Interface em Linguagem Natural para uma Sublinguagem de Câncer de Pele

Yago Henrique Pereira
Universidade Estadual de Londrina
Londrina, PR, Brasil
yago.henriquep@gmail.com

Cinthyan Renata Sachs Camerlengo de Barbosa
Universidade Estadual de Londrina
Londrina, PR, Brasil
cinthyan@uel.br

Arthur Alexandre Artoni
Universidade Estadual de Londrina
Londrina, PR, Brasil
arthurartoni@uel.br

José Luiz Villela Marcondes Mioni
Universidade Estadual de Londrina
Londrina, PR, Brasil
luiz.vmm@gmail.com

Jacques Duílio Brancher
Universidade Estadual de Londrina
Londrina, PR, Brasil
jacques@uel.br

ABSTRACT

This work presents an interface for a Skin Cancer Sublanguage, which uses a non-SQL oriented database technology, called MongoDB, implementing the concept of Natural Language Interface to Databases (NLIDB). When compared to relational databases, No-SQL (Not Only SQL) mechanism is more scalable, it provides superior performance and addresses several issues that a relational model is not designed to solve. All obtained words and their variations are labeled and stored in a MongoDB database, thus allowing customized queries to search and join morphemes. A dictionary will be used to build dialogs, narratives and diagnoses in a Serious Game with educational purposes that in addition obtains accurate answers to the project's database.

KEYWORDS

Natural Language Interface, MongoDB, Skin Cancer, Serious Game.

1 INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é uma área que utiliza métodos computacionais para o estudo de linguagem natural (falada ou escrita), a qual é dividida em múltiplas partes. Algumas dessas partes são descritas a seguir para explicar os processos utilizados neste trabalho.

Um componente essencial em muitos sistemas de PLN modernos são os bancos de dados lexicais, que frequentemente consistem em grandes quantidades de entradas altamente detalhadas e bem documentadas [1].

A Tokenização consiste em identificar letras maiúsculas e palavras compostas, remover caracteres especiais e *tags*, e também a verificação ortográfica. Todo processo pode ser feito com um simples código de computador, entretanto a supervisão humana não pode ser descartada, como no caso da utilização de cada palavra no contexto proposto é necessária. No léxico, os dados são armazenados como palavras, completas ou divididas, com informações sobre seu significado e uso. O conceito de "dividir palavras"permite a classificação dos verbos, conjugação e identificação de palavras morfemas, resultando na formação de outras palavras a partir da primeira palavra adquirida (palavra raiz).

O sucesso de um sistema de PNL dependerá do seu conhecimento sobre o domínio da aplicação, mais precisamente da abrangência do modelo de linguagem natural que engloba e da eficácia dos algoritmos que utiliza [2].

Segundo Schabes et al. [3], a lexicalização em uma perspectiva computacional é importante, uma vez que as gramáticas lexicalizadas tendem a ser analisadas de forma muito mais eficiente do que as não lexicalizadas. Comparando com a Forma Normal de Greibach (FNG), considerada como um tipo de lexicalização fraca, uma vez que não preserva a estrutura original da gramática, a Gramática Lexicalizada pode ser considerada uma versão mais forte da FNG, já que as estruturas são preservadas e não apenas os conjuntos de strings (capacidade gerativa fraca). A lexicalização é de interesse do ponto de vista linguístico, visto que a maioria das teorias linguísticas atuais fornece quantidade léxica de fenômenos que são considerados puramente sintáticos [3].

Com isso, as informações colocadas no léxico são aumentadas em quantidade e complexidade [4]. De acordo com Joshi et al. [4], alguns formalismos linguísticos ilustram o uso crescente da informação lexical, como: regras lexicais na Gramática Léxico-Funcionais, Gramática de Estrutura Frasal Generalizada (GPSG – *Generalized Phrase Structure Grammar*), Gramática de Estrutura de Frase Dirigida pela Cabeça (HPSG – *Head-Driven Phrase Structure Grammar*), Gramática de Combinação Categorial, Versão de Karttunen da Gramáticas Categoriais, algumas versões da Teoria da Regência e Ligação (teoria GB) de Chomsky e Gramáticas Léxicas.

Todo método para lexicalizar as Gramáticas Livres de Contexto (GLCs) tem compartilhado com a LTAG (*Lexicalized Tree Adjoining Grammar*), a infeliz característica de piorar o desempenho computacional ao invés de melhorá-lo [3]. Com isso, o foco principal deste artigo será nas Gramáticas Livres de Contexto Lexicalizadas (GLCL), originalmente introduzidas em [5], as quais combinam a elegância da LTAG e a eficiência da (GLC).

O presente artigo trabalha a utilização de GLCs para definições reais de uma interface na área médica com base em [6]. Contudo, consultas referentes ao diagnóstico de câncer de pele foram mapeadas para a GLCL. O subconjunto de linguagem utilizado inclui vozes ativas e passivas, orações relativas e interrogativas, suas combinações e pronomes.

O Serious Game (Jogo Sério) que está sendo criado utiliza-se de um dicionário e tem como objetivo auxiliar os médicos não dermatologistas a respeito do diagnóstico de Câncer de Pele. Ainda, envolve quais perguntas devem ser feitas durante o atendimento clínico, sendo esse o motivo da seleção da sublíngua do câncer de pele para este trabalho. A possível aplicação de uma Interface em Linguagem Natural para Banco de Dados (ILNBD) visa gerar consultas ricas sobre descrições diagnósticas.

Em síntese, o objetivo deste trabalho foi implementar um sistema de PLN e uma ILNBD em uma determinada sublíngua, utilizando uma base de dados não-SQL, denominada MongoDB[1]. A estrutura deste trabalho está dividida da seguinte forma: a segunda seção trata-se do processo de seleção das palavras, a terceira seção mostra como essas foram classificadas, a quarta seção apresenta a Interface em Linguagem Natural para Câncer de Pele utilizando da tecnologia MongoDB e a última contém as conclusões e trabalhos futuros.

2 SELEÇÃO DE PALAVRAS

O processo de seleção de palavras foi realizado em três etapas: reuniões com médicos e dermatologistas residentes no Hospital Universitário da Universidade Estadual de Londrina; análise das indicações bibliográficas (e.g., [7], recomendada por um especialista na área de dermatologia); consultas nos sites do INCA (*Instituto Nacional do Câncer do Brasil*)[2] e da SBD (*Sociedade Brasileira de Dermatologia*)[3].

Vinculando dois tipos de discursos, como o discurso formal (dos livros e artigos) e o discurso informal (das reuniões), tornou-se possível construir uma ontologia de palavras sobre o assunto, oferecendo várias formas de formular a mesma pergunta pertinente às consultas dermatológicas. Essas questões serão "ensinadas" para jogadores de um *Serious Game* visando o aprendizado da dermatologia. Algumas dessas questões na Língua Portuguesa estão listadas a seguir:

- *Ocorre com que frequência?*
- *Esteve em mais lugares?*
- *Você se expõe muito ao sol?*
- *Qual a sensibilidade da área da lesão?*
- *Alguém da família teve algo parecido?*
- *O tamanho da lesão mudou?*
- *A lesão possui simetria em formato de cruz?*
- *A lesão possui bordas irregulares?*
- *Possui várias tonalidades de cor?*
- *A lesão possui um brilho perolado?*
- *A mancha possui um brilho perolado?*
- *Há quanto tempo surgiu?*
- *Já furou a orelha?*
- *Está tomando algum remédio?*
- *Mora em campo ou cidade?*
- *A mancha aumentou de tamanho?*
- *A mancha é assimétrica?*
- *A lesão é normocrômica?*
- *Você se expõe muito ao sol?*
- *É recente?*

Um conjunto de regras que identifiquem referências aos vários tipos de diagnóstico e sintomas se faz necessário. Grupos gramaticais como:

- sentença;
- sujeito (e.g. *"a lesão"*);
- predicado (e.g. *"possui bordas irregulares?"*);
- frases adjetivais (e.g. *"a mancha é assimétrica?"*);
- frases preposicionais (e.g. *"a lesão possui uma simetria em forma de cruz?"*);
- frases adverbiais (e.g. *"se expõe muito ao sol?"*);
- cláusula relativa (e.g. *"faz quanto tempo que surgiu a lesão?"*);
- sentenças sim/não (e.g. *"possui várias tonalidades de cor?"*);
- sentenças *wh* (e.g. *"há quanto tempo surgiu?"*);
- sentenças alternativas na forma usual (não clivada) (e.g. *"mora em campo ou cidade?"*);
- sentenças de solicitação de explicação (e.g. *"por que a hipótese de câncer?"*);
- sentenças existenciais (e.g. *"alguém da família teve algo parecido?"*);
- sentenças na voz ativa (e.g. "*a lesão possui bordas irregulares?"*);
- sentenças na voz passiva (e.g. *"alguma região é comprometida pela lesão?"*);

[1] https://www.mongodb.com
[2] https://www.inca.gov.br
[3] http://www.sbd.org.br

- sentenças clivadas (e.g. *"é recente?"*);

A gramática proposta em [6], [8] e [9] propõe o tratamento das expressões de um vocabulário específico (palavras, expressões e jargões médicos [10]) utilizadas na radiologia.

O domínio de localidade de uma GLC é mais amplo por meio de um sistema de reescrita em árvore. Assim, as relações sintáticas descritas na GLC, apresentadas em [6], são necessárias também nas GLCLs. No entanto, essas relações são representadas por árvores (árvores iniciais e auxiliares) e operações de substituição e uma forma restrita de adjunção. Em [11] afirma-se que, em decorrência das propriedades formais da adjunção, o formalismo utilizado se torna mais poderoso que as GLCs.

Algumas das frases médicas podem ser vistas nas Figuras 1 e 2, as quais utilizam árvores elementares e itens lexicais do dicionário. As árvores que descrevem estruturas sintáticas da língua portuguesa são agrupadas em famílias, de acordo com critérios de regência verbal e transitividade [6].

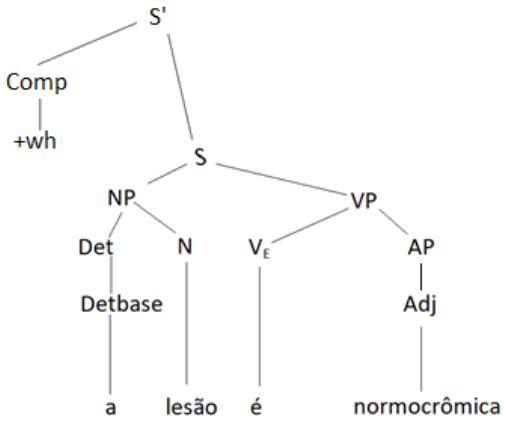

Figura 1: Substituição em a lesão é normocrômica

Até agora, as palavras coletadas eram escritas em um único arquivo de texto para o processo de "Tokenização", o qual extrai as palavras e faz a contagem de ocorrências de cada uma delas no texto. Os nomes compostos estão entre aspas para que o programa leitor de texto os reconheça (exemplo: "sulco transverso"). Todo o texto é segmentado para que as palavras sejam colocadas em uma posição de uma estrutura de dados, junto com sua quantidade de ocorrências, como atributo. Devido à simplicidade do script do programa, ele pode ser facilmente escrito em vários idiomas. A saída do programa resulta em uma lista de objetos JSON (*Javascript Object Notation*) em um arquivo JSON, o qual pode ser usado para transportar dados. Esse tipo de arquivo (JSON) foi selecionado uma vez que JSON é um formato leve e de fácil utilização tanto para leitura humana quanto para computadores, devido algumas características básicas compartilhadas.

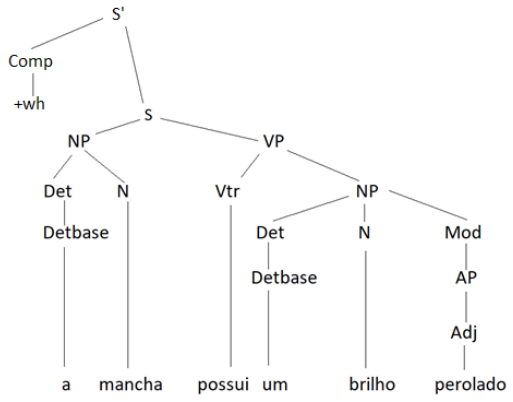

Figura 2: Substituição em a mancha possui um brilho perolado

3 CLASSIFICAÇÃO DE PALAVRAS

Neste processo as palavras foram separadas em grupos como *substantivos, verbos*, etc. Cada um desses grupos é representado como um arquivo JSON com uma estrutura particular de objeto. Por exemplo, estrutura de substantivo possui os seguintes campos: palavra (string); número (string "plu"para plural ou "sing"para singular); gênero ("masc"para masculino ou "fem"para feminino); grau (diminutivo, aumentativo ou nulo); cite_count (contagem de eventos). Todos esses são mostrados a seguir. A mesma estrutura de objeto está presente em quase todos os grupos, exceto em casos com variações diferentes como os adjetivos que têm campos adicionais. Um exemplo da estrutura é:

```
{
"palavra": "pintinhas",
"numero": "plu",
"genero": "fem",
"grau": "diminutivo",
"cite_count": 2
}
```

Os demais grupos são: Adjetivos Compostos, Advérbios, Artigos Demonstrativos, Contração, Quantificadores, Pronomes, Conjunções, Intensificadores, Preposições e Acrônimos. Considerando as classes de verbos, os dados referentes a eles foram segmentados respeitando o seguinte padrão:

- Infinitivo: radical + terminação *-ar, -er,* ou *-ir*.
- Tempo: para temporização quanto ao modo (*indicativo, imperativo* ou *subjuntivo*).
- Conjugação: identificada pelas vogais temáticas *a* (verbo na primeira conjugação), *e* (verbo na segunda conjugação), *i* (verbo na terceira conjugação).

- Pessoa: *1* (primeira), *2* (segunda), *3* (terceira).

Esse conjunto de verbos é segmentado em três seções (grupos de objetos JSON): radicais de verbos; terminações de verbos; desinência dos verbos.

- Desinência modo-temporais das formas nominais é dado pelo campo "*tempo*"que pode atribuir os seguintes valores:
  - "*ger*"(gerúndio) que tem a primeira conjugação terminada em -ando, segunda conjugação terminada em *-endo* e terceira conjugação terminada em *–indo*. Ou seja, todas terminações no gerúndio são em *ndo*;
  - "*par*"(particípio) que termina em *-do*;
  - "*inf*" (infinitivo). O atributo infinitivo armazena o verbo neste estado, que termina com *–r*.

  Por fim, foram colocados os atributos padrões já mencionados (palavra, número, gênero, pessoa);

- Radicais dos Verbos: Seus atributos são autoexplicativos (radical, verbo-regular, conjugação), mais o uso de vogal temática no campo de conjugação que vale a pena mencionar;

- Desinência de verbos: Indicam as flexões do verbo: número e pessoa, modo e tempo. Os finais são acompanhados por uma vogal temática (por exemplo, a desinência –o, presente em "mor-o", é uma desinência número –pessoal, pois indica que o verbo está na primeira pessoa do singular, "va", de "morava", é desinência modo-temporal que caracteriza uma forma verbal do pretérito imperfeito do indicativo, na 1ª conjugação. Desinências modo temporais é quando indicam os modos (indicativo, subjuntivo e imperativo) e os tempos (presente, passado e futuro) e foram avaliados neste trabalho:
  - *pr (presente do indicativo)*,
  - *prs (presente do subjuntivo)*,
  - *pti (pretérito imperfeito do indicativo)*,
  - *ptp (pretérito perfeito do indicativo)*,
  - *ptm (pretérito mais-que-perfeito do indicativo)*,
  - *pts (pretérito do subjuntivo)*,
  - *fpr (futuro do presente do indicativo)*,
  - *fpt (futuro do pretérito do indicativo)*,
  - *fs (futuro do subjuntivo)*,
  - *imp (imperativo)*.

4 INTERFACE EM LINGUAGEM NATURAL PARA BANCO DE DADOS DE CÂNCER DE PELE

Um sistema de ILNBD baseado em SQL é apresentado em [12] onde o idioma inglês foi traduzido para consultas de banco de dados. Também em [12] o tema do uso da Questão-Resposta (conhecido como QA-*Question-Answer*) foi considerado para este trabalho para construção de consultas para o MongoDB a partir de traduções para português.

A organização das palavras na estrutura do tipo JSON facilitou muito o processo de implementação do banco de dados. Devido ao tipo de banco previamente selecionado, a implementação não precisou de uma pré-modelagem, como é feita em Modelos de Entidade e Relacionamento. O banco selecionado foi o MongoDB.

MongoDB é um banco de dados não-relacional (NoSQL – *Not Only SQL*) de código aberto (open source) e multiplataforma. Sua orientação é baseada em documentos do formato JSON, o que ajuda várias aplicações a trabalhar com esses dados.

No entanto, MongoDB não possui uma modelagem de dados robusta, tem coleções voláteis e pode aceitar a adição de quaisquer dados na estrutura de um documento. Logicamente alguns dados importantes são mapeados e encapsulados. Esse mapeamento e a validação podem ser feitas na criação e consulta dos "*Collections*" (equivalentes às tabelas) dentro do banco de dados. Apesar da possibilidade de validação de campos, os *Collections* são voláteis, podendo aceitar a adição de qualquer outro dado em um *Document* (equivalente à tupla). Isso faz com que o banco se modele de acordo com as necessidades do software que o utiliza. NoSQL também suporta relacionamentos, mas de uma forma um pouco diferente. Portanto, como primeiro passo nessa metodologia, foi adotada a modelagem de dados da própria aplicação. Seus recursos atendem muitos aspectos consideráveis para superar o modelo clássico em vários tipos de aplicações. No caso do MongoDB, as grandes taxas de escritas são processadas com baixo custo, porém com menor segurança. Esse possui uma documentação[4] detalhada sobre suporte técnico, instalação e tutoriais de comandos e consultas.

Além do alto desempenho de consulta para dados massivos, o MongoDB possui um mecanismo de busca distinto, que permite construir expressões regulares e a utilização de funções como "$lookup", "$count", "$Match", "$group", "$project", "$sort", entre outras.

Esses comandos podem ser utilizados diretamente no código da consulta. Por exemplo, poderíamos fazer consultas com expressão regular nos radicais dos verbos, como no exemplo:

```
db.verbos.find(
{'$and':[
{radical: { "$regex": "String do Radical", "$options": "i"}}
]}).pretty()
```

As maiores vantagens do MongoDB, além da facilidade de trabalhar com uma carga imensa de dados são: Processo Sharding (divide os dados entre máquinas), Replicações simples, SchemaFree, Performance e uso do GridFS para gravar arquivos. Porém, como desvantagem sobre o MySQL temos o uso elevado de memória RAM e de espaço de armazenamento [13].

As questões que foram mapeadas são diferentes das do tipo "A lesão possui bordas irregulares?"e "O tamanho da lesão mudou?"(que respondem apenas sim/não), uma vez que as consultas à base de dados são o tema de discussão. Por exemplo, a frase "Quais lesões possuem bordas irregulares?" ou "Dê-me as lesões que são totalmente assimétricas, porém possuem bordas regulares".

A estratégia inicial consiste na aplicação de um dicionário para ambos os propósitos: um referente ao diálogo e à construção narrativa para o jogo e o outro para as consultas à base de dados. A Figura 3 apresenta o projeto do Processador de Linguagem Natural proposto neste trabalho, o qual foi baseado em [12] e [14].

[4] https://docs.mongodb.com/manual/installation/

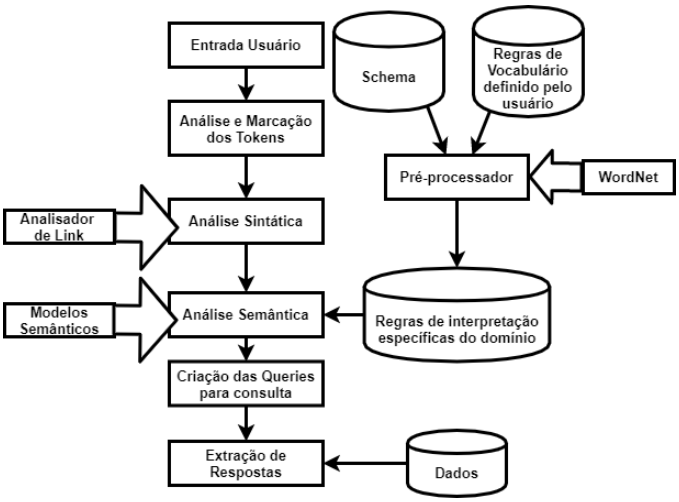

Figura 3: O projeto de um Processador de Língua Natural.

Um sistema de ILNBD baseado em SQL é apresentado em [12], onde o idioma inglês foi usado para consultas ao banco de dados. O resultado do processo pode talvez assemelhar-se a este trabalho, considerando que a frase (input) é convertida para os comandos de consulta como "$lookup", "$Match"e outros.

No referido projeto, durante a análise da imagem da lesão, como da presente na Figura 4, o usuário poderá utilizar uma caixa de diálogo para fazer perguntas ao paciente. Todas as informações a respeito da imagem e do paciente serão adquiridas do próprio banco de imagens.

Alguns atributos básicos foram considerados como importantes ao banco de lesões e são “member” (parte do corpo com a lesão), “benign_malignant” (benigno ou maligno), “diagnosis” (diagnóstico) e “diagnosis_type” (tipo de diagnóstico). Apenas o atributo “behaviour” (comportamento) ainda não foi testado para a ILNBD.

As imagens utilizadas neste trabalho foram coletadas de nove bancos de dados com imagens de lesões de pele do tipo melanoma e não melanoma, dos quais apenas cinco possuem acesso livre, requisitando apenas a referência do uso do banco de dados. Não foi encontrado nenhum banco de imagens brasileiro. Um exemplo de imagem de melanoma (à esquerda) e pinta (à direita) é demonstrado na Figura 5.

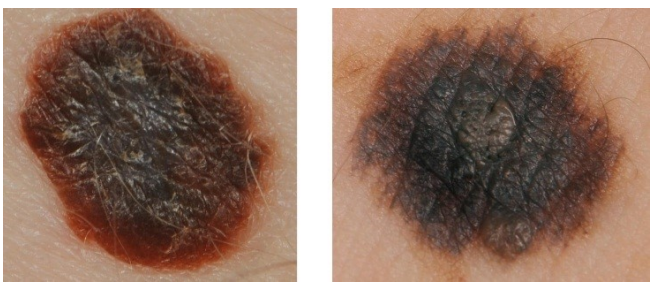

Figura 5: Melanoma à esquerda e Pinta à Direita.

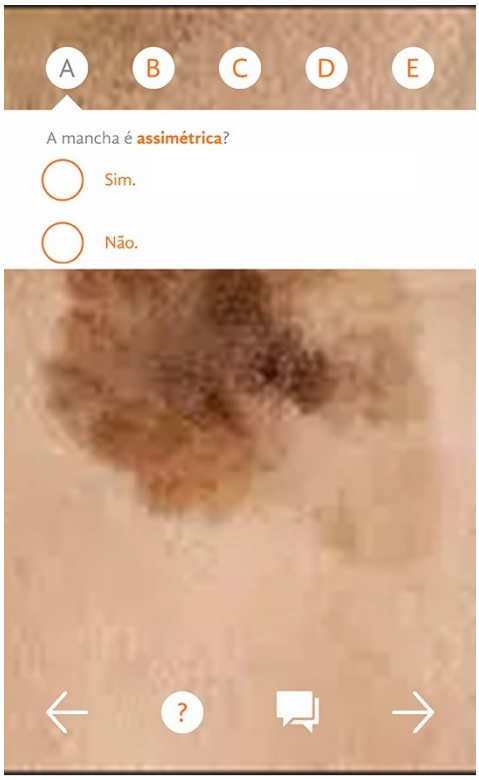

Figura 4: Ambiente 2D para examinar a imagem da lesão.

Com base na Figura 5 é possível notar características, as quais foram previamente perguntadas, como simetria, brilho perolado, característica da borda da lesão, etc.

A primeira base ISIC-ARCHIVE[5] possui 13.791 imagens com descrições bem detalhadas sobre a lesão e o paciente. Já a segunda base é do Departamento de Dermatologia da Universidade do Centro Médico de Groningen (UMCG)[6] e possui 70 imagens de melanoma e 100 imagens de pintas. A terceira base é a DermNet NZ[7], a qual não foi possível obter uma contagem exata, contudo, possui diversas imagens categorizadas em vários tipos como "melanoma", "tumores", "acne", entre vários outros. A quarta base é chamada Dermnet - Skin Disease Atlas[8], e como a terceira base de dados não forneceu uma quantidade exata de dados, porém existem várias imagens padronizadas e com poucos detalhes. A quinta base de dados, chamada de ADDI[9] (Automatic Computer-Based Diagnosis System for Dermoscopy Images), além de ser uma base específica para pesquisas, é de nacionalidade portuguesa, o que facilita a leitura.

5 CONCLUSÕES

Esta solução pode atingir maiores proporções quando se trata de análise semântica de texto, como em [14], ou geração de conteúdo para o jogo, utilizando abordagens como as apresentadas em [15].

[5] https://isic-archive.com/#images
[6] http://www.cs.rug.nl/~imaging/databases/melanoma_naevi/
[7] https://www.dermnetnz.org/
[8] http://www.dermnet.com/
[9] http://www.fc.up.pt/addi/index.html

No total foram adquiridas 3009 palavras. Espera-se melhorar o dicionário dermatológico durante o desenvolvimento do software, bem como questões que envolvem o PLN, como a análise semântica.

Além disso, uma vez que a ILNBD esteja completa, a área de dermatologia brasileira terá duas novas bases de dados, uma com imagens de lesões e descrições de diagnósticos e outra com uma ontologia de palavras sobre o câncer de pele. Essas estarão disponíveis ao público e seus dados serão utilizados no decorrer deste trabalho.

Infelizmente não foi possível obter descrições diagnósticas sobre lesões cutâneas a partir do hospital. Além disso, nenhuma base de dados brasileira com imagens e textos sobre lesões foi encontrada. Portanto, o trabalho final também visa criar um banco de dados, alimentada por uma comunidade de dermatologistas e jogadores do jogo por meio de um fórum interno. Uma vez que tal jogo utiliza-se de uma aplicação que implementa um processador de língua natural, isso torna-se o processo mais complexo, conforme demonstrado por [8].

REFERÊNCIAS

[1] Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard De Melo, Livy Real, and Anne Silva. Extending a lexicon of portuguese nominalizations with data from corpora. In *Proceedings of 11th International Conference on Computational Processing of the Portuguese Language*, (PROPOR '14), pages 114–124. Springer, Lecture Notes in Computer Science, 2014.

[2] Bill Z Manaris and Brian M Slator. Interactive natural language processing: building on success. *Computer*, (7):28–32, 1996.

[3] Yves Schabes and Richard C. Waters. Lexicalized context-free grammar: A cubic-time parsable, lexicalized normal form for context-free grammar that preserves tree structure. *Mitsubishi Electric Research Laboratories. Technical Report*, page 30, 1993.

[4] Aravind K Joshi and Yves Schabes. Tree-adjoining grammars and lexicalized grammars. *Proceedings of 4th European Summer School in Logic, Language and Information*, pages 1–23, 1991.

[5] Yves Schabes and Richard C. Waters. Lexicalized context-free grammars. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, page 121–129, USA, 1993. Association for Computational Linguistics (ACL '93). doi: 10.3115/981574.981591. URL https://doi.org/10.3115/981574.981591.

[6] Cinthyan Renata Sachs Camerlengo de Barbosa. Gramática para consultas radiológicas em língua portuguesa. Master's thesis, Instituto de Informática da Universidade Federal do Rio Grande do Sul., 1998.

[7] Sebastião A. P. Sampaio and Evandro A. Rivitti. Alterações na pele do idoso. *Dermatologia. 3ª. ed.*, pages 1313–22, 2007.

[8] Cinthyan Renata Sachs Camerlengo de Barbosa and José Mauro Volkmer de Castilho. Gramática Livre de Contexto Lexicalizada para a Análise Sintática da Língua Portuguesa: uma experiência na Geração de Consultas de uma Interface em Linguagem Natural para Banco de Dados. *Anais do 5º Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, pages 155–164, 2000.

[9] Cinthyan Renata Sachs Camerlengo de Barbosa, Davidson Cury, José Mauro Volkmer de Castilho, and Celso de Renna e Souza. Defining a lexicalized context-free grammar for a subdomain of Portuguese language. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks*, (TAG+6), pages 74–79, Universitá di Venezia, May 2002. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W02-2210.

[10] RH Baud, A-M Rassinoux, and J-R Scherrer. Natural language processing and semantical representation of medical texts. *Methods of information in medicine*, 31(02):117–125, 1992.

[11] Anne Abeillé and Yves Schabes. *Non-compositional discontinuous constituents in Tree Adjoining Grammar*, volume 6. Walter de Gruyter, 1996.

[12] Niculae Stratica and Bipin C. Desai. Schema-based natural language semantic mapping. In Farid Meziane and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pages 103–113, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-27779-8.

[13] Cornelia Győrödi, Robert Győrödi, George Pecherle, and Andrada Olah. A comparative study: MongoDB vs. MySQL. In *Proceedings of 13th International Conference on Engineering of Modern Electric Systems (EMES'15)*, pages 1–6, June 2015. doi: 10.1109/EMES.2015.7158433.

[14] Niculae Stratica, Leila Kosseim, and Bipin C. Desai. NLIDB templates for semantic parsing. In Antje Düsterhöft and Bernhard Thalheim, editors, *Natural language processing and information systems*, pages 235–241, Bonn, 2003. Gesellschaft für Informatik e.V.

[15] Tess Crosbie, Tim French, and Marc Conrad. Towards a model for replicating aesthetic literary appreciation. In *Proceedings of the Fifth Workshop on Semantic Web Information Management*, (SWIM '13), New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321945. doi: 10.1145/2484712.2484720. URL https://doi.org/10.1145/2484712.2484720.