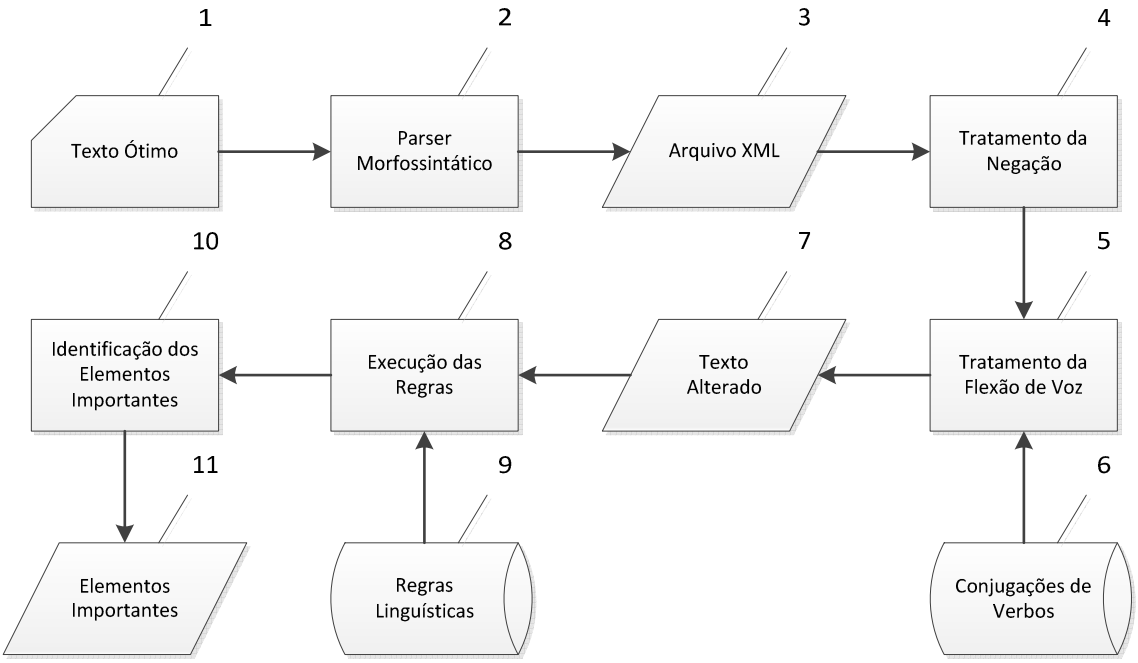
4.2.6 Tratamento de sinônimos

Para o tratamento de sinônimos é utilizado um repositório de informações provenientes do WordNet BR. Neste processo é verificado se o verbo apresentado no texto candidato é o mesmo, ou sinônimo, do apresentado no texto ótimo. Esse processo ocorre durante a avaliação do texto candidato, através de uma verificação no repositório de sinônimos. Caso os verbos apresentados tanto no texto ótimo quanto candidato pertencerem ao mesmo grupo de significância, os dois verbos são sinônimos.

4.2.7 Avaliação do texto ótimo

Neste trabalho denomina-se “texto ótimo” o texto fornecido pelo professor e contendo a resposta correta. A análise do texto ótimo segue o processo apresentado na Figura 9, que terá cada etapa explicada no que segue.

Figura 9: Processo de análise do texto ótimo



Fonte: Elaborada pelo autor.

Na etapa 1 deve ser recebido o texto ótimo, ou seja, a resposta do professor, que contém as informações importantes e que devem estar contidas no texto candidato. É preferível que este texto contenha a resposta correta de forma objetiva. Este texto segue então para a análise pelo *parser* morfossintático (etapa 2), que retorna um arquivo no formato XML (etapa 3) com as informações léxicas, sintáticas e morfológicas sobre cada palavra presente no texto e sobre a estrutura das sentenças.

Na etapa 4 ocorre o tratamento da negação, onde é verificado se há uma negação presente no texto e a qual verbo ela está associada. As informações deste verbo são então armazenadas para comparações futuras. Na etapa 5 ocorre o tratamento da flexão de voz (ativa ou passiva) utilizada no texto, onde todos os textos, caso não estejam na voz ativa, são modificados para a mesma, neste momento se faz uso também do repositório de verbos conjugados (etapa 6), que é consultado e retorna o verbo a ser alterado no tempo correto, e

que é colocado no texto ótimo, substituindo o verbo que estava na voz passiva. Os resultados destas alterações no texto ficam armazenados no modelo (etapa 7).

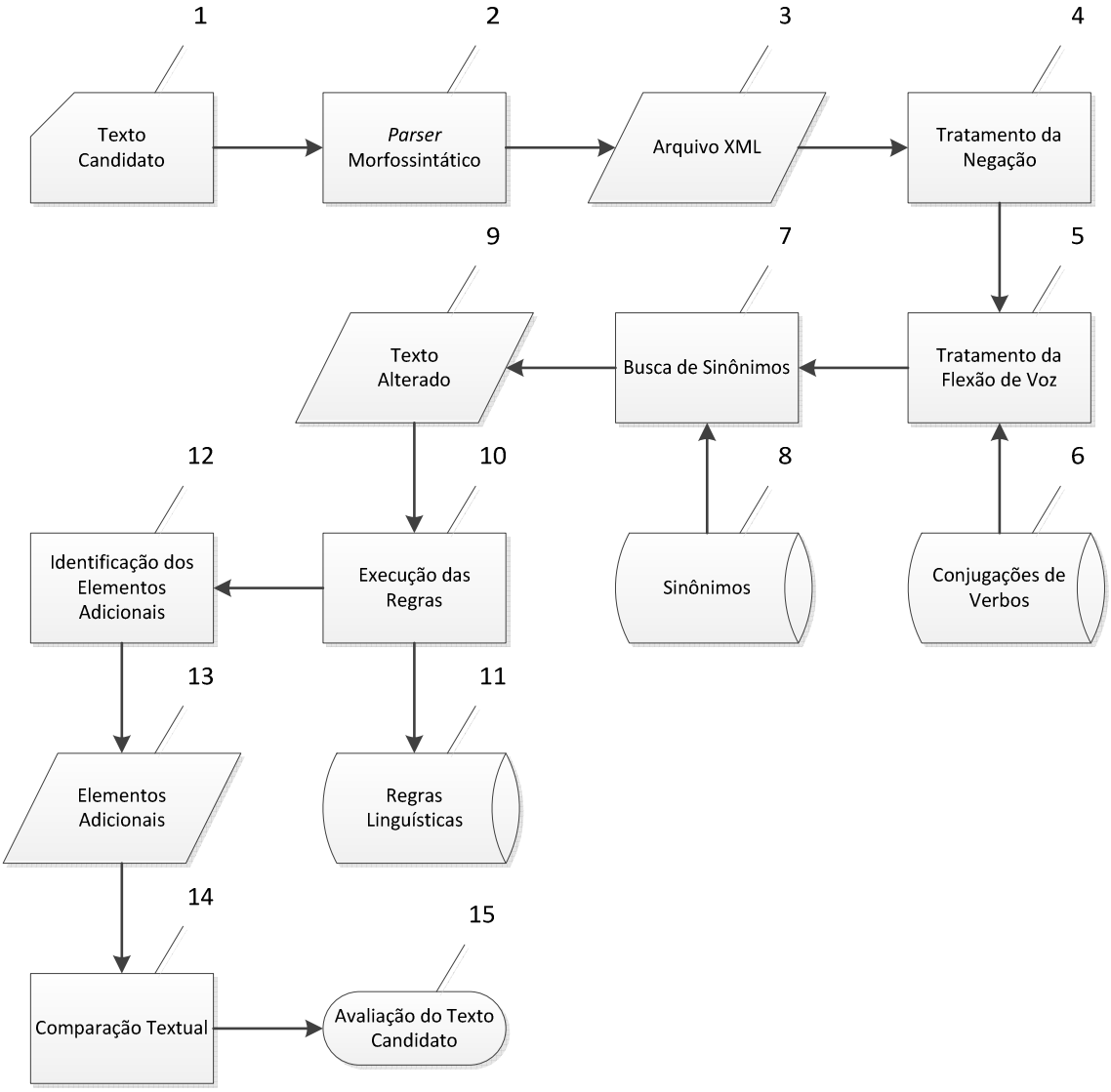
Durante a execução de regras (etapa 8), todas as regras linguísticas presentes no repositório de Regras Linguísticas (9) são avaliadas, identificando assim a regra que melhor abrange o texto que está sendo analisado. Após eleita a regra, a mesma é utilizada para subtrair do texto as informações desnecessárias, destacando assim os elementos importantes para a resposta (etapa 10). Elementos que são então armazenados, o que ocorre na etapa 11.

4.2.8 Análise do texto candidato

Análise do texto candidato ocorre de forma muito semelhante a do texto ótimo, porém estão presentes alguns processos adicionais onde determinadas informações do texto candidato são comparadas a do texto ótimo.

A Figura 10 apresenta todo o processo efetuado para a análise do texto candidato e cada etapa é explicada a seguir.

Figura 10: Processo de análise do texto candidato.



Fonte: Elaborada pelo autor.

As etapas de 1 a 3 apresentam exatamente as mesmas funções das realizadas na análise do texto ótimo. Na etapa 4 o processo de tratamento de negação do texto candidato também é o mesmo do texto ótimo, porém ao término da busca e identificação do verbo negado, esta informação é comparada a armazenada durante o tratamento da negação do texto ótimo. Caso ocorra a negação de um verbo que não está presente no texto candidato, ou vice-versa, o modelo indica o texto candidato como não vinculado ao texto ótimo. Caso isso não ocorra a análise segue normalmente.

A etapa 5 ocorre da mesma forma que a executada no texto ótimo, passando o texto para a voz ativa e alterando o verbo com base na informação encontrada no repositório de

verbos conjugados (etapa 6). Os verbos presentes no texto ótimo são então comparados ao do texto candidato (etapa 7), na mesma ordem em que aparecem no texto. Essa comparação é realizada com o objetivo de encontrar cada verbo do texto ótimo, representado diretamente ou através do mesmo grupo significância (etapa 8), dentro do texto candidato. Caso isso não ocorra, o modelo indica o texto candidato como não vinculado. Caso contrário, o texto em sua forma alterada é então armazenado pelo sistema (etapa 9).

Na etapa 10 todas as regras presentes no respectivo repositório (11) são avaliadas, para verificar qual delas melhor abrange o texto candidato. Após a regra linguística ser eleita, ela é utilizada para remover elementos do texto, restando assim apenas os elementos adicionais (etapa 12), e que são armazenados pelo modelo (etapa 13).

No passo seguinte os elementos importantes são buscados entre os elementos adicionais (etapa 14), o índice de elementos importantes encontrados entre os elementos adicionais, gera a avaliação sobre a vinculação textual entre os textos candidato e ótimo (etapa 15).

4.3 Exemplo com o resumo do funcionamento do modelo

Para a construção de um protótipo do ambiente proposto, nesta sessão serão apresentados alguns detalhes das etapas fundamentais e que realizam análises e avaliação de vinculação do texto candidato.

Para exemplificar cada processo, será considerada a pergunta hipotética: “Quais são as cores presentes no padrão CMYK?”. Como texto ótimo será considerada a frase: “São ciano, magenta, amarelo e preto” e como possíveis respostas, as presentes no quadro abaixo:

Quadro 1: Exemplo de pergunta e respostas

Pergunta hipotética: “Quais são as cores presentes no padrão CMYK?”
Texto ótimo: “São: ciano, magenta, amarelo e preto”
Possíveis respostas (textos candidatos):
a) “Preto, magenta, ciano e amarelo”;
b) “Magenta, amarelo e preto”;

c) “Preto”;

d) “Ciano, magenta, amarelo e preto, são as cores”;

e) “Preto, magenta, amarelo e ciano, são as cores presentes no padrão CMYK”;

f) “O padrão CMYK apresenta: preto, magenta, amarelo e ciano”.

Fonte: Elaborada pelo autor.

Os principais processos a serem executados, são apresentados no que segue.

A primeira atividade é a seleção das frases e análise pelo *parser* morfossintático. Após analisado pelo parser, o resultado é um arquivo XML contendo a anotação de cada palavra presente na amostra textual submetida, conforme suas informações morfossintáticas. O resultado pode ser observado na Figura 11.

As informações de interesse ao modelo estão presentes entre as *tags* “*terminals*”, sendo que o identificador *word* apresenta a palavra no estado em que estava presente no texto, *lemma* a palavra sem estar conjugada, *pos* a informação de parte no discurso e *morph* informações sobre o tempo e pessoa em que o verbo está conjugado.

Figura 11: Trecho de texto anotado com *parser* morfossintático.

```
<?xml version="1.0" encoding="UTF-8"?>
<xml>
<corpus>

    <body>
<s id="s1" ref="1" source="Running text" forest="1" text="São:">
    <graph root="s1_500">
        <terminals>
            <t id="s1_1" word="São" lemma="ser" pos="v-fin" morph="PR 3P IND VFIN" extra="fmc mv"/>
            <t id="s1_2" word=":" lemma="--" pos="pu" morph="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s1_500" cat="s">
                <edge label="UTT" idref="s1_501"/>
            </nt>
            <nt id="s1_501" cat="x">
                <edge label="STA" idref="s1_1"/>
                <edge label="PU" idref="s1_2"/>
            </nt>
        </nonterminals>
    </graph>
</s>

<s id="s2" ref="2" source="Running text" forest="1" text="ciano, magenta, amarelo e preto.">
    <graph root="s2_500">
        <terminals>
            <t id="s2_1" word="ciano" lemma="ciano" pos="adj" morph="M S" extra="col jh"/>
            <t id="s2_2" word="," lemma="--" pos="pu" morph="--" extra="--"/>
            <t id="s2_3" word="magenta" lemma="magenta" pos="n" morph="M S" extra="cjt-head cjt-head-X col"/>
            <t id="s2_4" word="," lemma="--" pos="pu" morph="--" extra="--"/>
            <t id="s2_5" word="amarelo" lemma="amarelo" pos="adj" morph="M S" extra="cjt-X col jh"/>
            <t id="s2_6" word="e" lemma="e" pos="conj" morph="--" extra="sam-"/>
            <t id="s2_7" word="preto" lemma="preto" pos="adj" morph="M S" extra="cjt-X col jh"/>
            <t id="s2_8" word="." lemma="--" pos="pu" morph="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s2_500" cat="s">
                <edge label="fCs" idref="s2_501"/>
            </nt>
            <nt id="s2_501" cat="adjp">
                <edge label="H" idref="s2_1"/>
                <edge label="PU" idref="s2_2"/>
                <edge label="X" idref="s2_502"/>
                <edge label="PU" idref="s2_7"/>
            </nt>
            <nt id="s2_502" cat="par">
                <edge label="CJT" idref="s2_3"/>
                <edge label="PU" idref="s2_4"/>
                <edge label="CJT" idref="s2_5"/>
                <edge label="CJT" idref="s2_6"/>
            </nt>
        </nonterminals>
    </graph>
</s>

    </body>
</corpus>
</xml>
```

Fonte: Elaborada pelo autor.

A segunda atividade comentada é o tratamento da flexão de voz. Esta validação é realizada através da extração das informações provenientes do analisador morfossintático e da verificação de ocorrência destas sequências de informações específicas para a identificação da voz passiva ou ativa. Utilizando a resposta do Quadro 1 identificada pela letra “e” como exemplo, neste texto encontra-se, depois do verbo “presentes“ o trecho “... no padrão CMYK”. A palavra “no” é formada pela preposição “em” e o artigo definido “o”, logo, por estar sendo antecedida pelo verbo e seguida do substantivo “padrão”, forma o agente da passiva.

Esta mesma frase poderia estar no formato: “No padrão CMYK estão presentes as cores: preto, magenta, amarelo e ciano” e ainda assim apresentaria o agente da passiva, caracterizando a voz passiva. Porém se a frase fosse: “O padrão CMYK apresenta as cores: preto, magenta, amarelo e ciano”, a ausência do agente da passiva caracterizaria a frase com portadora da voz ativa.

Em alguns casos, alunos e professores podem construir respostas que utilizam voz ativa ou voz passiva. Este fato prejudica a identificação dos elementos de interesse dentro do texto, e por este motivo se faz necessário o processo de detecção e alteração da estrutura do texto em alguns casos.

Podemos usar para este exemplo a resposta “f” do Quadro 1: “O padrão CMYK apresenta: preto, magenta, amarelo e ciano”. Após a identificação do objeto direto: “preto, magenta, amarelo e ciano”, este passa para o início da frase, seguido pela locução verbal, que é formada pelo verbo transitivo direto “apresenta”, antecedido por um verbo auxiliar; resultando em “são apresentados”. Finalmente o sujeito da ativa “O padrão CMYK”, recebe uma preposição que o antecede, neste caso “por”, resultando no agente da passiva “pelo padrão CMYK”, que é posicionado depois da locução verbal. Concluindo, esta frase “f” com a voz alterada para passiva passa a ser: “Preto, magenta, amarelo e ciano são apresentados pelo padrão CMYK”.

Para realizar o caminho contrário, alterando a voz passiva para voz ativa, podemos iniciar pelo agente da passiva, que perde a preposição e passa para o início da frase, se tornando o sujeito da ativa. No segundo passo a locução verbal perde o verbo auxiliar, cujo tempo verbal passa para o verbo transitivo direto e é posicionado após o sujeito da ativa. Para finalizar o sujeito da passiva passa a ser o objeto direto sem maiores alterações, e é posicionado após o verbo transitivo direto (NETO; INFANETE, 1998).

Utilizando a mesma frase anterior: “Preto, magenta, amarelo e ciano são apresentados pelo padrão CMYK”. Identificamos o agente da passiva “pelo padrão CMYK”, e após retirar a preposição obtemos o sujeito da ativa “o padrão CMYK”. Após é analisada a locução verbal “são apresentados”, onde o verbo no particípio recebe o tempo verbal do verbo auxiliar, que por sua vez é retirado da frase, restando apenas o verbo transitivo direto “apresenta”. Na última etapa, o sujeito da passiva: “preto, magenta, amarelo e ciano” passa para o objeto direto sem alterações. O resultado final deste processo é a frase “f” original: “O padrão CMYK apresenta preto, magenta, amarelo e ciano”.

O processo de execução das regras trata de identificar qual das regras presentes no banco de regras é capaz de representar o texto analisado. Para exemplificar consideraremos as regras presentes no quadro abaixo:

Quadro 2: Exemplos de regras linguísticas

I)	Artigo Definido” + “Substantivo” + “Verbo” + “*” + “Conjunção” + “*”;
II)	“Verbo” + “*” + “Conjunção” + “*”;
III)	“*” + “Conjunção” + “*” + “Verbo” + “Artigo” + “Substantivo” + “Preposição” + “Artigo” + “*”;
IV)	“*”;
V)	“*”+ “Conjunção” + “*”;
VI)	“*”+ “Preposição” + “Artigo”;
VII)	“*”+ “Conjunção” + “*”+ “Verbo” + “Artigo Definido” + “Substantivo”.

Fonte: Elaborada pelo autor.

Estas regras foram armazenadas em listas de acordo com o padrão utilizado no analisador PALAVRAS. O símbolo “*” representa o intervalo contendo as informações de interesse para a resposta.

O protótipo do sistema construído para a indicação da regra que melhor representa a frase está baseado em cálculo de uma pontuação numérica. Sendo assim cada regra precisa ser dividida em trechos sequenciais contendo as informações referentes a parte do discurso, ou seja, cada intervalo entre o elemento representado por “*” de cada regra deve ser representado separadamente e avaliada a sua ocorrência no texto analisado. A ocorrência de cada sequência acarreta na soma de 1 ponto para cada informação na sequência presente na frase. Por

exemplo, o Quadro 3 abaixo descreve os elementos e a sequência de operações realizadas se executarmos a regra “I” sobre a resposta “a” do Quadro 1.

Inicialmente obteremos a regra dividida em duas partes. A primeira parte seria composta pelos elementos: “Artigo Definido” + “Substantivo” + “Verbo”. A segunda parte seria composta pelo elemento: “Conjunção”. A primeira parte da regra é utilizada para uma busca de seus elementos nesta mesma sequência, ou seja, “Artigo Definido” + “Substantivo” + “Verbo”. Como esta sequência não é encontrada no texto, não há adição de pontos. Mas ao ser avaliada a segunda sequência, formada apenas pelo termo “Conjunção”, verifica-se a existência deste elemento na quarta posição da resposta. Como esta é a segunda sequência da regra, logo se sabe que há informação adicional entre as duas sequências. Portanto se houver a existência de qualquer informação anterior a conjunção encontrada na resposta, esta se refere ao símbolo “*” encontrado na regra, mesmo que a primeira sequência não esteja presente na resposta. Isso faz com que a ocorrência da informação anterior a conjunção some mais 1 ponto ao total desta regra, que apresenta um total de 2 no momento. A sequência de regras a ser verificada na resposta acaba, mas ainda há mais informações na mesma, pela existência de mais um “*” ao final da regra. Como ainda existem informações posteriores a conjunção na resposta, mais um ponto pode ser somado a esta regra, totalizando agora 3 pontos.

Quadro 3: Exemplo de aplicação das regras linguísticas

Regra: [“Artigo Definido” + “Substantivo” + “Verbo” + “*” + “Conjunção” + “*”]

Frase: [“Preto, magenta, ciano e amarelo”]

Análise morfológica da frase:[“Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”]

1ª sequência: [Artigo Definido” + “Substantivo” + “Verbo”]

Análise morfológica da frase:[“Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”]

Pontos = 0

2ª sequência: [“Conjunção”]

Análise morfológica da frase:[“Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”] (+ 1 ponto; + 1 ponto; + 1 ponto)

Pontos = 1+1+1 = 3

Fonte: Elaborada pelo autor.

Com este mesmo algoritmo, quando executada a regra “IV” do Quadro 2, nenhuma sequência será encontrada, resultando na resolução total da frase como compatibilidade com o “*” encontrado na regra, e somando 1 ponto a mesma.

Quando for avaliada a regra “V” do Quadro 2, será encontrada apenas uma sequência equivalente, contendo apenas a informação “Conjunção”, que pode ser encontrada na posição quatro da resposta, assim como ocorreu na avaliação da regra “I”, somando a esta regra 1 ponto. Como existem mais informações anteriores a conjunção e há também um símbolo “*” anterior a “Conjunção” na regra, mais 1 ponto é somado, totalizando 2 pontos. Assim não existem mais sequências a serem buscadas, mas existem mais informações na resposta e a existência de outro “*” após “Conjunção” na regra “V”, sendo assim mais um ponto é adicionado a regra, finalizando com um total de 3 pontos, armazenados junto a regra.

Após a execução de todas as regras, ocorre uma etapa de comparação entre as mesmas. Neste ponto a primeira regra verificada é armazenada como uma regra temporária, neste exemplo a regra “I”. Esta regra temporária é então comparada a segunda regra da seguinte forma: se o número de itens que compõem a regra temporária (pontos que podem ser realizados), menos os pontos somados pela mesma forem maiores ou iguais ao número de itens que compõem a segunda regra, menos os pontos somados pela mesma, e o número de total pontos armazenados pela regra temporária for menor ou igual ao número total de pontos armazenados pela segunda regra, então a segunda regra passa a ser a regra temporária. Sendo assim, os 6 itens que pertencem a regra “I” menos os 3 pontos somados pela mesma, resultam em 3, valor maior que o resultado da regra “IV”, onde 1 item que compõe a regra menos 1 ponto marcado pela mesma resultada em 0. Porém na segunda condição, o total de pontos somados pela regra temporária “I” são 3, valor maior que o da segunda regra “IV” que apresenta apenas 1 ponto, resultando na permanência da regra “I” como temporária.

Comparando esta regra temporária “I”, com a agora segunda regra “V”, obteremos o mesmo resultado na primeira condição (6-3=3), e para a segunda regra termos 3 elementos que a formam, menos os 3 pontos somados, resultando em 0, valor menor que a da regra temporária. Para a segunda condição temos os 3 pontos da regra temporária iguais aos 3 pontos da segunda regra. Sendo assim a regra “V” passa a ser a regra temporária, e, como não há mais regras para se comparar, esta regra temporária passa a ser a regra eleita, ou seja, a regra que melhor representa o texto analisado. Esta comparação pode ser visualizada no Quadro 4.

Quadro 4: Comparação de resultados de pontuação das regras

$[(I_{RT}\text{-}PS_{RT}) >= (I_{SR}\text{-}PS_{SR})]$ && $[(PS_{RT}) <= (PS_{SR})]$
1ª Comparação: [(6-3)>=(1-1)] && [(3)<=(1)]
Condições não encontradas, a regra temporária continua a mesma.
2ª Comparação: [(6-3)>=(3-3)] && [(3)<=(3)]
Condições encontradas, a segunda regra passa a ser a temporária.

Fonte: Elaborada pelo autor.

A comparação entre as regras foi construída com duas condições para priorizar duas informações de fundamental importância, a primeira se refere a qual fração da regra foi correlacionada ao texto analisado, e a segunda se refere ao quão variada é a informação descrita pela regra em questão. Isso faz com que a regra “IV”, por exemplo, tenha seu termo “*” presente no texto, ou seja, 1 termo na regra e 1 ponto somado, e mesmo assim em nenhum momento essa regra é considerada uma regra temporária, já que as outras duas regras utilizadas neste exemplo, abrangem mais informações sobre o texto analisado.

O processo de identificação de elementos ocorre depois da execução das regras e indexação da resposta ao texto ótimo. Embora ocorram da mesma maneira, os elementos identificados no texto ótimo e na resposta candidata recebem nomes distintos devido a sua proveniência, sendo os elementos retirados do texto ótimo, elementos importantes, e os da resposta candidata, elementos candidatos. Estas nomenclaturas são diferenciadas, pois os elementos importantes, presentes na resposta ótima, são os elementos que precisam estar presentes na lista dos elementos candidatos.

Para a identificação destes elementos a regra previamente indicada é fundamental, pois os elementos que indicam a parte no discurso, presentes na regra, são retirados do texto analisado, pois podem ser considerados como complementos de frase, ou seja, palavras utilizadas apenas para a construção de uma resposta.

Seguindo com o texto ótimo apresentado como exemplo no Quadro 1, que exibe as informações morfológicas: “Verbo” + “Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”, atribuímos a regra “II” do Quadro 2: “Verbo” + “*” +

“Conjunção” + “*”. Para a identificação dos elementos importantes, todas as informações presentes tanto na regra quanto no texto ótimo, são removidas. Estas informações se referem a verbo e conjunção, que representam no texto ótimo as palavras “São” e “e”, restando os quatro substantivos que representam as palavras: “ciano”, “magenta”, “amarelo” e “preto”. Estas informações restantes são os chamados elementos importantes, e são então armazenados para posterior comparação com os elementos candidatos que serão encontrados.

Para a análise do texto candidato “a” do Quadro 1, que apresenta as informações morfológicas: “Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”, foi indicada a regra “V” do Quadro 2: “*” + “Conjunção” + “*”. Ocorrendo então o mesmo processo de subtração das partes do discurso comuns entre regra e texto, retiramos a palavra “e” do texto candidato, restando apenas os quatro substantivos. Estes elementos restantes são armazenados como elementos candidatos, e serão cruzados com os elementos importantes para gerar a avaliação final. Este processo pode ser acompanhado no Quadro 5. Uma descrição detalhada de como esse processo pode ser implementado está presente no capítulo 5.

Quadro 5: Processo de identificação dos elementos

Texto ótimo: “São ciano, magenta, amarelo e preto”
Análise morfológica: “Verbo” + “Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”
Regra II: “Verbo” + “*” + “Conjunção” + “*”
Elementos importantes: “ciano” + “magenta” + “amarelo” + “preto”
Texto candidato: “Ciano, magenta, amarelo e preto”
Análise morfológica: “Substantivo” + “Substantivo” + “Substantivo” + “Conjunção” + “Substantivo”
Regra IV: “*” + “Conjunção” + “*”
Elementos candidatos: “ciano” + “magenta” + “amarelo” + “preto”

Fonte: Elaborada pelo autor.

Com apenas os elementos relevantes presentes tanto no texto ótimo como no candidato, agora se faz necessária a comparação entre os mesmos para indicar a

correspondência entre as amostras. Para realizar esta comparação são efetuados dois passos. No primeiro é realizada uma comparação quanto a informação morfológica apresentada nas duas amostras textuais. Quando é verificada a existência do mesmo dado nos dois textos, as palavras que originaram estas informações são armazenadas em suas respectivas listas, a que provém do texto ótimo é armazenada em uma lista de palavras importantes, e a que é derivada do texto candidato, em uma lista de palavras candidatas. No exemplo utilizado, todas as partes do discurso presentes nos elementos importantes também estão entre os elementos candidatos, resultando na adição de cada um deles a sua respectiva lista.

Para a segunda comparação, cada item presente na lista com os elementos importantes tem o seu lema (verbo no infinitivo, ou seja, sem nenhuma conjugação) comparado a cada item da lista contendo os elementos candidatos. Cada correlação pode ocorrer somente uma vez, e quando isto acontece, é somado 1 ponto ao total de pontos pertencentes ao texto candidato.

No exemplo que está sendo seguido, a lista de palavras importantes (“ciano” + “magenta” + “amarelo” + “preto”), contém exatamente o mesmo conteúdo da lista de palavras candidatas: “ciano” + “magenta” + “amarelo” + “preto”. Isto indica que a comparação textual encontrará cada uma das quatro palavras importantes entre as palavras candidatas, somando 1 ponto a cada palavra, e resultando e um total de 4 pontos.

Esta pontuação é dividida pelo número de elementos importantes identificados, neste caso 4 também. Este cálculo resulta em 1, e pode ser multiplicado por 100 para se obter o percentual de correspondência entre o texto candidato e o ótimo, neste caso 100%. O cálculo de correspondência textual pode ser apresentado de acordo com a Equação 1.

Equação 1: Correspondência entre textos

$$x = 100\left(\frac{\rho}{\beta}\right)$$

x: correspondência entre regra e texto

ρ: Pontuação total da regra (1)

β: Pontuação possível

Fonte: Elaborada pelo auto

Para avaliar uma vinculação textual que não apresente um resultado totalmente correto, podemos usar as respostas “b” e “c” do Quadro 1. Executando a verificação das

regras, encontraremos então a “V” para a frase “b” e “IV” para a frase “c”. No passo seguinte, quando retirados os complementos de frase, a “c” não sofrerá modificações, enquanto que a frase “d” perderá a conjunção “e”.

No exemplo a lista de palavras importantes possui: “ciano”, “magenta”, “amarelo” e “preto”, portanto teremos ao analisar a resposta “b” as palavras candidatas: “magenta”, “amarelo” e “preto”. Para a resposta “c” apenas “preto” como elemento candidato.

Sendo assim quando a comparação textual for realizada, teremos a existência de ¾ dos termos importantes no texto “b”, 75% de correspondência, e ¼ no texto “c”, equivalente a 25% de correspondência. Esta comparação pode ser acompanhada no Quadro 6.

Quadro 6: Comparação entre texto ótimo e frases “b” e “c” do Quadro 1.

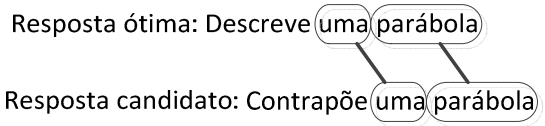
Lista de palavras importantes: “ciano” + “magenta” + “amarelo” + “preto”
Lista de palavras candidatos da frase “b”: “magenta” + “amarelo” + “preto”
Lista de palavras candidatos da frase “c”: “preto”
Xb = 100 (3/4)
Xc = 100 (1/4)

Fonte: Elaborada pelo autor.

Para realização da análise mais simples, a comparação textual estará somente baseada na análise léxica. Esta é baseada diretamente na informação referente a palavra em si, ou seja, o agrupamento das letras formando um conjunto único, sem avaliação de contexto, similaridade ou tempo verbal. Esta é uma análise simples de realizar, e pode ser efetiva em alguns casos. Porém quando ocorre a construção de uma ideia mais complexa para responder determinada pergunta, a análise léxica é incapaz de se adaptar a correlação de informações.

Por exemplo, se para determinada pergunta obtivermos a resposta ótima: “Descreve uma parábola”, a análise léxica realizará a busca unicamente por estas três palavras. Sendo assim, se obtendo como resposta candidata: “Contrapõe uma parábola”, a comparação que ocorre pode ser vista na Figura 12.

Figura 12: Análise léxica.



Fonte: Elaborada pelo autor.

Neste caso, ocorreu uma correlação de 2/3 palavras existentes no texto ótimo, sendo assim, de acordo com a análise léxica, o texto candidato apresenta 66% das informações colocadas pela resposta ótima. Porém podemos perceber que a resposta candidata traz junto ao verbo, um sentido para o texto totalmente diferente do apresentado pela resposta ótima.

5 IMPLEMENTAÇÃO

Para avaliar preliminarmente o funcionamento do modelo para identificação de vinculação textual utilizando regras linguísticas, descrito nesta dissertação, foi desenvolvido um protótipo em linguagem C# e que se vale das informações descritas no capitulo 4.

Este protótipo tem como objetivo principal avaliar um possível caso de desenvolvimento dos processos citados, realizando assim o tratamento de negação, de flexão voz, busca por sinônimos, execução das regras e comparação textual.

Para viabilizar a implementação e testes, neste protótipo foram utilizadas mensagens elaboradas por um especialista em linguística, presentes no tipo de pergunta Qual/Quais para extrair as regras linguísticas e realizar as primeiras avaliações sobre o funcionamento do protótipo, sendo assim o mesmo não se encontra ligado a nenhum AVEA. A consulta a sinônimos e conjugações de verbos também foi realizada em um repositório local, não ocorrendo assim a consulta diretamente aos do WordNet.

5.1 Visão geral da implementação do protótipo

Para realizar a análise de vinculação textual, o algoritmo do protótipo analisado segue a ordem de execução descrita no Quadro 7 sobre o texto ótimo já anotado pelo *parser* utilizado. A análise deste texto ótimo é realizada apenas uma vez para cada grupo de textos candidatos a ser analisado.

Quadro 7: Ordem de execução de ações sobre o texto ótimo.

Leitura arquivo XML
Tratamento da negação
Normalizador de voz
Tratamento de negação caso a frase estivesse na voz passiva
Execução das regras
Armazena elementos importantes
Escreve informações de saída

Fonte: Elaborada pelo autor.

Na etapa de “Leitura do arquivo XML” indicada no Quadro 7, o algoritmo faz a leitura inicialmente do texto ótimo e armazena as informações necessárias. Na sequência ocorre o “Tratamento de negação”. Com o objetivo de demonstrar e validar este processo, no protótipo desenvolvido para este trabalho, foi desenvolvido o tratamento de negação para os advérbios de negação que modificam o grupo verbal, ou seja, que estão ligados ao verbo principal.

O próximo passo é o “Normalizador de voz”, onde todos os textos que estiverem em voz passiva são alterados para a voz ativa. Na etapa de “Tratamento de negação caso a frase estivesse na voz passiva”, será avaliado se a voz do texto sendo analisado estiver descrita na voz passiva. Neste momento, já com a voz alterada, as informações do texto são armazenas pelo algoritmo.

Finalizadas todas as alterações no texto ótimo, ocorre a “Execução das regras”, onde todas as regras presentes no repositório de regras linguísticas são avaliadas e a regra que mais abrange as informações do texto sendo avaliado é eleita. Neste momento ocorre a identificação dos elementos. Para o texto ótimo são identificados e armazenados os elementos importantes. No passo final algumas informações são escritas em um arquivo de saída, para que possam ser avaliados os resultados obtidos pelo protótipo.

O Quadro 8 apresenta as informações referentes funções executadas sobre o texto candidato.

Para a análise dos textos candidatos os mesmos passos descritos no Quadro 7 são executados. Porém, após a normalização de voz, somente os textos candidatos passam pela “Busca de sinônimos”, na qual os verbos do texto candidato são pesquisados no mesmo grupo

de significância do texto ótimo. Para a implementação deste protótipo foi realizada uma consulta externa e suas informações foram armazenadas em um arquivo de texto, utilizado pelo protótipo para a consulta de sinônimos. Caso o verbo não seja encontrado no mesmo grupo de significância, o texto candidato é avaliado como não vinculado ao texto ótimo.

Depois de armazenadas as informações de negação, as do texto ótimo e do texto candidato são comparadas. Caso os dois textos apresentem verbos do mesmo grupo de significância e sejam ambos negados, ou ambos não apresentem verbos negados, então a análise continua. Caso contrário o texto candidato é considerado não vinculado ao texto ótimo.

Quadro 8: Ordem de execução de ações sobre o texto candidato

Leitura do arquivo XML
Tratamento de negação
Normalizador de voz
Busca de sinônimos
Tratamento de negação caso a frase estivesse na voz passiva
Execução das regras
Armazena elementos adicionais
Comparação textual
Escreve informações de saída

Fonte: Elaborada pelo autor.

A “Execução das regras” descrita no Quadro 8 ocorre sobre os textos candidatos da mesma forma que sobre o texto ótimo, destacando os elementos adicionais a serem armazenados pelo algoritmo. Estes elementos então são utilizados para a aplicação das métricas contidas na “Comparação textual”, onde cada palavra presente nos elementos importantes é pesquisada entre os elementos adicionais, resultando no percentual de vinculação textual do texto candidato.

Para finalizar esta análise algumas informações são gravadas em um arquivo, para que o funcionamento do protótipo seja verificado. Estes processos são efetuados sobre cada um dos textos candidatos indicados pelo usuário.

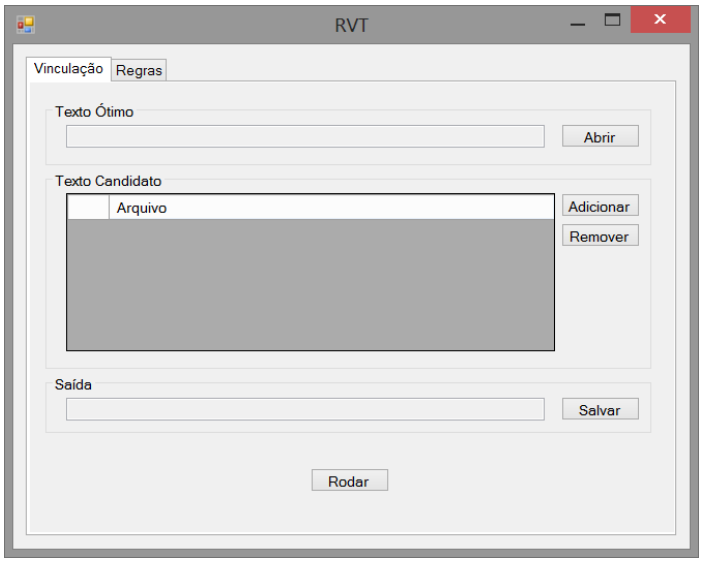
5.1.1 Interface

O protótipo desenvolvido possui três telas principais para o seu funcionamento e cada uma delas é apresentada no que segue.

A tela inicial, apresentada na Figura 13, é tela principal do programa. Ela possui duas abas, para separar as duas funções do protótipo. A aba inicial, “Vinculação”, realiza a análise de vinculação textual. Esta possui na parte superior da janela um campo identificado como “Texto Ótimo” para adição do texto ótimo em formato XML com as anotações realizadas pelo *parser* morfossintático utilizado.

No campo central da janela principal do protótipo, identificado como “Texto Candidato”, estão referenciados os arquivos que contem textos candidatos organizados em uma lista. Para adicionar mais arquivos a esta lista se faz necessário clicar no botão “Adicionar” localizado à direita deste campo e selecionar os arquivos desejados. Para remover, basta selecionar o arquivo desejado e clicar no botão “Remover” localizado também à direita do mesmo campo. O campo inferior, identificado como “Saída”, propicia ao usuário a escolha de pasta onde deseja salvar, e o nome do arquivo que será gerado pelo protótipo. Após completar estes campos o usuário pressiona o botão rodar, e quando a avaliação da vinculação textual estiver terminada o usuário recebe um aviso do fato ocorrido.

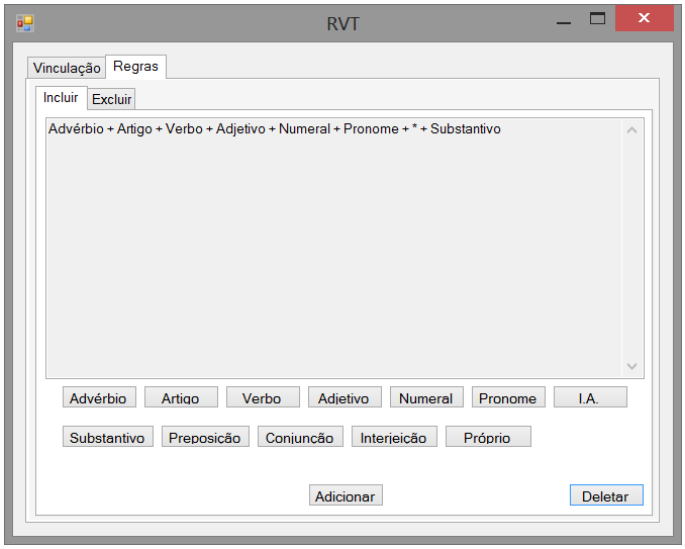
Figura 13: Tela inicial do protótipo.



Fonte: Elaborada pelo autor.

A segunda aba da tela inicial, “Regras”, permite realizar operações de inclusão e exclusão de regras sintáticas, sendo apresentada na Figura 14. A interface de inclusão de novas regras utiliza os botões de cada informação morfológica desejada, incluindo as regras cadastradas no repositório de regras sintáticas.

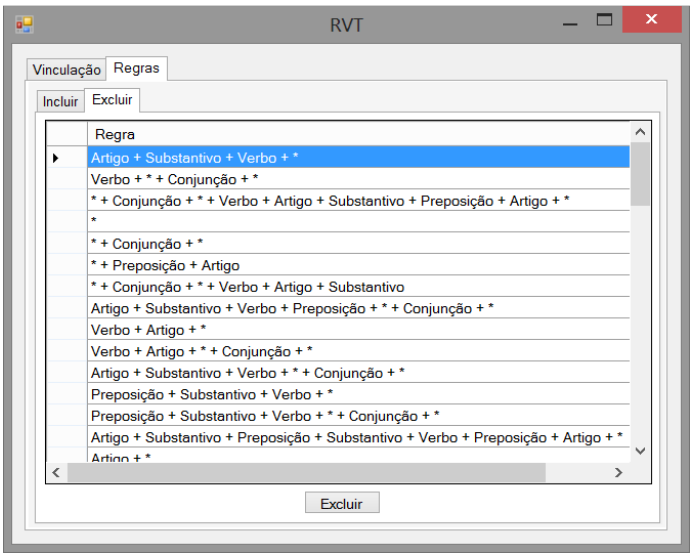
Figura 14: Tela para a inclusão de regras.



Fonte: Elaborada pelo autor.

A aba “Excluir” apresenta ao usuário as regras presentes no repositório local, como pode ser observado na Figura 15. Nesta aba é possível apagar do repositório qualquer uma das regras lá contidas. Para tanto, o usuário necessita clicar sobre a mesma e pressionar o botão “Excluir” posicionado na parte inferior da janela.

Figura 15: Tela para a exclusão de regras.



Fonte: Elaborada pelo autor.

5.1.2 Arquivo de saída

Para a avaliação do funcionamento do protótipo, a cada arquivo executado pelo algoritmo, algumas informações foram adicionadas a um arquivo no formato *Comma-Separated Values* (CSV), como pode ser visto na Tabela 3.

Tabela 3: Exemplo de resultado gerado pelo protótipo.

Texto ótimo:	A1.xml
Pontos possíveis:	4
Frase Original:	São: ciano, magenta, amarelo e preto.
Regra Eleita:	v- * conj *
Frase Alterada:	
Palavras Importantes:	Ciano magenta amarelo preto
POS:	adj n adj adj
Texto candidato:	A2.xml
Pontuação:	100%
Frase Original:	As cores padrão CMYK são: ciano, magenta, amarelo e preto.
Regra Eleita:	art n n prop v- * conj *
Frase Alterada:	
Palavras Candidatas:	Ciano magenta amarelo preto
POS:	adj n adj adj

Fonte: Elaborada pelo autor.

Nesta tabela estão contidas, de cima para baixo, iniciando pelo texto ótimo, o nome do arquivo identificado como texto ótimo; o número de pontos possíveis, ou seja, o número de palavras armazenadas em elementos importantes; a frase que compõe originalmente o texto

ótimo; a regra eleita dentre as presentes no repositório, que melhor representa esta amostra; a frase em sua forma alterada, caso isso ocorra; as palavras armazenadas como elementos importantes e por último as informações morfológicas destas palavras de acordo com a notação do *parser* Palavras (BICK, 2000).

Na sequência segue o nome do texto candidato indicado; a pontuação que o mesmo obteve, a frase em seu formato original; a regra eleita durante a execução das regras. A frase alterada, caso isso ocorra; as palavras armazenadas como elementos adicionais, e por último as informações morfológicas destas palavras.

5.2 Detalhamento do funcionamento interno do protótipo

A seguir serão detalhas as informações sobre o funcionamento interno do protótipo desenvolvido para avaliar o modelo descrito nesta dissertação.

5.2.1 Inclusão de Regras

Através de uma observação de informações morfológicas de padrões de frases construídas por alunos, um usuário, preferencialmente um especialista em linguística, pode montar a regra linguística utilizada na frase em questão para apresentar as palavras chaves e que constituem de fato a resposta. A Tabela 4 apresenta na coluna “Regra Linguística” a identificada por uma especialista em linguística a partir da observação das amostras textuais presentes na coluna “Frase”.

Tabela 4: Regras extraídas das frases.

	Frase	Regra Linguística
a)	"As cores do padrão CMYK são: ciano, magenta, amarelo e preto"	"Artigo" + "Substantivo" + "Preposição" + "*" + "Verbo" + "*" + "Conjunção" + "*"
b)	"São: ciano, magenta, amarelo e preto"	"Verbo" + "*" + "Conjunção" + "*"
c)	"Ciano, magenta, amarelo e preto"	"*" + "Conjunção" + "*"

d)	"Ciano, magenta, amarelo e preto são as cores do padrão CMYK"	"*" + "Verbo" + "Artigo" + "Substantivo" + "Preposição" + "*"
e)	“Ciano”	"*"
f)	“Magenta e Preto”	"*" + "Conjunção" + "*"

Fonte: Elaborada pelo autor.

Estas regras são então adicionadas ao repositório local, através da interface disponibilizada pelo protótipo.

5.2.2 Leitura do Arquivo XML

Foram desenvolvidas listas para o armazenamento das informações geradas pelo analisador morfossintático. Cada lista contém apenas um tipo de informação, mas de todas as palavras na frase. Com isso o mesmo índice representa, em cada lista, diferentes informações sobre a mesma palavra. Estas listas são chamadas de listas de terminais, pois se referem às informações presentes entre os elementos do tipo *terminals* nos arquivos de resultados da análise morfossintática.

A frase “A equação descreve uma parábola”, após analisada pelo *parser* morfossintático, tem suas informações organizadas de acordo com a Tabela 5. Neste exemplo a primeira coluna da tabela apresenta o nome da lista terminal, e as demais apresentam o conteúdo presente em cada, e armazenadas na mesma ordem em que um texto é lido, ou seja, o índice 0 na lista *word* contém o caracter “A”, e o índice 5 contém o caracter “.”.

Tabela 5: Listas de terminais.

Tipo de lista	Conteúdo das listas					
word	A	equação	descreve	uma	parábola	.
lemma	o	equação	descrever	um	parábola	--
pos	art	n	v-fin	art	n	pu
morph	F S	F S	PR 3s IND VFIN	F S	F S	--
extra	artd	cc-r	vH fmc mv	arti	ac-cat	--

Fonte: Elaborada pelo autor.

5.2.3 Tratamento da Negação

Para o tratamento da negação o algoritmo possui uma lista de advérbios de negação que podem modificar um grupo verbal. Advérbios são então pesquisados no texto que está sendo analisado. Se algum advérbio for encontrado na lista de “pos” do texto e o mesmo índice apresentar na lista “lemma” uma das palavras contidas na lista de advérbios de negação, então, há uma negação na amostra. O próximo passo é então descobrir se a negação se refere a um verbo.

Sendo assim, é verificada na posição seguinte a negação, se a informação de “pos” se refere a um verbo, caso o retorno seja negativo nada será feito, pois embora haja uma negação na amostra este protótipo trata apenas a negação de grupo verbal. Por isso caso a resposta seja positiva, o mesmo índice é buscado na lista de “lemma” e se a palavra retornada desta lista for uma conjugação do verbo ser, então o segundo índice após a negação é buscado na lista de “pos”. Esta busca é feita novamente por um verbo e caso o retorno seja positivo, então este é o verbo que está sendo negado e as suas informações das listas “pos”, “lemma”, “word” e “morph” serão armazenadas, caso o retorno seja negativo, o verbo que está sendo negado é justamente a conjugação do verbo ser, então será deste índice as informações a serem armazenadas.

5.2.4 Normalizador de Voz

No protótipo desenvolvido, todos os textos que estão na voz passiva são alterados para a voz ativa e para que isso ocorra é necessário identificar em qual flexão de voz se encontra o texto que está sendo analisado. Para rotular uma amostra com a flexão de voz passiva se buscou a presença de dois termos, o agente da passiva e a locução verbal.

Para identificar o agente da passiva é pesquisada na lista de “pos” do texto analisado uma preposição, caso encontrado, o mesmo índice na lista “lemma” é buscado, e a palavra presente neste é pesquisada em uma lista de preposições que podem compor um agente da passiva. Caso isso ocorra, é buscado na sequência do texto um substantivo, numeral ou nome próprio. Quando encontrado, deste índice até o índice da preposição, são considerados elementos do agente da passiva e seus índices são armazenados pelo algoritmo.

Para a identificação da locução verbal é realizado o mesmo processo, porém incialmente é pesquisado um verbo no particípio, caso encontrado é buscado nas posições anteriores a ele um verbo que seja a conjugação do verbo ser. Em caso positivo novamente, todos os índices entre estes dois elementos identificados, incluindo os mesmos são então armazenados como elementos da locução verbal. O índice em que se encontra o verbo ser possui informações importantes na lista “morph”, pois é ela que indica em qual pessoa e tempo o verbo está conjugado, e é nesta mesma conjugação que o verbo que atualmente está no particípio, deve ser conjugado, e por isto estas informações também são armazenadas, juntamente com a palavra presente na lista “lemma” do verbo que está no particípio. Este grupo de informações é identificado como verbo a ser conjugado. Caso haja uma negação na posição anterior ao verbo ser, este também é armazenado junto aos elementos da locução verbal.

Identificados os elementos de locução verbal e agente da passiva, os índices dos mesmos são reconfigurados de modo a tornar a frase ativa, ou seja, o agente da passiva passa para o início da frase, perdendo a preposição e se tornando o sujeito da ativa. A locução verbal perde o verbo auxiliar ser, e as informações identificadas como verbo a ser conjugado são então utilizadas para pesquisar no repositório de verbos conjugados, retornando assim o verbo principal no tempo correto para a voz ativa. As informações deste novo verbo substituem em todas as listas, as informações do que estava no particípio.

5.2.5 Busca de sinônimos

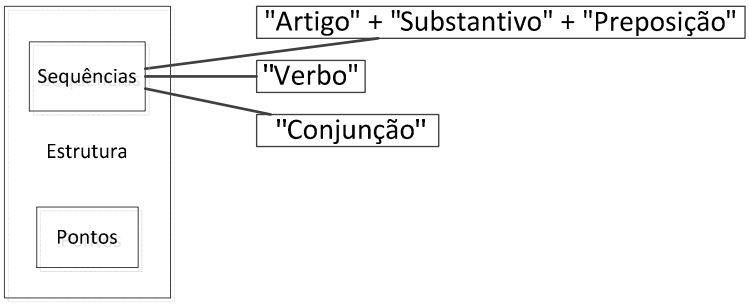
Neste ponto do protótipo os verbos presentes no texto ótimo são pesquisados de forma direta no texto candidato, ou seja, os índices que apresentarem na lista “pos” a informação verbo, terão o mesmo índice na lista “word” pesquisado na lista “word” do texto candidato. Essa pesquisa ocorre para todos os verbos presentes no texto ótimo. Caso não seja encontrado de forma idêntica, cada verbo do texto ótimo é pesquisado no repositório de sinônimos e com base do seu grupo de significância, o algoritmo busca cada um dos presentes no texto candidato dentro do mesmo grupo de significância. Caso algum verbo do texto ótimo não for encontrado, o texto candidato é avaliado como não vinculado ao texto ótimo. Porém, caso encontrado, o verbo presente no texto ótimo tem as suas informações copiadas em todas as listas, e estas então substituem as informações associadas ao texto candidato.

Estando a amostra textual na voz passiva, neste momento o tratamento da negação é novamente acionado, e as informações deste verbo conjugado corretamente são armazenadas pelo algoritmo.

5.2.6 Execução das Regras

Após executadas as alterações nas amostras, a primeira regra do arquivo de texto é pesquisada e armazenada em outra lista. Esta é então lida e repassada para outra estrutura. Esta estrutura contém listas e cada uma apresenta a sequência de informações morfológicas, entre os elementos importantes, identificados com o caractere "*", presentes nas regras. Essa quebra em sequências ocorre como, por exemplo, com a frase "a", onde antes do primeiro elemento "*" temos "Artigo" + "Substantivo" + "Preposição". Esta sequência é armazenada em uma lista. Logo depois temos "Verbo", que é colocado em uma segunda lista, e por último "Conjunção", acrescentado esta em uma última lista. O fato de cada sequência estar armazenada em uma lista diferente indica ao protótipo que entre elas existe um elemento do tipo "*", ou seja, uma informação a ser analisada. O protótipo também realiza a busca de dados antes de encontrar a primeira sequência, e após encontrar a última. Nesta mesma estrutura também são armazenados os pontos que a regra irá receber quando for avaliada. Na Figura 16 pode ser observada a forma como esta estrutura foi elaborada para armazenar as informações referentes a cada regra.

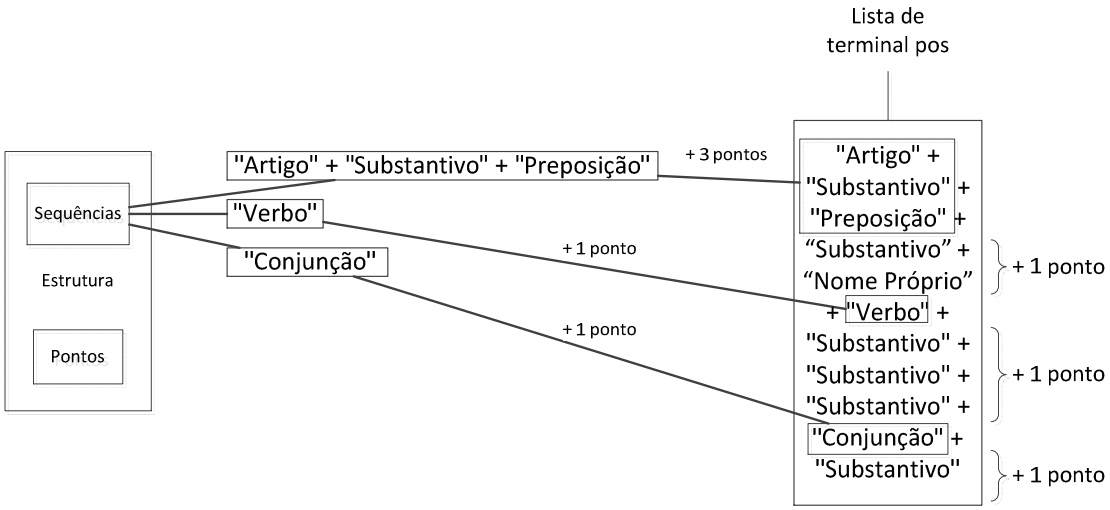
Figura 16: Estrutura e armazenamento da regra em sequências.



Fonte: Elaborada pelo autor.

Estas sequências são pesquisadas dentro da lista “pos”. E quando cada sequência é encontrada, a pontuação da regra é acrescida de 1 ponto a cada elemento presente na regra. O mesmo ocorre com os "*", caso sejam encontradas informações entre os elementos que compõem as sequências, estas são traduzidas como adicionais e somam também 1 ponto a regra. Este processo de busca e ponderação pode ser visualizado na Figura 17.

Figura 17: Busca e ponderação.



Fonte: Elaborada pelo autor.

A busca de sequência é iniciada pelo primeiro item da lista. No caso de haver a sequência "Artigo" + "Substantivo" + "Preposição", o primeiro elemento, "Artigo", é buscado dentro da lista “pos”. Quando encontrado, o segundo item da sequência passa a ser buscado, neste caso o "Substantivo", este elemento deve estar logo após a posição do "Artigo". Caso localizado, o próximo elemento da sequência é procurado logo após o "Substantivo". Se, e somente se, todos os elementos da lista de sequência forem encontrados na mesma ordem dentro da amostra textual é que a regra recebe a adição da quantidade de pontos correspondentes a cada elemento presente na sequência.

Ao final do processo de ponderação das regras, todas estas são comparadas com o intuito de se selecionar apenas uma para representar todo o conteúdo da frase.

5.2.7 Identificação dos Elementos

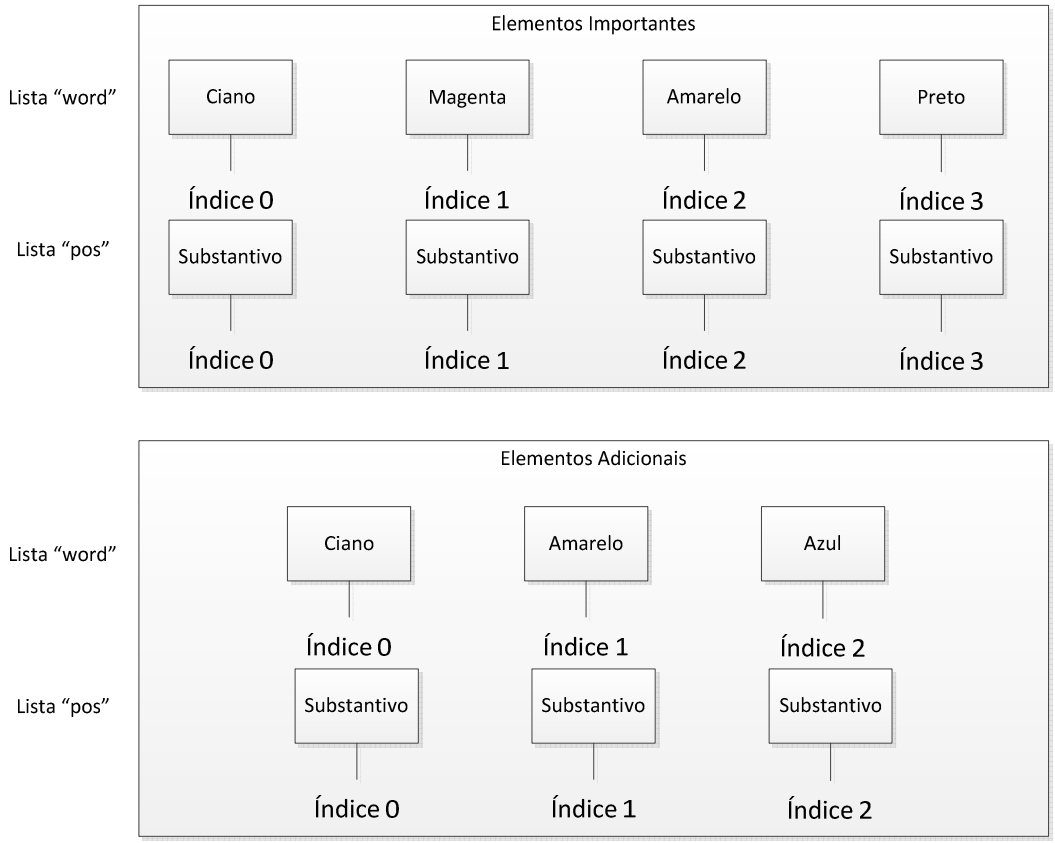
Após a eleição da regra, esta é utilizada para destacar os elementos de interesse. Estes elementos são armazenados em grupos distintos de acordo com o texto proveniente, sendo que os elementos provenientes do texto ótimo são chamados de elementos importantes, e os do texto candidato, elementos adicionais.

5.2.8 Comparação Textual

Com base nas listas de “word” e “pos” dos elementos importantes e adicionais (Figura 18), nesta etapa do algoritmo cada elemento importante tem sua informação na lista “pos” identificada dentro da lista “pos” dos elementos adicionais. O índice dos elementos importantes encontrados são então armazenados em uma lista paralela a lista de índices dos elementos adicionais.

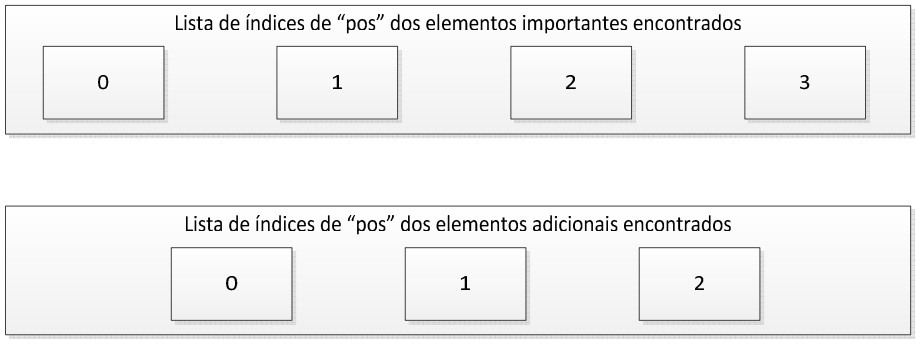
Após toda a lista “pos” de elementos importantes ser identificada dentre as do elementos adicionais (Figura 19), os índices dos elementos importantes encontrados e armazenados, tem sua respectiva posição na lista “word” buscada no índice indicado na lista paralela de elementos adicionais, também na lista “word”. Os índices encontrados em ambas as listas com informações iguais, são armazenados em uma última lista (Figura 20). Esta lista possui portanto, todos os índices da lista de elementos adicionais que estão corretos.

Figura 18: Listas de índices para elementos importantes e adicionais.



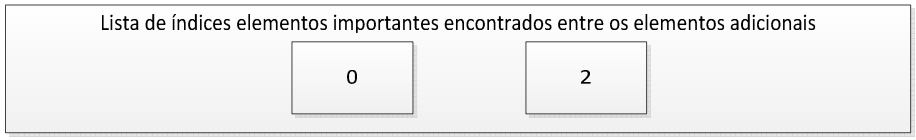
Fonte: Elaborada pelo autor.

Figura 19: Listas de índices das listas de "pos" para elementos importantes e adicionais.



Fonte: Elaborada pelo autor.

Figura 20: Listas de índices de elementos importantes encontrados entre os adicionais.



Fonte: Elaborada pelo autor.

Para realizar o cálculo de percentual de vinculação textual, o número de índices presentes na lista representada pela Figura 20 é dividido pelo número de índices presentes na lista de elementos importantes. No exemplo criado pelas figuras acima, o percentual de vinculação textual seria de 50%.

6 EXPERIMENTOS REALIZADOS

Neste capitulo serão apresentadas informações referentes às avaliações realizadas para verificar o funcionamento do protótipo desenvolvido, baseado no modelo apresentado nesta dissertação. As avaliações foram delimitadas com dois propósitos distintos: inicialmente para verificar a correção do protótipo dentro de um cenário controlado e a segunda para avaliar a capacidade de generalização do modelo, tratando de questões obtidas com usuários.

6.1 Estudo de caso inicial

No estudo de caso inicial houve a participação de um especialista em linguística. Este especialista desenvolveu trinta questões criadas especificamente para esta avaliação. Destas trinta questões, vinte e duas no formato “Qual/Quais”, cinco no formato “Quem”, duas no formato “Quando” e duas no formato “De que”. Todas as informações referentes ao estudo de caso inicial podem ser acompanhadas no Anexo B.

Junto a estas questões também foram desenvolvidas uma resposta ótima e quatro respostas candidatas para cada pergunta, totalizando cento e vinte grupos de respostas a serem avaliadas pelo protótipo. Estas respostas também foram observadas pelo especialista em linguística para a construção das regras linguísticas a serem cadastradas no protótipo, para ficarem disponíveis junto ao repositório de regras linguísticas. As regras presentes no repositório durante o estudo de caso inicial, podem ser observadas no Anexo C.

O especialista em linguística também indicou para cada resposta candidata uma nota em percentual, para possibilitar uma posterior comparação com o resultado gerado com o uso do protótipo. O critério usado pelo especialista para indicar a nota de cada resposta foi definido como sendo a medida dos elementos contidos na resposta para responder corretamente a pergunta. Como o grupo de perguntas tratado é voltado para situações objetivas e para respostas curtas, este critério é traduzido de forma bastante precisa e representa a quantidade de elementos corretos existentes em cada resposta.

A Tabela 6 indica um exemplo de um grupo de pergunta e respostas, com a anotação do especialista para cada uma das respostas.

Tabela 6: Exemplo de um grupo de pergunta e respostas.

		Avaliação do especialista (%)	Avaliação do protótipo (%)
Pergunta:	*Quais cores apresenta o padrão CMYK?*		
Resposta ótima:	São: ciano, magenta, amarelo e preto.	100	-
Respostas candidatas:	As cores padrão CMYK são: ciano, magenta, amarelo e preto.	100	100
	São as cores amarelo e preto.	50	50
	As cores são verde, vermelho e azul.	0	0
	São ciano e vermelho.	25	25

Fonte: Elaborada pelo autor.

A métrica de avaliação utilizada neste caso, para comparar as respostas anotadas pelo especialista manualmente com as respostas anotadas pelo protótipo foi a métrica de precisão (P), que é calculada dividindo-se o número de avaliações geradas pelo protótipo que são corretas e idênticas às do especialista (AC) pela soma deste número AC com o número de avaliações geradas pelo protótipo que são incorretas e diferentes das avaliações do especialista (AI) como pode ser observado Equação 2.

Equação 2: Cálculo da precisão

$$P = \frac{AC}{AC+AI} \tag{2}$$

O protótipo elaborado para avaliar o modelo proposto foi desenvolvido para questões do tipo “Qual/Quais” e “Quem”. Por este motivo analisaremos cada um dos resultados separadamente.

Das oitenta e oito respostas candidatas sobre a forma de pergunta “Qual/Quais”, o protótipo não apresentou exatamente o mesmo percentual que a avaliação do especialista em apenas três casos, resultando em uma precisão de 96,6%.

Das perguntas do tipo “Quem”, duas das vinte respostas candidatas não foram avaliadas da forma como indicada pelo especialista, resultando em uma precisão de 90%.

Depois de avaliado o sistema nos moldes das perguntas “Qual/Quais” e “Quem”, foram inseridos dois tipos de perguntas extras, apenas para avaliar o comportamento do protótipo. A pergunta do tipo “Que” teve suas respostas avaliadas corretamente em 100% dos

casos. Já as respostas do tipo “Quando” não puderam ser avaliadas pelo protótipo, isso porque o *software* Palavras não identificou as datas presentes nas amostras textuais, impossibilitando qualquer tipo de análise.

Sendo assim, do total de pares de respostas possíveis de serem analisadas pelo protótipo (cento e vinte), cento e duas puderam ser avaliadas pelo protótipo, e sendo que cinco análises não apresentaram exatamente o mesmo percentual de avaliação definido pelo especialista em linguística, o protótipo apresentou uma precisão de 95,7% de acerto no estudo de caso inicial.

Os resultados individuais e em sua totalidade podem ser observados na Tabela 7.

Tabela 7: Precisão por categoria.

Tipo de questão	**Total de respostas**	**Corretas**	**Incorretas**	**Precisão %**
Qual/Quais	88	85	3	96,6
Quem	20	18	2	90
De que	8	8	0	100
Total	**116**	**111**	**5**	**95,7**

Fonte: Elaborada pelo autor.

6.2 Segundo estudo de caso

Para o segundo estudo de caso foram criadas onze perguntas de estilos diferentes, sendo quatro relacionadas ao padrão “Qual/Quais”, quatro ao “Quem”, duas ao “Que” e a última ao “Como”. Estas perguntas foram apresentadas a dezoito pessoas selecionadas de forma aleatória, que as responderam de forma livre e anônima, ou seja, não havia a real necessidade de responder corretamente a questão ou mesmo construir uma resposta de forma coerente. Este segundo estudo de caso gerou então 198 respostas candidatas, que foram analisadas pelo especialista em linguística, e que indicou a cada resposta um percentual de acerto, conforme a resposta ótima. Estas mesmas respostas candidatas analisadas pelo especialista foram então utilizadas pelo protótipo e comparadas a resposta ótima, extraído assim o seu grau de vinculação textual. Os dados presentes neste segundo estudo de caso podem ser acompanhados no Anexo D. É importante salientar também que nenhuma regra linguística nova foi adicionada ao protótipo entre o primeiro e o segundo estudo de caso.

O protótipo foi desenvolvido para ser capaz de realizar a análise de vinculação textual para perguntas do tipo “Qual/Quais” e “Quem”. Foram adicionadas perguntas do tipo “De que” e “Como” apenas para analisar o comportamento do protótipo, mediante este tipo de resposta e utilizando as mesmas regras dos tipos de perguntas anteriores.

Para calcular a precisão do protótipo neste segundo estudo de caso foi utilizado também a Equação 2, porém o *parser* morfossintático utilizado não foi capaz de analisar os números presentes nas respostas da pergunta tipo “Como”, causando assim um erro como o visto no primeiro estudo de caso e inviabilizando a sua análise. Os resultados provenientes da análise das demais perguntas podem ser acompanhados na Tabela 8, e cada resultado será aprofundado no que segue.

Tabela 8: Precisão por categoria

Tipo de questão	**Total de respostas**	**Corretas**	**Incorretas**	**Precisão %**
Qual/Quais	72	72	0	100
Quem	72	57	15	79,16
De que	36	35	1	97,22
Total	**180**	**164**	**16**	**92,12**

Fonte: Elaborada pelo autor.

As quatro perguntas da categoria “Qual/Quais” apresentaram setenta e duas respostas candidatas, e que quando comparadas pelo protótipo a resposta ótima, apresentaram o mesmo percentual que o especialista indicou em seu estudo, sendo assim, nenhuma análise foi indicada como incorreta, e a precisão do algoritmo alcançou 100%.

As quatro perguntas da categoria “Quem” apresentaram também setenta e duas respostas candidatas, mas quinze delas obtiveram um percentual de vinculação textual diferente do indicado pelo especialista. Sendo assim, o protótipo atingiu neste tipo de pergunta uma precisão de 79,16%.

Por último foram analisadas as duas perguntas do tipo “De que”, que resultaram em trinta e seis respostas candidatas. Estas respostas, após avaliadas pelo protótipo, não apresentaram o mesmo percentual de vinculação textual que o especialista em linguística em apenas um dos casos. A avaliação destes resultados pode ser acompanhada no item seguinte.

6.3 Avaliação

O primeiro estudo de caso foi desenvolvido com o auxílio de um especialista em linguísticas, que criou as questões de acordo com os seus conhecimentos de padrões de respostas dadas pelos alunos em questionários. Destas mesmas respostas foram extraídas regras linguísticas que compuseram o repositório necessário para avaliar as respostas dadas pelos alunos.

Alguns dos casos onde houve divergência entre as avaliações do especialista e do algoritmo pelo protótipo, ocorreram por situações específicas, sendo que as mesmas serão analisadas no que segue.

Nos três casos relacionados ao padrão “Qual/Quais”, um está presente na pergunta 22 do Anexo B, e ocorreu pois o *software* Palavras rotulou a palavra “Terra” como substantivo, e não como nome próprio como era esperado. Nos outros dois casos, perguntas 15 e 28 do Anexo B, foram identificados corretamente os elementos importantes para a resposta, mas como o algoritmo conta cada palavra como um elemento independente, e o especialista considera o grupo de palavras uma informação independente, embora o funcionamento do protótipo esteja correto, o percentual de avaliação de ambos não é o mesmo.

Já os dois fatos relacionados ao padrão “Quem”, ocorreram nas perguntas 12 e 19 do Anexo B pelo mesmo motivo, o *software* Palavras identificou o par de palavras na resposta ótima como um nome próprio (exemplo “Charles Darwin”), e depois o algoritmo buscou esta informação por completa entre as respostas candidatas, e como não a encontrou, não inferiu como correta a resposta.

Sendo assim o funcionamento do protótipo ocorreu dentro do esperado, apresentando um grau de precisão total de 95,7%, que mesmo gerado dentro de um ambiente controlado do estudo de caso inicial, comprova o funcionamento efetivo das regras linguísticas como recurso viável para a construção de ferramentas de análise de vinculação textual, na língua portuguesa brasileira.

O segundo estudo de caso também foi trabalhado com o auxílio de um especialista, que foi responsável por definir o percentual de vinculação textual de cada resposta dada quando comparada a resposta ótima. Algumas destas respostas quando avaliadas pelo

protótipo, receberam um grau de vinculação textual diferente do definido pelo especialista, e cada uma destas situações serão analisadas no que segue.

Todos os quinze fatos relacionados ao tipo de pergunta “Quem”, ocorreram pelo mesmo motivo do relatado no primeiro estudo de caso, ou seja, o *parser* morfossintático uniu todas as palavras referentes ao nome, fazendo com que o protótipo não encontrasse a pessoa mencionada na frase. Assim como no primeiro estudo, caso houvesse uma forma de apresentar os diversos nomes possíveis para se identificar o sujeito de interesse, este tipo de pergunta resultaria em 100% de precisão.

Já a resposta da pergunta “Que” não apresentou o mesmo grau de vinculação textual e apresenta informações importantes a serem estudadas. Um dos fatos que acarreta no erro da avaliação é ausência de uma regra capaz de destacar a informação “sinalizar” na resposta ótima. Só esta ação, no entanto, ainda não torna o protótipo capaz de realizar está análise com precisão, já que a palavra importante na resposta candidata é “sinalizações”. As duas palavras, no entanto, apresentam na lista de “lemma” a palavra “sinalizar”, o que indica que em algumas situações o algoritmo deve realizar a comparação textual baseado nas informações contidas nesta lista.

Assim como no primeiro estudo de caso, o funcionamento do protótipo ocorreu dentro do esperado, apresentando um grau de precisão total de 92,12%, agora trabalhado com respostas não controladas pelo especialista e que apresentaram comportamentos distintos entre si, o que fortalece o modelo proposto nesta dissertação, e por este motivo reforçando o funcionamento efetivo das regras linguísticas como recurso viável para a construção de ferramentas de análise de vinculação textual, na língua portuguesa brasileira.

7 CONSIDERAÇÕES FINAIS

Neste capítulo são descritas as conclusões do trabalho e apontadas as contribuições do mesmo, juntamente com trabalhos futuros.

7.1 Conclusões

Esta dissertação apresenta um modelo que busca contribuir para a área de identificação de vinculação textual, mais especificamente a área de *Question Answering* e que apresenta como principais componentes a utilização de regras linguísticas em combinação com um *parser* morfossintático, identificação de voz das frases, negação e busca de sinônimos. Para a avaliação deste modelo, foi desenvolvido um protótipo com todas as etapas apresentadas para avaliar respostas para perguntas dos tipos Qual/Quais e Quem. Na sequência da pesquisa, foram realizados dois testes para avaliação dos resultados.

O primeiro teste realizado sobre uma base desenvolvida propositalmente para ser avaliada pelo protótipo, resultando em um nível de precisão alto e também em algumas informações específicas inesperadas como a anotação inconsistente do *parser* morfossintático utilizado e a necessidade de se ter uma espécie de repositório de significância para nomes próprios ou termos específicos. Este último evitaria que na avaliação de respostas contendo nomes próprios, por exemplo, o algoritmo possa compreender quais são as formas possíveis de se referir a aquela pessoa em específico, seja pelo primeiro, último ou todos os nomes de uma pessoa, e ainda assim ser avaliado como correta a sua resposta.

A primeira avaliação também levanta uma questão importante sobre o método utilizado para ponderar os elementos importantes. O fato do percentual de vinculação textual ser calculado com base no número de elementos importantes e não no número de conhecimentos importantes, faz com que a avaliação, embora correta, fique restrita a uma avaliação próxima da avaliação léxica. No exemplo do texto ótimo “São minerais metálicos e minerais não metálicos”, temos dois elementos de conhecimento: “minerais metálicos” e “minerais não metálicos”, e não apenas cinco palavras desligadas entre si. Sendo assim, na análise atual caso a resposta candidata contenha apenas “minerais metálicos”, a análise de vinculação textual resultará em 40%, e caso ela contenha “minerais não metálicos” o grau de

vinculação será de 60%. Caso alterado o método de ponderação de elementos, a avaliação deve ser 50% de vinculação após analisar qualquer uma destas duas respostas candidatas.

O segundo teste confirmou os resultados observados no primeiro teste, apresentando uma boa precisão geral e também a mesma flexibilidade para o seu uso em diversos contextos, além de indicar que o uso das regras linguísticas pode ser considerado em diversos contextos, de forma efetiva.

Visto que há ocorrências onde tanto o *parser* morfossintático utilizado quanto o modelo apresentam ainda pontos a serem trabalhados, é notória a precisão alcançada pela ferramenta em ambos os testes, demonstrando assim que o modelo descrito nesta dissertação atinge o seu objetivo e responde a questão de pesquisa apresentada nesta dissertação pois é possível realizar a identificação correta e flexível de respostas breves para perguntas sobre autoria e composição através de um modelo para reconhecimento de vinculação textual com base no conjunto de informações sintáticas, morfológicas, regras linguísticas e sinônimos.

7.2 Contribuições

Como pode ser verificado nas referências citadas ao longo deste documento, a área de identificação de vinculação textual está sendo amplamente explorada ao longo dos anos. O trabalho proposto neste documento apresenta uma contribuição quanto ao seu foco, pois poucos trabalhos desenvolvidos nessa área podem ser adaptados para a língua portuguesa.

Foi demonstrado que este modelo proposto com enfoque no uso das regras linguísticas, relativamente pouco exploradas pelos pesquisadores na área, pode viabilizar a construção de um sistema de vinculação textual voltado para a língua portuguesa, que possua real valia no processo de auxílio ao corpo docente que precisa acompanhar grandes quantidades de dados textuais.

O uso da detecção e tratamento da flexão de voz, como proposto, também é um diferencial na área, pois poucos trabalhos pesquisados apresentam tal preocupação.

A necessidade de um grande número de regras linguísticas pode parecer no primeiro momento como um ponto fraco no projeto, porém como visto no segundo estudo de caso, mantendo as mesmas regras, foi possível identificar com precisão a vinculação textual entre

textos provenientes de fontes aleatórias. Este fato corrobora a validação do modelo em que este documento está focado.

Assim sendo, considera-se que o Modelo de reconhecimento de vinculação textual baseado em regras linguísticas e informações morfossintáticas voltado para ambientes virtuais de ensino e aprendizagem, possa beneficiar o cenário atual onde grandes volumes de dados textuais são gerados através da crescente utilização das modalidades de Educação a Distância e Educação Presencial, MOOCs e AVEAs em geral.

Uma das linhas de trabalho desenvolvidas durante este período foi submetida ao Simpósio Brasileiro de Informática na Educação (SBIE) e aceita para a publicação em Novembro de 2013. O assunto descrito no artigo “O papel do Processamento de Linguagem Natural e da Representação de Conhecimento na extração de informações em mensagens textuais na Educação a Distância”, embora não esteja plenamente ligado ao apresentado nesta proposta, foi o responsável por promover experiência e vários conhecimentos ao autor, acerca de como trabalhar com dados textuais e fóruns utilizados em EaD. A referência do artigo consta em anexo.

Em Abril de 2014 foi submetido ao Centro Latino-americano de Estudos em Informática (CLEI) o assunto presente nesta dissertação. O artigo de nome “Modelo de reconhecimento de vinculação textual baseado em regras linguísticas e informações morfossintáticas voltado para ambientes virtuais de ensino e aprendizagem”, descreve de forma sucinta o modelo apresentado nesta dissertação, o protótipo e os resultados alcançados. Até a data de entrega desta dissertação não havia ocorrido o retorno de seu aceite.

7.3 Trabalhos futuros

Os resultados obtidos pelo protótipo construído para avaliar o modelo proposto nesta dissertação obteve bons resultados, apresentando uma precisão superior a 90% para a avaliação de vinculação textual realizada em ambos os estudos de caso. Esse resultado demonstra o potencial no uso das regras linguísticas para a construção de ferramentas com este objetivo. Durante a fase de testes foi possível verificar também que tanto o protótipo quanto o *parser* morfossintático utilizados apresentam falhas e pontos a serem melhorados.

Como trabalhos futuros estão previstos, a curto prazo, o remodelamento de alguns pontos como o tratamento para nomes próprios e os elementos de conhecimento, que devem ser aprofundados para o funcionamento futuro do protótipo. Além disso, o protótipo construído vislumbrou alcançar apenas duas categorias de perguntas, se fazendo assim necessária o tratamento das demais categorias. Tal tipo de tratamento poderá ser realizado adicionando mais regras ao repositório, ou adicionando etapas específicas de tratamento ao modelo proposto. Para que essa definição ocorra são necessários também alguns estudos relacionados especificamente a estes casos.

A longo prazo estão previstos testes a serem realizados em outras línguas, neste caso o Inglês e o Espanhol, para verificar o quão flexível o modelo é no caso de tratamento multilíngue. Isso indicaria também se de alguma forma o modelo proposto pode auxiliar na avaliação de reconhecimento de vinculação textual em línguas, que assim como a portuguesa, possuem certa dificuldade para possuírem um repositório semântico adequado às necessidades desta área.

REFERÊNCIAS

Associação Brasileira de Educação a Distância (ABED). **Relatório analítico da aprendizagem a distância no Brasil.** Editora Pearson Education do Brasil, São Paulo. Disponível em <http://www.abed.org.br/site/pt/midiateca/censo_ead/>. Acesso em: 13 nov 2013.

ADACHI, Ana Amélia Chaves Teixeira, (2009). **Evasão e evadidos nos cursos de graduação da Universidade Federal de Minas Gerais**, Dissertação de Mestrado. Faculdade de Educação, UFMG, 2009.

ADAMS, Rod et al. **Textual entailment through extended lexical overlap and lexico-semantic matching.** ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 2007.

ALVES, Ieda Maria. **A integração dos neologismos por empréstimo ao léxico português**. **A**LFA: Revista de Linguística, v. 28, p. 119–126, 1984.

ANDROUTSOPOULOS, Ion; MALAKASIOTIS, Prodromos. **A survey of paraphrasing and textual entailment methods**. Journal of Artificial Intelligence Research, v. 38, p. 135–187, 2010.

BAKER, Ryan, ISOTANI, Seiji, CARVALHO, Adriana. **Mineração de Dados Educacionais: oportunidades para o brasil**. Revista Brasileira de Informática na Educação, 19(02). P. 15-35, 2011.

BARBUR, Gabriel; BLAGA, Bogdan.; GROZA, Adrian. **OntoRich - A support tool for semi-automatic ontology enrichment and evaluation**. 2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing, p. 129–132, 2011.

BAR-HAIM, Yair et. al. **Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study.** 2007.

BICK, Eckhard. **The parsing system Palavras**.Tese de Doutorado, Universidade de Aarhus, 2000.

BIDERMAN, Maria Tereza Camargo. **Léxico e vocabulário fundamental**. ALFA: Revista de Linguística, p. 27–46, 1996.

BLOOM, Benjamin S. **The 2 sigma problem: The search for methods of group instruction as effective as one–to–one tutoring**. Educational Researcher, volume 13, number 6, pp. 4–16, 1984.

BLOOM, Benjamin S.; DURHAM, N.C. **Learning for mastery.** Regional Education Laboratory for the Carolinas and Virginia, 1968.

Censo EAD.BR. (2011). Relatório analítico da aprendizagem a distância no Brasil, Associação Brasileira de Educação a Distância (ABED), Editora Pearson Education do Brasil, São Paulo.

CHEN, Liang; LIU, Yajun. **Automated Scoring System Using Dependency-Based Weighted Semantic Similarity Model**. 2009 Second International Symposium on Knowledge Acquisition and Modeling, v. 1, p. 241–244, 2009.

CHOMSKY, Noam. **Three models for the description of language**. IRE Transactions on Information Theory, v. 2, 1956.

CHUONG, B. Do et al. **Self-Driven Mastery in Massive Open Online Courses. MOOCs FORUM**. September 2013, 1(P): 14-16. doi:10.1089/mooc.2013.0003.

CLARK, Charlotte R.; GUSKEY, Thomas R.; BENNINGA, Jacques S. **The effectiveness of mastery learning strategies in undergraduate education courses.** *Journal of Educational Research*, volume 76, number 4, pp. 210–214, 1983.

CLARKE, Daoud. **Meaning as Context and Subse- quence Analysis for Entailment.** In: the Second PASCAL Recognising Textual Entailment Chal- lenge, Venice, Italy, 2006.

CUNNINGHAM JR., R. Daniel. **Modeling Mastery Learning Through Classroom Supervision**. NASSP, v. 75, p. 83–87, 1991.

DAGAN, Ido et al. **Guest Editors Introduction: Recognizing Textual Entailment: Rational, Evaluation and Approaches**. Journal of Natural Language Engineering,2009.

DAGAN, Ido; GLICKMAN, Oren; MAGNINI, Bernardo. **The pascal recognising textual entailment challenge**.Recognising Textual Entailment Challeng, 2006.

DAGAN, Ido; GLICKMAN, Oren. **Probabilistic textual entailment: Generic applied modeling of language variability**. PASCAL Workshop on Learning Methods for Text Understanding and Mining, 2004.

DAGAN, Ido et al. **Recognizing Textual Entailment – Models and Applications**. Synthesis Lectures on Human Language Technologies, Toronto. Morgan & Claypool Publishers. 2013

DALFOVO, Michael Samir; JOS, Maria; DOMINGUES, Souza. **O Ambiente Virtual de Aprendizagem (AVA) na Universidade Regional de Blumenau (FURB)**. **EnADI**, p. 1–13, 2007.

DEDEK, Jan; VOJTÁS, Peter. **Semantic Annotation Semantically: Using a Shareable Extraction Ontology and a Reasoner**. Proceedings of SEMAPRO 2011, The Fifth International Conference on Advances in Semantic Processing, p. 29-34, Lisbon, 2011.

DEKANG Lin; PATRICK Pantel. Discovery of Inference Rules for Question Answering. Natural Language Engineering, 7(4):343-360, 2001.

FABRÍCIO, Breno Terra Azevedo; BEHAR, Patricia Alejandra; REATEGUI, Eliseo Berni. **Análise das mensagens de fóruns de discussão através de um software para mineração de textos** SBIE. n. 2004, p. 20–29, 2011.

FABRÍCIO, Breno Terra Azevedo. **MINERAFÓRUM: Um recurso de apoio para análise qualitativa em fóruns de discussão**. Programa de Pós-Graduação em Informática na Educação. UFRGS. Tese de Doutorado, 2011.

HERRERA, Jesús et al. **Textual Entailment Recognition Based on Dependency Analysis and WordNet.** In: the First Challenge Workshop Recognising Textual Entailment, pp. 21-24, 33–36 Southampton, U.K, 2005HUANG, Zhiheng; THINT, Marcus; CELIKYILMAZ, Asli. **Investigation of question classifier in question answering**. Association for Computational Linguistics, 2009.

HUTCHINS, John. **The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954**. Association for Machine Translation in the Americas, 2004.

IFTENE, Adrian. **Building a Textual Entailment System for the RTE3 Competition. Application to a QA System.** 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2008.

JOHNSON, Larry et al**. NMC Horizon Report: 2013 Higher Education Edition**. Austin, Texas: The New Media Consortium, 2013.

KIM, Jihie. et al. **An Intelligent Discussion-Bot for Guiding Student Interactions in Threaded Discussions**. Proceedings of the AAAI 2007 Spring Symposium on Interaction Challenges for Intelligent Assistants, 2007.

LAMJIRI, A. K.; KOSSEIM, L.; RADHAKRISHNAN, T. **Comparing the Contribution of Syntactic and Semantic Features in Closed versus Open Domain Question Answering**. International Conference on Semantic Computing ICSC 2007, p. 679–685, 2007.

LESMO, Leonardo et al. **Tulsi, P.: an NLP system for extracting legal modificatory provisions**. Artificial Intelligence and Law. v. 21, p. 139-172. Springer Netherlands. 2013.

LIN, Fu-Ren; HSIEH, Lu-Shih; CHUANG, Fu-Tai. **Discovering genres of online discussion threads via text mining**. Computers & Education, v. 52, n. 2, p. 481–495, 2009.

LITKOWSKI, Ken. **Overlap Analysis in Textual Entailment Recognition.** In: TAC 2009 Workshop, National Institute of Standards and Technology Gaithersburg, Maryland USA, 2009.LIU, Peng-Yuan; ZHAO, Tie-Jun; YU, Xiao-Feng. **Application-Oriented Comparison and Evaluation of Six Semantic Similarity Measures Based on Wordnet**. 2006 International Conference on Machine Learning and Cybernetics, p. 2605–2610, 2006.

MACEDO, Alexandra et al. **Using Text-Mining to Support the Evaluation of Texts Produced Collaboratively**. Education and Technology for a Better World, IFIP Advances in Information and Communication Technology, Volume 302. ISBN 978-3-642-03114-4. Springer Berlin Heidelberg, 2009, p. 368.

MAJUMDAR, Debarghya; BHATTACHARYYA,Pushpak. **Lexical Based Text Entailment System for Summarization Settings of RTE6.** In: the Text Analysis Conference (TAC 2010), National Insti- tute of Standards and Technology Gaithersburg, Maryland, USA, 2010.

MILLER, George A. et al. **A semantic concordance.** HLT '93 Proceedings of the workshop on Human Language Technology, 1993.

MONTALVO-HUHN, Orlando; TAYLOR, Stephen. **Textual Entailment - Fitchburg State College.** Text Analysis Conference, 2008.

MOSCHITTI, Alessandro; ZANZOTTO; Fabio Massimo. **Encoding Tree Pair-based Graphs in Learning Algorithms:the Textual Entailment Recognition Case.** Coling 2008: Proceedings of 3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing, pages 25–32, 2008.

NADKARNI, Prakash M.; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. **Natural language processing: an introduction**. Journal of the American Medical Informatics Association, v. 18, p. 544–551, 2011.

NETO, Pasquale Cipro; INFANTE, Ulisses. **Gramática Da Língua Portuguesa**. São Paulo: Scipione, 1998.

OLIVEIRA, Roberto. L.; ESMIN, Ahmed. **Monitoramento Automático de Mensagens de Fóruns de Discussão Usando Técnica de Classificação de Texto Semi-Supervisionado**. Anais do 23º Simpósio Brasileiro de Informática na Educação (SBIE 2012), ISSN 2316-6533 Rio de Janeiro, 26-30 de Novembro de 2012

RAY, Santosh Kumar; SINGH, Shailendra; JOSHI, B. P. **A semantic approach for question classification using WordNet and Wikipedia**. Pattern Recognition Letters, v. 31, n. 13, p. 1935–1943, out. 2010.

REU-DEBOVE, Josette; MORAIS, Clóvis Barleta de. **Léxico e dicionário**. ALFA: Revista de Linguística, p. 45–69, 1984.

RIOS, Miguel. **Recognizing Textual Entailment with a Semantic Edit Distance Metric.** 11th Mexican International Conference on Artificial Intelligence, 2012.

ROTH, Dan; SAMMONS, Mark.**Semantic and Logical Inference Model for Textual Entailment.** Workshop on Textual Entailment and Paraphrasing, 2007.

RUS, Vasile. **A study on textual entailment**. 17th IEEE Internacional Conference on Tools with Artificial Intelligence, p. 333–341, 2005.

SHARIATMADARI, Shahdad; MAMAT, Ali. **Discovering semantic similarity association in semantic search system**. Proceedings of the 11th WAS, p. 331, 2009.

SILVA, Júlia Kikuye Kambara et al. **Automatização do Processo de Identificação de Presença Social em Fóruns e Chats** . Anais do 23º (SBIE 2012, ISSN 2316-6533 Rio de Janeiro, 26-30 de Novembro de 2012.

TSUCHIDA, Masaaki; ISHIKAWA, Kai. **IKOMA at TAC2011: A Method for Recognizing Textual Entailment using Lexical-level and Sentence Structure-level features.** Fourth Text Analysis Conference, 2011.

VANDERWENDE, Lucy et al. **Syntactic contributions in the entailment task: an implementation.** In Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop, 2006.

VARMA, Vasudeva et al. **IIIT Hyderabad at TAC 2009.** In: TAC 2009 Workshop, National Institute of Standards and Technology Gaithersburg, Maryland USA, 2009.

WANG, Rui; NEUMANN, Günter. **Recognizing textual entailment using sentence similarity based on dependency tree skeletons.** ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 2007.

WIMALASURIYA, Daya C.; DOU, Dejing, **Ontology-based information extraction: an introduction and a survey of current approaches**. Journal of Information Science, v. 36, n. 3, p. 306–323, 2010.

ZHANG, Shangqing et al. **Recognizing Textual Entailment with synthetic analysis based on SVM and feature value control.** IEEE 3rd International Conference on Software Engineering and Service Science, 2012.

ANEXO A ARTIGOS PUBLICADOS

METZ Evandro; RIGO, Sandro José; ABECH, Márcia; BARBOSA, Jorge; COSTA, Cristiano. **O papel do Processamento de Linguagem Natural e da Representação de Conhecimento na extração de informações em mensagens textuais na Educação a Distância** (aceito para publicação em novembro/2013). Simpósio Brasileiro de Informática na Educação (SBIE), 2013, Campinas - SP. Simpósio Brasileiro de Informática na Educação (SBIE). Campinas - SP: SBC, 2013. v. 1. p. 1-10.

ANEXO B TABELA DE DADOS DO PRIMEIRO ESTUDO DE CASO

		Avaliação sobre a vinculação do texto realizada pelo especialista (%)	Avaliação sobre a vinculação do texto realizada pelo protótipo (%)
Pergunta 1	*Quais cores apresenta o padrão CMYK?*		
Resposta ótima	São: ciano, magenta, amarelo e preto	100	
Respostas candidatas	As cores padrão CMYK são: ciano, magenta, amarelo e preto	100	100
	São as cores amarelo e preto	50	50
	As cores são verde, vermelho e azul	0	0
	São ciano e vermelho	25	25
Pergunta 2	*Quem resolveu o problema do paciente?*		
Resposta ótima	O médico estagiário	100	
Respostas candidatas	O problema foi resolvido pelo médico	50	50
	O estagiário resolveu o problema	50	50
	O cardiologista.	0	0
	A enfermeira e o médico resolveram o problema	50	50
Pergunta 3	*Quais são os princípios da administração pública direta e indireta de qualquer dos Poderes da União, mencionados no artigo 37 da Constituição Federal Brasileira?*		
Resposta ótima	Legalidade, impessoalidade, moralidade, publicidade e eficiência	100	
Respostas candidatas	São: Legalidade, impessoalidade e moralidade.	60	60
	Os princípios consistem em moralidade e publicidade.	40	40
	São: segurança e ética.	0	0
	Legalidade, impessoalidade, ética e segurança.	40	40
Pergunta 4	*Quais são as premissas do Endomarketing?*		

Resposta ótima	Transparência, qualidade e afeto.	100	
Respostas candidatas	São transparência e ética.	33,33	33,33
	As premissãs são: transparência, qualidade e ética	66,66	66,66
	Qualidade e diversidade.	33,33	33,33
	São segurança e ética	0	0
Pergunta 5	*Qual o tipo de marketing mais utilizado?*		
Resposta ótima	O tipo de marketing mais utilizado é o Marketing direto.	100	
Respostas candidatas	É o marketing direto	100	100
	O mais utilizado é o marketing multinivil	0	0
	O marketing viral.	0	0
	O tipo de marketing mais utilizado é o político	0	0
Pergunta 6	*Quais são os P's do marketing?*		
Resposta ótima	Praça, preço, produto e promoção.	100	
Respostas candidatas	São praça, preço e produto.	75	75
	Praça e propaganda.	25	25
	Propaganda e publicidade.	0	0
	Preço e popularização.	25	25
Pergunta 7	*No que consiste uma análise de Participação relativa de mercado?*		
Resposta ótima	A análise de participação consiste nos concorrentes	100	
Respostas candidatas	Consiste nos concorrentes e nos consumidores.	100	100
	Nos consumidores.	0	0
	Nos concorrentes e mercado interno dependem a análise de participação relativa do mercado	100	100
	Na economia mundial.	0	0

Pergunta 8	*O desenvolvimento sustentável procura atender as necessidades atuais sem comprometer a capacidade de quais gerações?*		
Resposta ótima	As gerações futuras.	100	
Respostas candidatas	Os baby boomers	0	0
	A geração atual.	0	0
	A e B são as gerações que o desenvolvimento sustentável atende às necessidades	0	0
	C e D são essas gerações	0	0
Pergunta 9	*Quem gritou "Independência ou Morte" às margens do Ipiranga?*		
Resposta ótima	Foi D. Pedro I	100	
Respostas candidatas	Quem gritou foi Pedro Alvares Cabral	0	0
	Foi D. Pedro II quem griou "Independência ou morte"	0	0
	Foi Pedro de Alcântara e D. Pedro I	100	100
	Libero Badaró e Pedro de Alcântara	0	0
Pergunta 10	*A indústria cultural consiste na reprodução em série de elementos culturais, além de quais outros?*		
Resposta ótima	Além de arte e entretenimento.	100	
Respostas candidatas	Além dos mercadológicos	0	0
	Além dos elementos tradicionais e artísticos.	50	50
	Entretenimento e gastronomia são elementos culturais	50	50
	Gastronomia, tradição e entretenimento	50	50
Pergunta 11	*Qual o principal benefício da água?*		
Resposta ótima	O principal benefício da água é hidratar	100	
Respostas candidatas	Hidratar e curar doenças	100	100
	Curar doenças é a função da água	0	0
	Transportar líquidos no corpo é uma das funções da água	0	0

	Auxiliar na formação de água no corpo	0	0
Pergunta 12	*Quem desenvolveu a teoria da relatividade?*		
Resposta ótima	Quem desenvolveu foi Albert Einstein	100	
Respostas candidatas	Albert Einstein e Darwin	100	100
	Darwin inventou a teoria da relatividade	0	0
	Foi Freud quem inventou essa teoria	0	0
	Foram Freud, Einstein e Darwin	100	0
Pergunta 13	*Qual seleção derrotou o Brasil na final da copa de 98?*		
Resposta ótima	Foi a seleção da França.	100	
Respostas candidatas	A seleção da Inglaterra venceu o Brasil em 98	0	0
	A seleção da Itália	0	0
	A Alemanha venceu o Brasil na copa de 98.	0	0
	A Holanda	0	0
Pergunta 14	*Quais estados a linha do equador atravessa?*		
Resposta ótima	Os estados brasileiros que são cortados pela linha do Equador são: Pará, Amapá, Amazonas e Roraima.	100	
Respostas candidatas	São: Pará e São Paulo	25	25
	Rio de Janeiro e Tocantins são atravessados pela linha do Equador	0	0
	Pernambuco e Goiás	0	0
	Brasília, Amapá e Rondônia	25	25
Pergunta 15	*Quais são os grupos de minerais?*		
Resposta ótima	São minerais metálicos e minerais não metálicos	100	
Respostas candidatas	São: Minerais, Sais e rochas magmáticas	0	0
	Sedimentares e Metamórficas	0	0
	minerais metálicos e magmáticos	50	40
	sais	0	0

Pergunta 16	*Qual a maior estrela conhecida?*		
Resposta ótima	É o sol	100	
Respostas candidatas	O sol e a lua	100	100
	A lua é a maior estrela conhecida	0	0
	A maior estrela é Júpiter	0	0
	O sol e Plutão são as maiores estrelas que existem	100	100
Pergunta 17	*Quando surgiu a internet?*		
Resposta ótima	A internet surgiu na década de 1960.	100	
Respostas candidatas	A internet surgiu no ano de 1992.	0	0
	Surgiu entre1998 e 2001	0	0
	Em 1982	0	0
	Entre 1974 e 1976	0	0
Pergunta 18	*Qual era o nome da missão do primeiro homem a pisar na lua?*		
Resposta ótima	O nome da missão era Apollo 11.	100	
Respostas candidatas	O nome era Houston	0	0
	Águia era o nome da missão	0	0
	O nome era Mar da Tranquilidade	0	0
	Era Columbia	0	0
Pergunta 19	*Quem desenvolveu a teoria da evolução?*		
Resposta ótima	Foi Charles Darwin	100	
Respostas candidatas	Foram Darwin e Laplace	100	0
	Henri Poincaré e Einstein	0	0
	Foi Galileu Galilei	0	0
	Laplace	0	0
Pergunta 20	*Quais são os 3 países mais populosos do mundo?*		
Resposta ótima	Os países mais populosos são: China, Índia e EUA	100	
Respostas candidatas	São a China e a Indonésia	33,33	33,33
	A China, o Brasil e a Hungria	33,33	33,33
	O México e a Alemanha são os países mais	0	0

	populosos		
	Portugal e Irã	0	0
Pergunta 21	*Quais são oontinentes?*		
Resposta ótima	África, América, Antártida, Ásia, Europa e Oceania	100	
Respostas candidatas	São África e Europa	33,33	33,33
	São Antartica e Indico	0	0
	São américa do sul e américa do norte	0	0
	África, América, Antártida e Ásia são os nomes dos continentes	83,33	83,33
Pergunta 22	*Quais são os principais planetas do sistema solar?*		
Resposta ótima	Mercúrio, Vénus, Terra, Marte, Júpiter, Saturno, Urano e Netuno.	100	
Respostas candidatas	São: Mercúrio, Marte, Júpiter e Saturno	50	50
	Mercúrio, Vênus, Terra, Marte e Júpiter são os principais planetas do sistema solar	62,5	62,5
	Sãos os planetas Plutão, Vênus e Terra	25	10
	Os planetas são Asgard e Plutão	0	0
Pergunta 23	*Qual é o maior rio do mundo?*		
Resposta ótima	É o Rio Amazonas	100	
Respostas candidatas	É o Rio Nilo	0	0
	O maior rio é o Rio Negro	0	0
	Rio Solimões	0	0
	Rio Yangtze é o maior rio do mundo	0	0
Pergunta 24	*Em que ano foi fundada a cidade de Brasília?*		
Resposta ótima	A cidade de Brasília foi fundada em 1960	100	
Respostas candidatas	Foi fundada em 1962	0	0
	Brasília foi fundada em 1970	0	0
	O ano de fundação de Brasília é 1965	0	0

	Foi fundada entre 1960-1962	100	0
Pergunta 25	*Quais são os 3 poderes do Brasil?*		
Resposta ótima	Executivo, Legislativo e Judiciário	100	
Respostas candidatas	São: Executivo, Legislativo e Hierárquico	66,66	66,66
	Disciplinar, regulamentar, e de polícia são os poderes executivos do Brasil	0	0
	São os Poderes federal e judicial	0	0
	São judiciário, executivo, legislativo e hierárquico	100	100
Pergunta 26	*Quem escreveu o livro O Guarani?*		
Resposta ótima	Quem escreveu o livro O Guarani foi José de Alencar	100	
Respostas candidatas	Foi Castro Alves	0	0
	O Guarani foi escrito por Aluisio de Azevedo e José de Alencar	100	100
	Quem escreveu o livro foi Cecilia Meireles	0	0
	O livros foi escrito por Machado de Assis e Cecilia Meireles	0	0
Pergunta 27	*De que cores derivam as cores quentes?*		
Resposta ótima	Amarela, Laranja e Vermelha.	100	
Respostas candidatas	Das cores azul, verde e vermelha.	33,33	33,33
	As cores quentes derivam do amarelo e do branco	33,33	33,33
	Do preto e do roxo	0	0
	Derivam das cores cinza e preta	0	0
Pergunta 28	*De que se constitui o vidro?*		
Resposta ótima	areia, sódio, cálcio e componentes químicos	100	
Respostas candidatas	De sais e água	0	0
	De cimento e pedras é constituído o vidro	0	0
	O vidro é constituído de cálcio e terra	0	0

	Sódio e cimento	25	20
Pergunta 29	*Quais as principais características do romantismo?*		
Resposta ótima	Nacionalismo, historicismo e medievalismo	100	
Respostas candidatas	As principais são romance, ternura e historicismo	33,33	33,33
	São amor, patriotismo e ternura	0	0
	Criticismo e romance são as principais características do romantismo	0	0
	Nacionalismo e romance	33,33	33,33
Pergunta 30	*Quais são as gerações de autores presentes no romantismo?*		
Resposta ótima	A primeira, a segunda e a terceira geração de autores.	100	
Respostas candidatas	No romantismo, há apenas a primeira geração de autores.	33,33	33,33
	A primeira, a segunda, a terceira e a quarta geração;	100	100
	Somente a primeira e a segunda gerações	66,66	66,66
	Há a primeira, a segunda, a terceira, a quarta e a quinta geração de autores no romantismo	100	100

ANEXO C TABELA DE REGRAS LINGUISTICAS UTILIZADAS NO PRIMEIRO ESTUDO DE CASO

Nº	Elementos da regra									
1	art	n	v-	*						
2	v-	*	conj	*						
3	*	conj	*	v-	art	n	prp	art	*	
4	*									
5	*	conj	*							
6	*	prp	art							
7	*	conj	*	v-	art	n				
8	art	n	v-	prp	*	conj	*			
9	v-	art	*							
10	v-	art	*	conj	*					
11	art	n	v-	*	conj	*				
12	prp	n	v-	*						
13	prp	n	v-	*	conj	*				
14	art	n	prp	n	v-	prp	art	*		
15	art	*								
16	prp	n	*							
17	art	adj	n	prp	art	n	v-	*		
18	*	v-	art	n						

19	art	n	prp	art	*					
20	v-	*								
21	prp	*								
22	*	v-	art	n	adj	prp	art			
23	art	n	prp	n	adv	v-	v-	art	*	
24	art	n	v-	v-	prp	art	*			
25	art	*	v-	art	n					
26	art	*	conj	art	*	v-	art	n		
27	art	adv	v-	v-	art	*				
28	art	prop	prp	art	n	v-	*	conj	*	
29	v-	prp	art	n	prp	art	*			
30	prp	art	*							
31	*	conj	*	v-	pron	n				
32	*	v-	pron	n						
33	adv	prp	art	*						
34	*	conj	*	v-	n					
35	*	conj	*	v-	n	adj				
36	*	v-	n	adj						
37	*	v-	n	adj	adv					
38	*	v-	art	n	prp	art	n			
39	*	v-	num	prp	art	n	prp	art	n	

40	art	*	v-							
41	art	*	v-	art	adj					
42	art	adj	n	v-	*					
43	art	n	adv	adj	v-	*				
44	art	n	adv	adj	v-	*	conj	*		
45	*	v-	art	n	adv	adj				
46	*	art	n	prp	art	n				
47	*	v-	art	adj	n	prp	art	n		
48	*	v-	art	adj	n					
49	*	v-	art	adj						
50	art	adj	n	v-	art	*				
51	art	adj	n	v-	art	*	conj	*		
52	art	adj	adj	prp	art	n	v-	*		
53	art	adj	adj	prp	art	n	v-	*	conj	*
54	art	adj	v-	*						
55	art	adj	v-	*	conj	*				
56	*	v-	art	adj	adj	prp	art	n		
57	prp	art	n	v-	*					
58	prp	art	n	v-	*	conj	*			
59	adv	*	n							
60	adv	*	conj	*	n					

61	*	prp	n							
62	*	prp	n	prp	art	n				
63	art	n	n	prop	v-	*				
64	art	n	n	prop	v-	*	conj	*		
65	v-	art	n	*						
66	v-	art	n	*	conj	*				
67	*	conj	*	v-	art	adj	adj	prp	art	n
68	v-	prp	art	*	conj	pron	art	*		
69	prp	*	conj	*						

Legenda:

Abreviação	Morfologia
art	Artigo
n	Substantivo
v-	Verbo
conj	Conjunção
prp	Preposição
adj	Adjetivo
prop	Nome Próprio
pron	Pronome
num	Numeral
adv	Advérbio
*	Informação Adicional

ANEXO D TABELA DE DADOS DO SEGUNDO ESTUDO DE CASO

		Avaliação sobre a vinculação do texto realizada pelo especialista (%)	Avaliação sobre a vinculação do texto realizada pelo protótipo (%)
Pergunta 1	**Quais as cores da bandeira do Brasil?**		
Resposta Ótima	Amarelo, azul, verde e branco.	100	
Respostas Candidatas	amarela azul e verde branco	100	100
	as cores são: verde, amarelo, azul e branco	100	100
	Verde, Amarelo, Azul	75	75
	Verde, Amarelo, Braco, Preto, Azul	75	75
	As cores são Azul, Verde, Amarelo e Branco	100	100
	As cores da bandeira do Brasil são: verde, amarelo, azul e branco.	100	100
	As cores da bandeira brasileira são azul, verde e amarelo	75	75
	Verde, amarelo, azul e branco	100	100
	Verde, amarelo e azul.	75	75
	Verde, amarelo, azul e branco	100	100
	Verde,Amarelo,Azul,Branco	100	100
	Azul, verde, branco e amarelo	100	100
	Verde, amarela e azul. Se considerar estrelas e tarja, também tem o branco.	100	100

	azul, amarelo, branco e verde	100	100
	Verde, amarelo, azul e branco	100	100
	As cores da bandeira são amarelo,verde,azul, preto e branco.	100	100
	azul, amarelo e verde	75	75
	Verde e Amarela	50	50
Pergunta 2	**Quem é o presidente do Brasil?**		
Resposta Ótima	Dilma	100	
Respostas Candidatas	dilma russef	0	0
	é a dilma russef	0	0
	Galvão Bueno	0	0
	Dilma Rousseff	100	0
	A Presidente do Brasil, infelizmente, é a Dilma Rousseff	100	0
	A presidente do Brasil é a Dilma Rousseff.	100	0
	Dilma	100	100
	Dilma	100	100
	Dilma Rouseff	100	0
	Dilma	100	100
	Dilma Russef	100	0

	Dilma	100	100
	Dilma	100	100
	Dilma	100	100
	Dilma	100	100
	O presidente do Brasil no momento é a Dilma.	100	100
	Dilma Roussef	100	0
	Dilma Rouseff	100	0
Pergunta 3	**Como é caracterizada a febre no ser humano?**		
Resposta Ótima	Temperatura acima de 37 graus.	100	
Respostas Candidatas	aumento da temperatura do corpoa acima de 37 graus	100	100
	temperatura acima de 37 graus e calafrios	100	100
	Alta Temperatura	50	25
	É a elevação da temperatura do corpo humano para cima dos limites considerados normais	75	25
	A febre é caracterizada por alta temperatura no corpo	75	25
	Febre ou pirexia é a elevação da temperatura do corpo humano para cima dos limites considerados normais (36 a 37,4 °C).	100	25
	o aquecimento do corpo	50	0

	Possuir a temperatura corporal acima de 37,5 graus	75	75
	Testa muito quente.	50	0
	Aumento de temperatura do corpo	75	25
	Aumento na Temperatura do Corpo,Cansaço	75	25
	Sintomas como temperatura do corpo acima de 37 graus e sensação de frio	100	100
	Temperatura corporal acima dos 37°C	100	75
	aumento de temperatura corporal	75	25
	é caracterizada quando uma pessoa fica com o corpo com a temperatura acima de 37° aproximadamente	100	75
	O corpo estar acima de 38 graus.	100	50
	Aumento da temperatura corporal acima dos níveis normais	75	50
	Temperatura alta	50	25
Pergunta 4	**Quem é o atual Papa da igreja católica?**		
Resposta Ótima	Francisco	100	
Respostas Candidatas	Papa Francsico	0	0
	é o Papa Francsico	0	0
	Neto Fagundes	0	0
	Jorge Mario Bergoglio	0	0

	Papa Francisco, com nome Jorge Mario Bergoglio	100	0
	O atual papa da igreja católica é o Papa Francisco.	100	0
	Francisco	100	100
	Francisco. Além de papa é Argentino.	100	100
	Não sei.	0	0
	Gregorio	0	0
	João Francisco Pedro II	0	0
	Um argentino chamado Bergoglio	0	0
	Francisco	100	100
	não lembro o nome	0	0
	Francisco	100	100
	O papa é Francisco.	100	100
	Um velho reacionário com tendências à exposição de humildade	0	0
	Francisco	100	100
Pergunta 5	**Quais são os pontos cardeais?**		
Resposta Ótima	Norte, sul, leste e oeste.	100	
Respostas Candidatas	norte sul leste oeste	100	100
	os pontos cardeais são: norte, sul, leste e oeste	100	100

	Opala	0	0
	Norte, Sul, Leste, Oeste, Sudeste, Sudoeste, Noroeste, Nordeste	100	100
	Os pontos Cardeais são Norte, Sul, Leste, Oeste	100	100
	Os pontos cardeias são: norte, sul, leste e oeste.	100	100
	Os pontos cardeais são quatro: norte, leste, sul, oeste	100	100
	Norte, sul, leste o oeste.	100	100
	Norte, Sul, Leste, Oeste.	100	100
	Norte, sul, leste e Oeste	100	100
	Norte,Sul,Leste,Oeste	100	100
	Norte, sul, leste e oeste	100	100
	Norte, sul, lest e oeste.	75	75
	não não lembro	0	0
	Norte, Sul, Leste e Oeste	100	100
	Os pontos são sul, leste, oeste e norte.	100	100
	Norte, Sul, Leste e Niv-Mizzet	75	75
	Norte, Sul, Leste e Oeste	100	100
Pergunta 6	**O que o motorista deve fazer caso seja o primeiro a chegar a um local de acidente?**		
Resposta Ótima	Sinalizar o local.	100	

Respostas Candidatas	sinalizar o local	100	100
	ele deve sinalizar o local	100	100
	Chamar a Ambulância	0	0
	Deve chamar a emergência	0	0
	O motorista deve chamar imediatamente a Ambulância em caso de ser o primeiro a chegar a um local de acidente, e só tentar salvar o acidentado em caso de realmente saber o que está fazendo ou caso de risco de explosão do veículo acidentado	0	0
	O motorista deve prestar os primeiros socorros.	0	0
	Caso alguém, que não seja motorista, ainda não tenha ligado para a emergência, ligar para uma ambulância.	0	0
	Chamar ajuda.	0	0
	Parar o veiculo, ajudar o acidentado, chamar ambulancia.	0	0
	Estacionar em lugar seguro, sinalisar o local e prestar os primeiros socorros	100	100
	Parar o Carro	0	0
	Se possível nada	0	0
	Chamar a ambulancia e por as devidas sinalizaçoes.	100	0
	socorrer se possivel e chamar ajuda profissional	0	0
	Ligar para o pronto socorro, caso tenha feridos, policia etc.	0	0
	Chamar ajuda policial.	0	0

	Ligar para a SAMU	0	0
	Chamar por socorro	0	0
Pergunta 7	**Quem foi o primeiro homem a pisar na lua?**		
Resposta Ótima	Neil Armstrong	100	
Respostas Candidatas	lui amstrong	0	0
	foi neil amstrong	0	0
	Silvio Santos	0	0
	Foi Neil Alden Armstrong	100	0
	Neil Alden Armstrong	100	0
	O primeiro homem a pisar na lua foi Neil Armstrong.	100	100
	A lua foi pisada primeiramente por Nicolas Cage.	0	0
	Neil Armstrong	100	100
	Armstrong.	100	0
	Louis Anstrong	0	0
	Neil Armstrong	100	100
	O americano Armstrong	100	0
	Neil Armstrong	100	100
	não sei o nome	0	0

	Não lembro	0	0
	Maicon Jordan	0	0
	Dr. Manhattan	0	0
	Neil Armstrong	100	100
Pergunta 8	**Qual ser mitológico é parte homem parte touro?**		
Resposta Ótima	Minotauro.	100	100
Respostas Candidatas	monotauro	0	0
	é o minotauro	100	100
	Minotauro	100	100
	Minotauro	100	100
	Minotauro	100	100
	Minotauro.	100	100
	Minotauro, e por isso, há um lutado de UFC que tem este apelido.	100	100
	Minotauro	100	100
	Centauro	0	0
	Minotauro	100	100
	Minotauro	100	100
	Centauro	0	0

	Minotauro	100	100
	centauro	0	0
	Centauro	0	0
	Ele se chama Minotáuro	100	100
	Bucentauro	0	0
	Centauro	0	0
Pergunta 9	**O que representa o circulo vermelho na bandeira do Japão?**		
Resposta Ótima	Sol.	100	100
Respostas Candidatas	paz	0	0
	representa o sol nascente	100	100
	Sangue	0	0
	Representa o Sol	100	100
	Representa o Sol	100	100
	O sangue derramado para criar o país.	0	0
	Não sei. Porém, deve haver uma explicação ocidental para isso.	0	0
	Sol	100	100
	Não sei.	0	0
	Sol	100	100

	Sol	100	100
	Sol nascente	100	100
	O sol nascente.	100	100
	não sei	0	0
	não sei	0	0
	O circulo representa o sol.	100	100
	O percentual de Japão no Japão. É um gráfico em pizza.	0	0
	Sol	100	100
Pergunta 10	**Quem desenvolveu a teoria da relatividade?**		
Resposta Ótima	Albert Einstein	100	
Respostas Candidatas		0	0
		0	0
	Albert Einstein	100	100
	Quem desenvolveu foi Henri Poincaré, Hendrik Lorentz e posteriormente foi publicada por Albert Einstein	100	100
	O Gênio Albert Einstein	100	100
	Albert Einstein.	100	100
	Albert Einstein foi a pessoa quem desenvolveu tal teoria.	100	100

	Albert Einstein	100	100
	Isac Newton.	0	0
	Isac Newton	0	0
	Albert Einstein	100	100
	Eisntein	100	0
	Einsten	100	0
	Albert Einstein	100	100
	não sei	0	0
	Foi Hiestein	0	0
	Depende de pra quem vc perguntar	0	0
	Albert Einstein	100	100
Pergunta 11	**Quais estados brasileiros são atravessados pela linha do equador?**		
Resposta Ótima	Pará, Amapá, Amazonas e Roraima.	100	
Respostas Candidatas		0	0
		0	0
	Rio Grande do Sul	0	0
	No Brasil, os estados atravessados pela linha do Equador são: Amapá, Para, Roraima e Amazonas.	100	100
	O Estado Província Rio Grande do Sul	0	0

	http://lmgtfy.com/?q=Quais+estados+s%C3%A3o+atravessados+pela+linha+do+equador%3F	0	0
	São vários de diversos países, mas não encontrei fonte segura com a relação de todos.	0	0
	Amazonas, Pará, Amapá e Roraima	100	100
	São 4 mas não sei quais são.	0	0
	Acre, Anazonas e Pará	25	25
	Mato Grosso do Sul, Minas Gerais,São Paulo	0	0
	Amapá, Roraima...	50	50
	Estados do norte e nordeste em peso.	0	0
	não sei de cor	0	0
	Amazonas, Para, e mais algumas cidades do Norte	50	50
	Amapá e Macapá	25	25
	Nenhum. Ela passa por cima, não por dentro.	0	0
	Pará, Amazonas, Amapa e Roraima	100	100