DOI: 10.17851/2237-2083.23.3.599-640

Extração de relações hiponímicas em um *corpus* de língua portuguesa

Hyponym relation extraction in a Portuguese language corpus

Pablo Neves Machado
Pontifícia Universidade Católica (PUC-RS), Porto Alegre, RS, Brasil
machadoum@gmail.com

Vera Lúcia Strube de Lima
Pontifícia Universidade Católica (PUC-RS), Porto Alegre, RS, Brasil
vera.strube@pucrs.br

Resumo: As relações hiponímicas são importantes na construção de estruturas de conhecimento, tais como ontologias ou taxonomias, para melhorar o processo de busca. O presente trabalho estuda em detalhe padrões para extração de relações hiponímicas com base em um *corpus* de língua portuguesa. Para tanto, toma como base os padrões específicos propostos por Hearst (1992), Freitas e Quental (2007) e Taba e Caseli (2014). Constrói, a partir desses padrões, regras que alimentam um protótipo, o qual as aplica a um *corpus* e extrai, como resultado, relações hiponímicas. Avaliadores humanos avaliam as relações extraídas, utilizando a escala proposta por Freitas e Quental. A precisão das extrações é compatível com as da literatura. O trabalho ainda apresenta um minucioso estudo quanto à produtividade dos padrões e quanto à avaliação das extrações.

Palavras-chave: relações hiponímicas; extração de relações; padrões de extração de relações.

Abstract: Hyponym relations are important in the building of knowledge structures such as ontologies or taxonomies to enhance the search process. This work studies patterns for extraction of hyponym relations from a specific Portuguese language corpus. It starts from selected patterns proposed by Hearst (1992), Freitas and Quental (2007), and Taba and Caseli (2014). From these

patterns, it builds up rules that feed a prototype. This prototype applies them to the corpus and extracts, as a result, hyponym relations. Human evaluators assess the extracted relations, using the scale proposed by Freitas and Quental. Precision of the extractions is consistent with the literature. The paper also describes a detailed study on the productivity of the patterns and the assessment of the extractions.

Keywords: hyponym relations; relation extraction; relation extraction patterns

Recebido em: 29 de julho de 2015.
Aprovado em: 26 de outubro de 2015.

1 Introdução

O Processamento de Língua Natural (PLN) é área destacada por sua relevância no que tange ao processamento de grandes quantidades de documentos. O crescimento rápido da *World Wide Web* teve como consequência um desafio na compreensão do conteúdo das informações. Hoje, o acesso à informação na *web* é realizado, prioritariamente, por meio da busca de palavras-chave, e essa busca é realizada por mecanismos de comparação lexical. Devido ao gigantesco tamanho que a *web* apresenta atualmente e a sua contínua expansão, quando são realizadas buscas de palavras-chave, diversos conteúdos irrelevantes para o usuário são encontrados. Algumas fontes de dados manualmente estruturadas foram surgindo, mas, devido à grande quantidade de conteúdo existente na rede, fica evidente a importância de ferramentas para extração da informação disponível em língua natural.

Entre as informações a ser extraídas de textos, revestem-se de especial importância as relações hiponímicas, as quais podem ser usadas na construção de estruturas ontológicas e outras estruturas de representação do conhecimento.

O presente trabalho estuda a extração de relações com base em textos escritos em língua portuguesa. A abordagem inicial baseia-se na organização das contribuições presentes nos trabalhos de Marti Hearst (1992), Freitas e Quental (2007) e Taba e Caseli (2014), mas a arquitetura da solução e a prototipação seguem organização própria.

Também são aproveitados outros estudos, como o de Baségio (2007), que realizou a adaptação de padrões da língua inglesa para a língua portuguesa. Foi desenvolvido um protótipo de extração de relações hiponímicas de *corpus* em língua portuguesa, o qual permitiu a realização de experimentos com diferentes conjuntos de padrões e regras de extração. Os resultados obtidos com a execução do protótipo são analisados, avaliados e comparados com os registrados na literatura. Além disso, discute-se o processo de avaliação manual e analisa-se os erros frequentes.

O texto do trabalho está organizado em cinco seções, incluindo esta introdução. A seção 2 aborda, brevemente, a temática e os trabalhos correlatos ao presente estudo. A seção 3 descreve o trabalho realizado, com o detalhamento das regras de extração, as suas fontes e o processo de extração de relações. A seção 4 trata da avaliação e discussão dos resultados, incluindo uma análise dos erros encontrados. A seção 5 traz um fechamento do trabalho.

2 Extração de relações semânticas e trabalhos correlatos

2.1 A extração de relações

De acordo com o indicado no estudo de Jurafsky e Martin (2009), o significado de uma palavra pode ser expresso como sendo a sua relação com outras palavras. Tal como proposta por Gruber (1992), uma relação é um conjunto de n-uplas que representam um relacionamento entre objetos no universo do discurso. Cada n-upla é uma sequência finita e ordenada de objetos. Nesse contexto, a n-upla pode ser representada pela expressão (nome-da-relação, *arg*1, *arg*2,…, *arg*n), na qual *arg*1 é um objeto na n-upla. Um exemplo de formato de uma relação binária pode ser dado por (*arg*1, nome-da-relação, *arg*2), com pequena alteração na ordem dos componentes da n-upla.

Entre as relações semânticas, especialmente interessantes são as relações hierárquicas representadas pela hiperonímia e hiponímia. A hiperonímia expressa uma relação de significado geral, enquanto a hiponímia representa um significado hierárquico restrito. Alguns

exemplos de relações citadas nesta seção podem ser vistos no Quadro 1. No presente trabalho, focamos nas relações hiponímicas binárias, que representamos por "Hiponímia (*arg*1, *arg*2)".

Quadro 1 – Exemplos de relações semânticas

Nome da relação	**arg1**	**arg2**
Sinonímia	claro	alvo
Antonímia	claro	escuro
Hiperonímia	animal	cachorro
Hiponímia	cachorro	animal

Fonte: nossa autoria.

Existem outras relações semânticas que ligam argumentos no texto, verbais ou não verbais. Por exemplo, da oração "Alexandre adora fritas", pode ser extraída a n-upla (adora, Alexandre, fritas), na qual "Alexandre" e "fritas" são argumentos e "adora" representa a relação.

Os primeiros trabalhos relacionados à extração automática de relações abordam, principalmente, relações hiponímicas e meronímicas. Isso se deve ao fato de essas relações serem a base para a construção de ontologias. As relações hiponímicas são comumente representadas por "é um" e expressam relações entre instâncias e classes, como também entre classes. Mais recentemente as pesquisas nessa área têm incluído também as relações verbais, em geral representadas por verbos e seus argumentos.

Esses modelos de relação podem ser extraídos de textos em língua natural com base no processamento de *corpora*. Para Banko e coautores (2007), os sistemas de extração de relações normalmente buscam satisfazer determinadas demandas previamente especificadas, por exemplo, extrair o local e horário de um evento por meio de um conjunto de anúncios. Quando se tem de extrair relações de um novo domínio, costuma ser necessário um retrabalho. Pode ser preciso refazer algumas das tarefas, como o estabelecimento da heurística empregada na extração e a etiquetagem de um novo *corpus* de treino.

No atual estado da arte, existem trabalhos, principalmente para a língua inglesa, que abordam o tema da extração de relações. Há também ferramentas e recursos disponíveis que são interessantes. Introduzimos, a seguir, alguns desses trabalhos referentes ao tema. Entre as abordagens que estudam a extração de relações em *corpora* textuais, duas são as mais comuns: o aprendizado de máquina e a extração baseada em regras. Na exposição que segue, é dada ênfase maior à segunda abordagem.

2.2 Trabalhos com foco em língua estrangeira

Hearst (1992) propõe um método de aquisição de relações hiponímicas entre sintagmas nominais para a língua inglesa, com base em seis padrões simples que podem ser encontrados com frequência em textos. Esses padrões podem ser vistos no Quadro 2.

Quadro 2 – Padrões apresentados por Hearst

i.	NP such as {NP ,}* {(or | and)} NP
ii.	such NP as {NP ,}* {(or | and)} NP
iii.	NP {, NP}* {,} or other NP
iv.	NP {, NP}*{,} and other NP
v.	NP {,} including {NP ,}* {or | and} NP
vi.	NP {,} especially {NP ,}* {or | and} NP

Fonte: Hearst (1992).

Um dos objetivos que conduziu Hearst a essa abordagem foi criar um método aplicável a grandes quantidades de textos. A importância do trabalho de Hearst se deve ao fato de ser um dos primeiros a propor padrões lexicais na extração de relações semânticas. Um exemplo é aquele em que a autora mostra uma aplicação do padrão (vi) (HEARST, 1992):

> "... most European countries, especially France, England and Spain."

Aplicando o padrão "NP {,} especially {NP ,}* {or | and} NP", apresentado como (vi) no Quadro 2, no qual NP é um sintagma nominal (*noun phrase*), as seguintes relações são extraídas:

Hiponímia ("France", "European country")
Hiponímia ("England", "European country")
Hiponímia ("Spain", "European country")

Hearst aplicou seus padrões em *corpora* enciclopédicos e jornalísticos, avaliando que 63% das relações identificadas eram de boa qualidade.

Os padrões textuais criados por Hearst são utilizados em diversos trabalhos, por exemplo, os de Maedche e Staab (2002), Cederberg e Widdows (2003), Degeratu e Hatzivassiloglou (2004), Baségio (2007) e Freitas e Quental (2007).

Cederberg e Widdows (2003), por exemplo, utilizam os padrões propostos por Hearst com um método denominado *Latent Semantic Analysis* (LSA), para filtrar as relações incorretas, aumentando a precisão das extrações em 30%. Relações corretamente extraídas podem ser usadas como "sementes" para a extração de diversas outras relações, aumentando, assim, a cobertura.

Morin e Jacquemin (2003) apresentam padrões para a aquisição de relações hiponímicas em *corpora* de língua francesa, tal como pode ser visto no Quadro 3. O exemplo a seguir, dado pelos mesmos autores, mostra como esses padrões se comportam.

Se o padrão "{deux|trois...|2|3|4...} NP1 (LIST2)" é aplicado ao trecho:

"... analyse foliaire de quatre espèces ligneuses (chêne, frêne, lierre et cornouiller) dans...",

é possível identificar as seguintes relações:

Hiponímia ("chêne", "espèce ligneux")
Hiponímia ("frêne", "espèce ligneux")

Hiponímia (“lierre”, “espèce ligneux”)
Hiponímia (“cornouiller”, “espèce ligneux”)

Quadro 3 – Padrões para a língua francesa propostos por Morin e Jacquemin

i.	deux|trois...|2|3|4...} NP1 (LIST2)
ii.	{certain|quelque|de autre...} NP1 (LIST2)
iii.	{deux|trois...|2|3|4...} NP1: LIST2
iv.	{certain|quelque|de autre...} NP1: LIST2
v.	{de autre} NP1 tel que LIST2
vi.	NP1, particulièrement NP2
vii.	{de autre} NP1 comme LIST2
viii.	NP1 tel LIST2
ix.	NP2 {et|ou} de autre NP1
x.	NP1 et notamment NP2

Fonte: Morin e Jacquemin (2003).

Uma proposta mais recente para a extração de relações é a denominada *Open Information Extraction* (OpenIE), que visa à extração aberta e em grande escala, sem se preocupar em tipificar as relações extraídas. Nessa proposta, Corro e Gemulla (2013) apresentam o sistema ClausIE (*Clause-based Open Information Extraction*). Os experimentos de tais autores sugerem que o sistema obtenha os melhores resultados, tornando-se referência na OpenIE, dada a característica de utilizar uma abordagem baseada em cláusulas (orações), de mais forte cunho linguístico. O sistema ClausIE identifica conjuntos de orações e o seu tipo (de acordo com a função gramatical do conteúdo), sendo baseado em um *parser* (“analisador sintático”) de dependências e em um pequeno conjunto de léxicos independentes de domínio. Essa organização permite ao sistema o processamento em paralelo e, desse modo, o processamento de grandes coleções de conteúdo de maneira escalável. Assim como

ocorre no presente trabalho, ClausIE não necessita de pós-processamento e de dados de treino para sua execução. Contudo, os autores reportam que as principais incorreções nas relações extraídas são provenientes de erros de *parsing* ("análise sintática").

Em seu trabalho, Gamallo e coautores (2012) descrevem um método que igualmente utiliza o paradigma OpenIE para a extração de triplas baseadas em verbos de *corpora* multilíngues. O método extrai relações em *corpora* nos idiomas português, inglês, espanhol e galego. Segundo os autores, o método apresenta resultados superiores aos alcançados pelos trabalhos no estado da arte, devido principalmente ao fato de utilizar análise sintática profunda e um tokenizador[1] robusto e rápido.

2.3 Trabalhos com foco na língua portuguesa

O *software* PALAVRAS, proposto e desenvolvido por Bick (2000), reúne diversas ferramentas para o processamento de línguas naturais que aceitam como entrada textos em língua portuguesa, e pode ser utilizado para etiquetagem de *corpus*, processamento léxico-morfológico, geração de árvores sintáticas e reconhecimento de entidades nomeadas, entre outros. É relatada precisão maior que 97%, tanto em termos de morfologia quanto de sintaxe. O *parser* PALAVRAS é baseado em regras e está disponível por meio do projeto VISL.[2]

Freitas e Quental (2007) adaptam dois padrões de Hearst para a língua portuguesa ("such as" e "and/or others") e criam outros três padrões, com base em análise de ocorrências no texto, capazes de detectar relações hiponímicas. O trabalho utiliza o *parser* PALAVRAS para a identificação de sintagmas nominais. As regras foram aplicadas ao *corpus* CORSA (*CORpus* da SAúde Pública), que contém cerca de 2 milhões de palavras. Os resultados foram compatíveis com os de Hearst, mostrando um percentual de 73% de relações consideradas de boa qualidade. O Quadro 4 ilustra os dois padrões de Hearst adaptados por

---

1 Será usado o termo "tokenizador" em lugar de "*tokenizer*", por ser encontrado em outros trabalhos da área em língua portuguesa.

2 Sítio *web* do projeto VISL: <http://beta.visl.sdu.dk>. Acesso em: 26 jul. 2015.

Freitas e Quental (i(a), i(b) e ii), assim como os três padrões propostos pelas autoras (iii, iv e v).

Quadro 4 – Padrões extraídos de Freitas e Quental (2007)

i(a)	SN HHiper (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) Sni
i(b)	SN Hiper, (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) Sni
ii	SN HHipo { ,SN Hipoi } * { , } e|ou outros SN Hiper
iii	tipos de SN Hiper: SN1 { , SN2 ... ,} (e | ou) Sni
iv	SN HHiper chamado/s/a/as (de) SN Hipo
v	SN Hiper conhecido/s/a/as como SN Hipo

Fonte: Freitas e Quental (2007), adaptado de Hearst (1992).

O excerto de texto apresentado por Freitas (2007, p. 76) e reproduzido a seguir mostra a aplicação do padrão (iv):

> "e nele existe uma [substância] chamada [benzopireno]."

Com a aplicação, a seguinte relação deve ser extraída: "Hiponímia (benzopireno, substância)". Como exposto por Freitas, HHiper representa a configuração na qual o termo hiperônimo é o primeiro substantivo à esquerda.

O trabalho de Baségio (2007), da mesma época que o de Freitas e Quental, visualiza a construção semiautomática de ontologias com base em textos na língua portuguesa do Brasil. Para esse fim, o trabalho emprega uma abordagem que inclui extração de relações hiponímicas. O autor traduziu, para a língua portuguesa do Brasil, relações propostas em outros trabalhos consolidados, como, principalmente, o de Hearst (1992), o que pode ser visto no Quadro 5.

Quadro 5 – Padrões de Hearst adaptados por Baségio (2007)

NP such as {(NP,)*(or|and)} NP	SUB como {(SUB,)*(ou|e)} SUB SUB tal(is) como {(SUB,)*(ou|e)} SUB
NP such as {(NP,)*(or|and)} NP	SUB como {(SUB,)*(ou|e)} SUB
such NP as {(NP,)*(or|and)} NP	tal(is) SUB como {(SUB,)*(ou|e)} SUB
NP {, NP}* {,} or other NP	SUB {, SUB}* {,} ou outro(s) SUB
NP {, NP}* {,} and other NP	SUB {, SUB}* {,} e outro(s) SUB
NP {,} including {NP,}*{or|and} NP	SUB {,} incluindo {SUB,}*{ou|e} SUB
NP {,} especially {NP,}*{or|and} NP	SUB {,} especialmente {SUB,}*{ou|e} SUB SUB {,} principalmente {SUB,}*{ou|e} SUB SUB {,} particularmente {SUB,}*{ou|e} SUB SUB {,} em especial { SUB,}*{ou|e} SUB SUB {,} em particular { SUB,}*{ou|e} SUB SUB {,} de maneira especial { SUB,}*{ou|e} SUB SUB {,} sobretudo { SUB,}*{ou|e} SUB

Fonte: Baségio (2007).

Para obter resultados mais efetivos, Baségio implementou um processo de remoção de palavras entendidas como pouco relevantes para o domínio, fixando-se nos substantivos, como se observa nas regras propostas pelo autor. Esse processo removeu cerca de 70% das palavras analisadas. Observa-se, contudo, que a contribuição do trabalho de

Baségio reside mais no estudo realizado e na transposição de regras para o português do que na análise da produtividade das regras empregadas.

Batista e coautores (2013) trazem para a língua portuguesa uma proposta de uso de aprendizagem de máquina voltada à classificação de relações, mais do que à extração propriamente dita. O trabalho faz uso de métodos de aprendizagem semisupervisionada para identificar o tipo das relações, em vez de usar regras específicas e palavras-chave como ocorre nos estudos com regras de extração. Para tanto, emprega a similaridade entre elementos de um conjunto e, assim, identifica grupos de elementos similares entre exemplares de triplas representando relações. Nos experimentos relatados pelos autores, essas triplas correspondem a frases extraídas da Wikipedia que expressam relações entre pares de entidades extraídas da DBPedia. O algoritmo *k-nearest neighbors* (*k*-vizinhos mais próximos) é usado para construir os grupos, com base, principalmente, nas classes gramaticais das palavras que ocorrem antes, depois e entre duas entidades mencionadas. Embora o foco do trabalho e dos experimentos relatados esteja na classificação de relações entre entidades mencionadas, sem identificar relações hiponímicas, a abordagem descortina novas alternativas para estudos na área para a língua portuguesa, que efetivamente vêm a ser desenvolvidos por Taba e Caseli.

Em seus estudos, Taba e Caseli (2014) fazem uso de múltiplas abordagens de aprendizagem de máquina, bem como de regras, na extração automática de relações semânticas em textos em português. As abordagens são comparadas em diferentes experimentos, com utilização de dois *corpora* anotados pelo *parser* PALAVRAS. O primeiro, o CETENFolha, *corpus* de caráter jornalístico, é composto por 24 milhões de palavras de artigos do jornal Folha de São Paulo; o segundo, de caráter científico, conta com 870 mil palavras e é proveniente de textos de uma revista de divulgação científica (FAPESP). Os autores buscam extrair as seguintes relações:

- is-a
- part-of
- location-of
- effect-of
- property-of
- made-of

- used-for

No que tange à relação *is-a*, Taba e Caseli (2014) empregaram padrões provenientes de Hearst (1992) e de Freitas e Quental (2007), além de introduzir novos padrões manualmente definidos, que podem ser vistos no Quadro 6, no qual o termo T1 representa o hiperônimo de uma relação e os termos T2, T3 representam possíveis hipônimos.

Quadro 6 – Padrões para a relação *is-a* conforme Taba e Caseli (2014)

i	T1 (tais como|como) T2 {, T3}* (e|ou) TN
ii	T2 {, T3}* ,? (e|ou) outros T1
iii	tipos de T1: T2 {, T3}* (e|ou) TN
iv	T1 chamad(o|a|os|as) de? T2
v	T2 {, T3}* ,? (e|ou) (qualquer|quaisquer) outro{s} T1
vi	T2 é (o|a|um|uma) T1
vii	T2 são T1

Fonte: Taba e Caseli (2014).

Os autores relatam quatro experimentos: o primeiro com o método baseado em regras, e os três outros usando aprendizagem de máquina para a extração das relações. No primeiro caso, a precisão reportada para as relações *is-a* é de 61,1%, porém com muito baixa cobertura, de 1,2%. Os experimentos com aprendizagem de máquina visam, além dos resultados das extrações, estudar a influência das informações sintáticas, morfológicas ou superficiais no processo de extração. Os resultados apresentados mostram que o aprendizado de máquina pode trazer aumentos significativos na cobertura. A técnica conhecida como máquinas de vetores de suporte alcança a melhor precisão para as relações *is-a*, chegando a 78,2% nos *corpora* utilizados. A comparação de tais resultados com outros disponíveis na literatura não é possível devido às diferenças nos *corpora* e *parsers*, entre outras. Mas a técnica revela-se bastante promissora.

3 O trabalho realizado

Visando a extrair relações hiponímicas por meio de *corpora* de língua portuguesa, foram realizadas adaptações de padrões propostos pelos autores estudados, o que gerou regras. As regras foram inseridas em um programa de computador desenvolvido especialmente para esse fim.

O formato das regras segue o das expressões regulares, de uso corrente nas Ciências da Computação. As regras são escritas em linguagem formal e são interpretadas por um processador de expressões regulares, tal como ocorre no presente trabalho. O objetivo de uma regra é prover uma forma concisa e flexível de representar um padrão que identifica cadeias de caracteres, ou caracteres de interesse. A linguagem formal adotada representa sintagmas nominais por "SN" e utiliza os parênteses para agrupar as expressões; o símbolo "*" indica que uma expressão pode não ocorrer, pode ocorrer uma ou mais vezes. A interrogação significa nenhuma ou uma repetição. Outro símbolo comumente utilizado é a barra vertical "|", que representa um "ou exclusivo". Também foi utilizada a notação "<sn PALAVRA-CHAVE sn>" para identificar a ocorrência de uma palavra-chave que está contida dentro de um *chunk* ("porção de texto, sintaticamente analisada"). Como um Sintagma Nominal pode ser formado por outros SNs, o símbolo "sn" (minúsculo) foi empregado para identificar um Sintagma Nominal que é um dos elementos de um "SN". Caso a palavra-chave se encontre logo após o símbolo "<" ou antes do símbolo ">", significa que ela é, respectivamente, a primeira ou a última palavra do *chunk*, como em: "<outros sn>" (a palavra-chave é representada por "outros"). O Quadro 7 relaciona as regras definidas neste trabalho e empregadas no estudo realizado. Todas se originam de padrões específicos, compilados da literatura estudada, com breves adaptações e melhorias introduzidas. A numeração, de 1 a 11, visa a facilitar sua exposição; observe-se que a regra 10 subdivide-se em duas, 10A e 10B, levando a um total de doze regras implementadas.

Quadro 7 – Regras de extração empregadas no presente trabalho

Regra	Descrição
1	SN(,)? como (SN ,)*(SN (e|ou))*SN
2	SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
3	SN(,)? incluindo (SN ,)*(SN (e|ou))*SN
4	SN(,)? especialmente (SN ,)*(SN (e|ou))*SN
5	(SN (ou|e|,))*<outr(a|o)(s)? sn>
6	<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN
7	SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN
8	SN((,)? também)?(,|é|são|foram)? conhecid(o|a|os|as) como (SN ,)*SN (e|ou))*SN
9	(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>
10A	SN é < (o|a) sn>
10B	SN é < (um|uma) sn
11	SN são SN

Fonte: nossa autoria.

As seções 3.1 a 3.4 detalham essas regras, sua origem e sua aplicação.

3.1 Extração com padrões de Hearst

Os padrões propostos por Hearst (QUADRO 2) foram criados com o intuito de extrair relações hiponímicas em *corpus* de língua inglesa. Para a utilização desses padrões na língua portuguesa do Brasil, partiu-se de trabalhos prévios de tradução e contextualização de tais relações.

Por exemplo: dado o excerto de texto: “Países como o Brasil, Equador e os EUA”, o padrão SN(,)? como (SN ,)*(SN (e|ou))*SN pode extrair as seguintes relações: Hiponímia (Brasil, País), Hiponímia (Equador, País) e Hiponímia (EUA, País).

O padrão em questão é o referente ao “such as”, proveniente dos estudos de Hearst. Este corresponderia ao “como” em português, que pode exercer diversas funções sintáticas em uma sentença, o que causa dificuldade em obter altos níveis de precisão, como já indicado por Freitas e Quental (2007). Baségio (2007) também havia empreendido esforços para adaptar o padrão “como” para a língua portuguesa. Entretanto considerava apenas substantivos, simplificando a ideia de sintagma nominal presente nos padrões de Hearst. Em nosso trabalho, escolhemos utilizar SNs, empregando padrões mais complexos. Já na adaptação por Freitas e Quental, foram utilizadas regras levando em conta a existência de SNs, mas ocorreu, assim como no caso de Baségio, a flexibilização, nesse caso, visando ao uso apenas da palavra mais à direita dentro do sintagma nominal.

Uma melhoria introduzida em relação aos trabalhos de Freitas e Quental (2007), e detalhada mais adiante, foi o tratamento da vírgula, que pode ocorrer antes da palavra “como”, por exemplo, em:

“... [outras falhas], como [dois nomes para um mesmo fator]...”

Essa alteração aumentou em torno de 40% o número de relações extraídas com o padrão “como” em relação aos resultados anteriores. Utilizando uma abordagem semelhante, foi possível adaptar os padrões de Hearst identificados no Quadro 2, como (ii), (v) e (vi):

SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
SN(,)? incluindo (SN ,)*(SN (e|ou))*SN
SN(,)? especialmente (SN ,)*(SN (e|ou))*SN

Já o padrão a seguir, inspirado nos padrões (iii) e (iv) de Hearst, precisou de uma implementação alternativa:

(SN (ou|e|,))*<outr(a|o)(s)? sn>

O *parser* PALAVRAS, ao processar um texto como "Brasil, Equador, EUA e outros países", identifica diversos SNs, um dos quais inclui o determinante "outros":

"[Brasil], [Equador], [EUA] e [outros países]"

O presente trabalho propõe uma adaptação para encontrar SNs nessa situação. Com tal adaptação, as relações que podem ser extraídas com o padrão para o texto do exemplo são: Hiponímia (Brasil, País), Hiponímia (Equador, País), Hiponímia (EUA, País). O Quadro 8 associa os padrões propostos por Hearst com as regras propostas neste trabalho e apresentadas no Quadro 7. Pode-se observar que a regra 5 foi utilizada para expressar dois padrões propostos.

Quadro 8 – Associação entre padrões de Hearst e regras propostas neste trabalho

Regra	Padrão de Hearst
1	NP such as {NP ,}* {(or | and)} NP
2	such NP as {NP ,}* {(or | and)} NP
3	NP {,} including {NP ,}* {or | and} NP
4	NP {,} especially {NP ,}* {or | and} NP
5	NP {, NP}* {,} or other NP
	NP {, NP}*{,} and other NP

Fonte: nossa autoria, com base em Hearst (1992).

3.2 Extração com padrões de Freitas e Quental

Os padrões (iii), (iv) e (v) do Quadro 4, provenientes dos estudos específicos de Freitas e Quental, dão origem a regras desenvolvidas neste trabalho e numeradas como 6 a 8, no Quadro 7.

O padrão (iii), denominado "tipos de", busca extrair relações com base nas palavras-chave que dão origem ao seu nome. O excerto de texto

a seguir, proveniente do *corpus* CORSA, visa a demonstrar as relações que a regra correspondente é capaz de extrair:

> "desenvolver [dois tipos de dengue] : [dengue clássica] e
> [dengue hemorrágica]".

Desse trecho, a regra deve ser capaz de extrair as relações: Hiponímia (dengue clássica, dengue), Hiponímia (dengue hemorrágica, dengue). O resultado da adaptação criada para realizar tal tarefa é descrito na regra 6 do Quadro 7, a saber:

> <... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN

A semelhança com o padrão de Freitas e Quental se dá na utilização dos símbolos "<" e ">", para representar um sintagma nominal que contém em seu interior as palavras-chave da regra. Isso se deve ao fato de o *parser* PALAVRAS definir que a expressão "tipos de" faz parte de um *chunk* com outras palavras que podem vir antes ou depois do padrão, por exemplo, "[todos os tipos de cortes]" e "[os principais tipos de tifo]". Para maximizar o número de relações extraídas, a regra foi flexibilizada para aceitar a expressão "tipo de", sem a utilização do plural. Essa regra apresenta um alto grau de confiança, como pontua Freitas:

> "... o padrão 'tipos de' não apresenta problemas de ambiguidade relativos ao sintagma preposicionado, nem particularidades de natureza discursiva ou coesiva – o que significa que as relações identificadas são altamente confiáveis." (FREITAS, 2007, p. 75).

Outra adaptação, dos estudos de Freitas e Quental, foi a do padrão denominado "chamado/a/os/as", representado como (iv) no Quadro 4, que deve extrair relações de textos como:

> "... e nele existe uma [substância] chamada [benzopireno]."

Nesse caso, a relação extraída é Hiponímia (benzopireno, substância). A regra encarregada de tal extração é identificada como regra 7 no Quadro 7, a saber:

> SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN

Para maximizar o número de relações extraídas, foram flexibilizados o uso do verbo "ser" em quatro formas – "é", "são", "foi", "foram"– e a utilização de vírgula. Foi também permitida a ocorrência de uma lista de sintagmas nominais após a palavra-chave "chamado". Esse formato de lista já está presente nas regras 1, 2, 3 e 4 do Quadro 7; e permite a extração de relações de excertos de textos como:

> "... vem estudando profundamente [o fenômeno] , chamado de [sinantropia] ou [domiciliação]..."

O padrão (v) do Quadro 4 é o último adaptado do trabalho de Freitas e Quental. As autoras o denominam "conhecido/a/os/as como", e deve extrair relações de excertos como:

"[vesículas esféricas de gordura] , conhecidas como [lipossomas]", obtendo a relação Hiponímia (lipossomas, vesículas esféricas de gordura).

Após o processo de adaptação, a regra criada, numerada como regra 8 no Quadro 7, ganhou a seguinte representação:

> SN((,)? também)?(,|é|são|foram)? conhecid(o|a|os|as) como (SN,)*SN (e|ou))*SN

Para maximizar o número de relações extraídas pela regra 8, foram realizadas alterações, permitindo a presença de vírgula e das formas verbais "é", "são", "foi" e "foram" antes da expressão "conhecido como", além de uma lista de sintagmas nominais ter sido acrescentada após a expressão. Ainda, a presença da palavra "também" após o primeiro sintagma nominal passou a ser prevista.

No Quadro 9, esses são associados ao presente trabalho.

Quadro 9 – Associação entre os padrões de Freitas e Quental e regras do presente trabalho

Regra	Padrão de Freitas e Quental
6	tipos de SN Hiper: SN 1 { , SN 2 ... ,} (e | ou) Sni
7	SN HHiper chamado/s/a/as (de) SN Hipo
8	SN Hiper conhecido/s/a/as como SN Hipo

Fonte: nossa autoria, baseado em Freitas e Quental (2007).

Analisando o trabalho de Freitas e Quental, nota-se um formato de sintagma nominal que está ausente nas regras do presente trabalho: "SN HHiper". As autoras utilizaram esse prefixo para os sintagmas nominais com o objetivo de melhorar a precisão das extrações. O SN HHiper é utilizado para identificar apenas a primeira palavra encontrada mais à direita de um sintagma nominal. Já os "SN Hipo" e "SN Hiper" são utilizados para representar um sintagma nominal como elemento hiponímico ou hiperonímico da relação.

3.3 Extração com padrões de Taba e Caseli (2014)

O trabalho de Taba e Caseli (2014) também estuda o emprego de padrões para extração automática de relações semânticas de *corpus* de língua portuguesa. Em sua pesquisa, os autores utilizam os padrões criados por Freitas e Quental, assim como outros de sua própria autoria. Destes, abordaremos apenas os padrões (v), (vi) e (vii) do Quadro 6, que realizam extração de relações hiponímicas (denominadas por Taba e Caseli de relações "is-a"). O primeiro padrão adaptado foi o padrão (v), que busca extrair relações de textos como:

> "... apresentar [febre] ou [qualquer outro sintoma da doença de Chagas]..."

Com base no padrão (v), obtém-se a relação Hiponímia (febre, sintoma da doença de Chagas). A sua adaptação, apresentada na regra de número 9 no Quadro 7, pode ser vista a seguir:

(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>

No padrão original (TABA; CASELI, 2014, p. 2741), são permitidas apenas as palavras "outro" ou "outros" antes do último SN. No corrente trabalho, esse modelo foi flexibilizado para que a palavra no gênero feminino também fosse válida ("outra", "outras"). Assim como em outras regras, foram utilizados os sinais ">" e "<" para indicar que as palavras-chave são encontradas dentro de um *chunk*, e uma parte desse *chunk*, que é representada por "sn", será considerada nas relações extraídas.

Já o padrão (vi) do Quadro 6 é capaz de extrair relações de sentenças como:

> "por [a agência local de a Fundação Instituto Brasileiro de Geografia e Estatística] , [Pelotas] é [uma cidade] [cuja zona urbana comporta 297.825 habitantes]"

No caso, é obtida a relação Hiponímia (Pelotas, cidade).

O padrão em questão gerou duas regras, respectivamente numeradas no Quadro 7 como 10A e 10B:

SN é < (o|a) sn>
SN é < (um|uma) sn>.

Como pode ser observado, as regras 10A e 10B são semelhantes, mas o motivo da criação de duas regras, nesse caso, é o fato de ambas serem generalistas. Como extraem um grande número de relações, essa separação visa a que futuras análises possam determinar a precisão das regras individualmente. Ambas as regras apresentam a estrutura que indica que as palavras-chave estão dentro do *chunk*.

O padrão (vii) de Taba e Caseli (QUADRO 6) objetiva extrair relações de textos como no exemplo a seguir:

> "[as hemoglobinopatias] são [doenças geneticamente determinadas] e apresentam [morbidade significativa] em todo o mundo",

obtendo a relação Hiponímia (as hemoglobinopatias, doenças geneticamente determinadas).

Por fim, podemos ver a última regra adaptada de Taba e Caseli (2014), referente ao padrão (vii) do Quadro 6 e descrita no Quadro 7 como regra 11. A construção dessa regra reflete basicamente a transcrição do padrão desses autores para a sintaxe utilizada neste trabalho:

SN são SN

No Quadro 10, podem ser vistos os padrões adaptados de Taba e Caseli (2014), com suas correspondências para quatro regras do presente trabalho.

Quadro 10 – Associação entre padrões de Taba e Caseli e as regras do presente trabalho

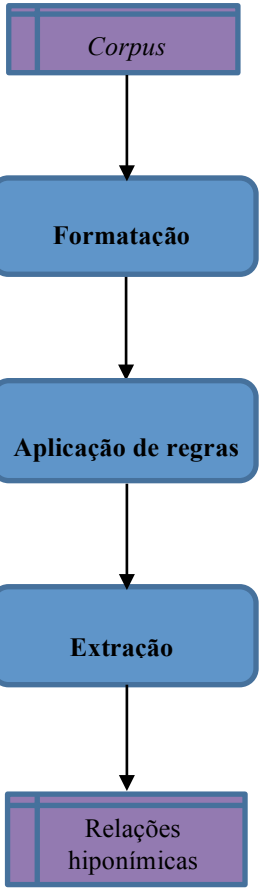
Regra	Padrão de Taba e Caseli
9	(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>
10A	SN é < (o|a) sn>
10B	SN é < (um|uma) sn>
11	SN são SN

Fonte: nossa autoria.

3.4 Protótipo e aplicação das regras

Para implementar e testar um extrator de relações hiponímicas de textos em português com base nos padrões trabalhados, foi desenvolvido um protótipo funcional, cuja arquitetura é descrita na Esquema 1.

Esquema 1 – Arquitetura utilizada na construção do protótipo



Fonte: nossa autoria.

Como mostra o Esquema 1, o fluxo inicia pela inserção do *corpus* como um parâmetro de entrada. O processo de formatação age sobre todo o *corpus* de entrada e produz, como parâmetro de saída, uma nova versão desse mesmo *corpus*, agora formatado adequadamente para a extração das relações. Então, o processo de aplicação de regras entra em ação, executando sobre cada sentença as regras do Quadro 7. Como

resultado, esse processo retorna todos os trechos de sentenças que foram identificados pelas regras. Na última etapa, esses trechos são inseridos como parâmetro de entrada para o processo de extração. Nesse processo, as relações resultantes são criadas e, então, é produzida uma lista com todas as extrações obtidas pela execução do protótipo.

As regras, na forma de expressões regulares, são escritas em linguagem formal e são interpretadas pelo processador de expressões regulares, que examina o texto e procura por trechos que atendam às regras determinadas pela expressão. A escolha desse método se deu por sua simplicidade e expressividade, assim como por estar disponível para uso em diversas linguagens de programação.

O *corpus* utilizado para experimentar o protótipo desenvolvido foi o CORSA, que é formado por 1.846.502 palavras armazenadas em um arquivo de 11Mb. O CORSA foi criado com base em textos da área de Saúde Pública, incluindo artigos acadêmicos, cartilhas, manuais, textos didáticos e também textos jornalísticos. A diversidade das fontes é proposital, com o objetivo de agregar variadas formas de escrita, assim como diferentes níveis de aprofundamento técnico. Esses textos foram analisados previamente pelo *parser* PALAVRAS. Em seguida, os sintagmas nominais foram etiquetados de acordo com as indicações propostas por Santos e Oliveira (2005). No *corpus*, cada linha apresenta uma palavra seguida de sua etiqueta POS. A palavra é separada de sua etiqueta pelo símbolo "_". Ainda, no final de cada linha é encontrada uma etiqueta do tipo "BIO": "I" para representar o início de um sintagma nominal, "O" para representar o fim, ou ainda "B", representando a ocorrência conjunta do fim do SN anterior e início de um novo. Segue um exemplo de trecho etiquetado do CORSA:

```
...
o_ART_I
Ministério=da=Saúde=do=Brasil_NPROP_I
(_(_O
MS_NPROP_I
)_)_O
lançou_V_O
uma_ART_I
campanha_N_I
nacional_ADJ_I
para_PREP_O
promover_V_O
o_ART_I
uso_N_I
```

```
de_PREP_I
preservativos_N_I
entre_PREP_I
meninas_N_I
adolescentes_ADJ_I
._._O
entre_PREP_I
13_NUM_I
e_KC_I
19_NUM_I
anos_N_I
,_,_O
um_ART_I
grupo_N_I
social_ADJ_I
que_PRO-KS_O
havia_VAUX_O
registrado_PCP_O
crescimento_N_I
em_PREP_O
o_ART_I
número_N_I
de_PREP_I
casos_N_I
de_PREP_I
AIDS_NPROP_I
e_KC_O
outras_PROADJ_I
doenças_N_I
sexualmente_ADV_I
transmissíveis_ADJ_I
._._
```

Nesse formato de *corpus*, não é possível que um sintagma nominal contenha outro, ou seja, não se pode representar aninhamentos de SNs nem, por consequência, empregar regras recursivas ou reentrantes. A escolha de um *corpus* já etiquetado foi realizada com o intuito de diminuir a influência do erro na fase de pré-processamento. Assim, possíveis erros nessa fase não são propagados para a fase de avaliação das extrações, evitando o prejuízo à análise dos resultados.

Com o objetivo de possibilitar o funcionamento com diferentes formatos de *corpus* e ainda facilitar a criação das regras, o *corpus* de entrada é convertido para um formato específico, o qual permite aplicar as expressões regulares, agregando a origem do padrão por autor de referência e número de relações geradas pelo padrão.

O formato adotado aceita sentenças descritas textualmente, com apenas um destaque para os sintagmas nominais. Esses estão entre colchetes, como pode ser visto a seguir:

> "... entre [os municípios maiores] , [Cáceres] e [Rondonópolis] são..."

Esse formato é aplicado a todo *corpus* e, na sequência, cada sentença é adicionada a uma lista para se dar início à próxima etapa, a de aplicação das regras. Nessa etapa a lista de sentenças é percorrida e, para cada sentença, todas as regras são aplicadas em forma de expressões regulares. Quando uma expressão "casa", ou seja, combina com uma sentença, dá-se início à etapa de identificação dos termos da relação.

Ao longo do trabalho de prototipação, foi preciso adicionar diversas regras e alterá-las. Percebeu-se que era preciso simplificar esse processo, já que, até então, era necessário escrever todo o código para a criação e aplicação de cada regra. Assim, foi adotado o conceito do armazenamento de regras em arquivo externo. As regras foram escritas em um arquivo externo, que foi usado como entrada na etapa de aplicação das regras. O arquivo de entrada consiste de um documento JSON (*Java Script Object Notation*), com todas as regras listadas e respectiva identificação da fonte. Esse formato de documento foi adotado por ser um padrão leve, de simples implementação e alta expressividade.

Nessa etapa, a relação já foi identificada na sentença, mas ainda se deve verificar quais dos SNs compõem cada relação extraída, posto que uma regra pode identificar mais de uma relação binária. Além disso, é necessário detectar qual sintagma nominal é o termo hiponímico e hiperonímico da relação. Por fim é gerada uma lista com todas as relações encontradas, no seguinte formato:

> Sentença: {Sentença analisada}
> Extrações: {Autor}-{Padrão} {Nome da Relação}({Argumento1}, {Argumento2}) ...

O campo {Autor} contém a identificação da fonte do padrão que deu origem à regra.

4 Avaliação e discussão dos resultados

4.1 Modelo de processo avaliativo

Durante a execução das etapas de avaliação, diversas dificuldades foram encontradas, entre as quais o grande número de relações extraídas pelo protótipo, ao todo 8.601, o que impossibilitou a análise manual de todas as extrações. Descartou-se a possibilidade de avaliação automática, devido à indisponibilidade de um *Gold Standard* na língua portuguesa com o qual os resultados pudessem ser comparados. Outro motivo que dificultou a execução de uma análise manual foi a falta de uma equipe com o número apropriado de avaliadores para realizar o processo avaliativo. Dois avaliadores realizaram a tarefa de avaliação, ambos com dedicação parcial.

Durante o processo de avaliação de resultados, torna-se necessário situar o trabalho em relação à bibliografia e comparar os resultados com os de outros autores. Na literatura, encontramos poucos trabalhos que realizam a extração de relações em *corpora* de língua portuguesa e, entre esses, não foi possível encontrar resultados que possam ser considerados um *baseline* (base para avaliação) para comparar a precisão e a cobertura. Mesmo assim, para situar o leitor, o texto apresenta na Seção 4.5 alguns dados de precisão e cobertura anunciados em pesquisas nessa área.

Para realizar os testes e a avaliação, o *corpus* CORSA serviu como entrada do protótipo construído. Essa escolha permitiu a comparação de resultados com os descritos por Freitas e Quental (2007), que realizaram a avaliação em dois formatos. No primeiro, as autoras analisaram os resultados dos padrões por elas propostos, individualmente, em busca de erros sintáticos. O objetivo era a eliminação dos erros mais frequentes para cada padrão. Já no segundo formato de avaliação, que se toma como principal referência, foi realizada uma validação humana, em que o foco era tornar os resultados "mais comparáveis" e "mais significativos". A avaliação se deu, no âmbito do trabalho de Freitas e Quental, após essa validação. As relações resultantes receberam notas de 0 a 3, com base nos critérios apresentados pelas autoras e indicados no Quadro 11.

Quadro 11 – Critérios de avaliação empregados por Freitas e Quental

Nota	Descrição
3	*A relação está correta da forma como foi extraída.*
2	*A relação está "um pouco" correta, isto é, o substantivo núcleo está correto, mas preposições, adjetivos etc. que o acompanham deixam a relação estranha.*
1	*A relação está correta em termos gerais; isto é, é muito geral ou muito específica para ser útil.*
0	*A relação está errada.*

Fonte: baseado em Freitas e Quental (2007).

No processo de avaliação desenvolvido por Freitas e Quental, três avaliadores realizaram em conjunto a análise de 436 relações, que representaram um terço das relações extraídas. O consenso entre os três avaliadores determinou a nota atribuída a cada relação. Esses avaliadores tinham formação em biologia, educação física e direito, ou seja, bastante diversificada.

Para avaliar as extrações, a análise foi realizada sobre um subconjunto do total de relações extraídas, composto por todas as 218 extrações obtidas pelas regras indicadas como 6, 7 e 8 no Quadro 7. Essas regras foram escolhidas por terem extraído uma quantidade aceitável de relações (considerando-se o processo avaliativo) e por pertencerem ao conjunto de regras adaptadas do trabalho de Freitas e Quental (2007), o que viabilizaria, em alguma medida, uma comparação. Para esse propósito, dois juízes humanos, sem treinamento prévio, analisaram as 218 relações sob os mesmos critérios utilizados no processo avaliativo de Freitas e Quental. Cada um dos dois avaliadores humanos realizou individualmente a avaliação e atribuiu notas de 0 a 3 às extrações.

As seções 4.2, 4.3 e 4.4 apresentam diferentes propostas de análise quanto à avaliação e às extrações. A Seção 4.5 se detém na

avaliação dos resultados em que houve concordância entre os avaliadores.

4.2 Distribuição das extrações

A aplicação das regras 1 a 11 (Quadro 7) sobre o *corpus* CORSA extraiu 8.601 relações, organizadas em três grupos. Cada grupo compõe-se das relações obtidas com a aplicação das regras sugeridas pelos autores de referência, a saber: grupo 1, regras 1 a 5, provenientes dos estudos de Hearst (1992); grupo 2, regras 6 a 8, exclusivamente de Freitas e Quental (2007); e grupo 3, regras 9 a 11, exclusivamente dos estudos de Taba e Caseli (2014). Nos grupos 2 e 3, tomou-se o cuidado de não se retomarem os padrões de Hearst, implementados pelos respectivos autores de referência. A constituição dos três grupos de regras visou a evidenciar a contribuição específica dos padrões trazidos pelos trabalhos de referência na adição de propostas voltadas à língua portuguesa. A Tabela 1 mostra o número de relações extraídas a partir de cada grupo de regras e o percentual com que cada grupo contribui para o total das relações extraídas. Conforme a tabela, o grupo das regras provenientes de Hearst trouxe a maior contribuição relativa para o resultado, gerando 69,02% das 8.601 relações obtidas, o que já era esperado. As regras propostas por Taba e Caseli constituíram o segundo grupo mais representativo, com 28,45% do total de relações obtidas. Por fim, o grupo das regras sugeridas por Freitas e Quental gerou 2,53% do total das 8.601 relações extraídas.

Tabela 1 – Número de relações extraídas do *corpus* CORSA por grupo de referência

Referência	Número de relações	Percentual
Grupo 1	5.936	69,02%
Grupo 2	218	2,53%
Grupo 3	2.447	28,45%
TOTAL	8.601	100%

Fonte: nossa autoria.[3]

[3] Todas as Tabelas neste trabalho são de nossa autoria.

Ressalta-se, contudo, que tal distribuição não pode ser entendida como uma medida da produtividade das regras, uma vez que se trata de uma ótica de extensão a padrões já consolidados e amplamente empregados, que são os de Hearst, com a capacidade de somar novas relações às já extraídas por aqueles padrões. As extrações produzidas, independentemente dos grupos de regras que as originam, constituem relações igualmente importantes às aplicações, por exemplo, de construção de estruturas de conhecimento, tais como ontologias, ou de extração de informações a partir de bases de dados textuais, tais como repositórios de artigos científicos. Ainda, a distribuição leva em conta exclusivamente o estudo sobre o *corpus* CORSA e pode sofrer influências tanto da adaptação das regras como das próprias ferramentas de processamento que promoveram a etiquetagem do CORSA.

A Tabela 2 exibe o número de relações obtidas com as regras que adaptam os padrões de Hearst e o percentual que essas relações representam em relação ao total de 5.936 extrações que se encontram nesse grupo (cf. TABELA 1). A Seção 2.2 apresenta os padrões originais. A regra 1, que busca extrair relações por meio da palavra-chave "como", gerou um número grande de relações, representando 76,9% das extrações. O resultado já era esperado, pois a palavra-chave em questão é comum na língua portuguesa. Esse grande número de relações extraídas influenciou no resultado obtido com base em padrões de Hearst.

Tabela 2 – Distribuição das relações extraídas por regras adaptadas de Hearst

Regra	Número de relações	Percentual
1	4.565	76,90 %
2	351	5,91 %
3	578	9,74 %
4	376	6,33 %
5	63	1,06 %
TOTAL	5.936	100 %

As regras adaptadas de Freitas e Quental (2007) são baseadas em termos com menor frequência em textos em língua portuguesa e extraem 218 relações. A Tabela 3 mostra, por exemplo, que a regra 6, representada como "<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN", extrai 44,95% das relações.

Tabela 3 – Distribuição das relações extraídas por regras adaptadas de Freitas e Quental

Regra	Número de relações	Percentual
6	98	44,95%
7	75	34,40%
8	45	20,64%
TOTAL	218	100%

A Tabela 4 mostra a distribuição das 2.447 extrações obtidas com a adaptação dos padrões de Taba e Caseli (2014). Algumas regras são abrangentes, por exemplo as que se baseiam em expressões como "é um" e "são", obtendo alto número de relações.

Tabela 4 – Distribuição das relações extraídas por regras adaptadas de Taba e Caseli (2014)

Regra	Número de relações	Percentual
9	23	1,00%
10A	920	37,59%
10B	694	28,36%
11	810	33,10%
TOTAL	2.447	100 %

Ainda analisando as relações extraídas com base em Taba e Caseli (2014), observa-se que a regra 9 apresenta uma quantidade de extrações muito inferior às demais. Essa se baseia na combinação das palavras "qualquer" e "outros", que é menos comum na língua portuguesa, tornando-se uma regra de aplicação menos frequente.

4.3 Avaliação dos resultados

Esta seção apresenta o resultado da avaliação de um subconjunto das extrações obtidas pela aplicação das regras. Embora empregando uma metodologia de avaliação já aplicada, o sentido, mais do que buscar comparações, é prover análises dos resultados obtidos, sob diferentes prismas.

Foi utilizado o processo de avaliação manual dos resultados, também relatado na literatura. Devido ao fato de o total de resultados em foco ser superior a 8 mil extrações, a análise manual tornou-se inviável no tempo disponível, sendo estabelecido um subconjunto de relações para análise. Foram escolhidas as 218 relações extraídas com base nas regras adaptadas de Freitas e Quental (2007) e, com essas, foi possível realizar a avaliação manual. Outro motivo importante para a escolha das relações utilizadas nessa etapa foi a possível comparação de resultados com os relatados em Freitas e Quental (2007).

O Avaliador 1 classificou cada resultado em um de quatro grupos que são representados por notas de 0 a 3, gerando os dados presentes na Tabela 5.

Tabela 5 – Resultado da Avaliação 1: total de relações por nota de avaliação

Nota	Número de relações	Percentual
0	29	13,3%
1	41	18,8%
2	46	21,1%
3	102	46,8%
TOTAL	218	100 %

Analisando a Tabela 5, observa-se que um total de 46,8% de relações extraídas com 100% de correção (nota 3) não é um valor alto. Por outro lado, apenas 13,3% das relações foram consideradas totalmente erradas, o que é um resultado promissor.

Na avaliação feita pelo Avaliador 2 (TABELA 6), os resultados se assemelham com os obtidos na primeira avaliação, com um leve desvio nas relações classificadas com nota 1 e 2, o que pode demonstrar

alguma dificuldade em trabalhar com a escala proposta por Freitas e Quental.

Tabela 6 – Resultado da Avaliação 2: total de relações por nota de avaliação

Nota	Número de Relações	Percentual
0	26	11,9%
1	53	24,3%
2	41	18,8%
3	98	45,0%
TOTAL	218	100, %

Na busca de melhor entendimento do processo avaliativo, a Tabela 7 apresenta uma distribuição das notas atribuídas, considerando o total de 436 avaliações (218 de cada avaliador). Por exemplo, observa-se que, das notas atribuídas, 21,6% foram notas 1, valor que representa a proporção de notas 1 atribuídas no contexto das 436 avaliações.

Tabela 7 – Notas atribuídas pelos avaliadores

Nota	Percentual
0	12,6%
1	21,6%
2	19,9%
3	45,9%
TOTAL	100,%

Outro ponto interessante é a diferença entre os julgamentos atribuídos pelos avaliadores, como mostra a Tabela 8. Ou seja, das 218 avaliações em tela, 109 são coincidentes. É sobre elas que se desenvolve, mais adiante, a avaliação consolidada do trabalho.

Tabela 8 – Comparação entre resultados de julgamento pelos avaliadores

Nota	Avaliações idênticas	Percentual
0	13	11,9%
1	14	12,8%
2	13	11,9%
3	69	63,3%
TOTAL	109	100%

Constata-se que o número de relações que receberam a mesma nota pelos avaliadores é consideravelmente baixo. Esse resultado demonstra uma diferença nos critérios de cada avaliador ao determinar se uma relação está correta, fator que tem um significado importante na avaliação. A Tabela 9 apresenta um exemplo dessa diferença, para as extrações A, B, C e D a seguir:

A Hiponímia (transtorno de a compulsão alimentar periódica; transtorno alimentar)
B Hiponímia (questionário individual de homens; questionários)
C Hiponímia (questionário individual de mulheres; questionários)
D Hiponímia (colinesterase verdadeira; colinesterases)

Tabela 9 – Comparação entre julgamentos para cinco relações específicas

Relação	Avaliador 1	Avaliador 2
A	3	1
B	2	3
C	2	3
D	3	1

Todas as relações de A até E foram avaliadas com nota 3 no processo avaliativo conduzido por Freitas e Quental. Já no processo de avaliação realizado no presente trabalho, essas relações receberam notas distintas. Na Tabela 9, pode-se notar que apenas a relação A obteve o mesmo resultado nas três avaliações (a saber, Avaliador 1, Avaliador 2 e

avaliação relatada por Freitas e Quental). A discordância entre as avaliações sugere que os critérios de julgamento possam ser ambíguos.

Outra forma utilizada para analisar os resultados, provavelmente a mais expressiva, é a que leva em consideração apenas os resultados em que há concordância entre as avaliações providas pelos avaliadores humanos. A título de exemplo, a Tabela 10 traça uma distribuição das extrações conforme notas atribuídas pelos avaliadores exclusivamente nos casos em que houve concordância. Nas colunas, encontram-se as notas de 0 a 3. As linhas agrupam as extrações por regra (6, 7 ou 8) cuja aplicação as originou.

Tabela 10 – Percentual médio de relações encontradas por avaliação e por regra, segundo critério de concordância entre avaliadores

Regra/Nota	0	1	2	3
6	11,1%	9,3%	14,8%	64,8%
7	8,1%	16,2%	10,8%	64,9%
8	22,2%	16,7%	6,25%	55,6%

Considerando os dados na Tabela 10, as regras 6 e 7 apresentam resultados muito semelhantes no que se refere à nota 3. Já a regra 8 apresenta percentuais um pouco inferiores nas notas mais altas, o que pode indicar que ela apresente uma precisão também inferior se comparada com as regras 6 e 7.

É importante ressaltar que, para além dos resultados da avaliação humana (TABELAS 5, 6), optou-se por examinar aqui, cuidadosamente, o comportamento das notas atribuídas pelos avaliadores, tanto no geral (TABELAS 7, 8) quanto em extrações a partir de regras específicas (TABELA 9). A Tabela 10 mostra o comportamento dos dados nos casos em que há concordância entre os avaliadores, situação que será retomada na seção 4.5.

4.4 Uma análise de erros encontrados

Analisando as relações que obtiveram nota 0 de ambos os avaliadores, podemos destacar alguns motivos de erros mais frequentes. Um desses é o erro de *chunking* quando o *parser* realiza uma identificação incorreta. Esse mesmo tipo de erro foi apontado em Corro e Gemulla (2013) como uma das principais fontes de incorreções naquele trabalho. O erro ocorre após a etapa de tokenização, quando o *chunker* identifica os sintagmas nominais. A seguir consta de um exemplo em que o *parser* identificou, incorretamente, a letra "o" como sendo um sintagma nominal:

> "... [dois tipos de modelos] : [o] logístico e [o] hierárquico..."

Em alguns casos, um sintagma nominal pode ser subdividido em SNs menores, sem de fato gerar um erro sintático. Esse comportamento não pode ser considerado uma falha no *chunker.* Tecnicamente, tanto a identificação de um Sintagma Nominal composto (formado por um grupo de SNs) quanto a identificação de apenas um elemento desse conjunto estão corretas, mas esse comportamento gera resultados incoerentes. O trecho a seguir provê alguns exemplos:

> "[o aparecimento de anticorpos] em [o sangue] , chamado de [janela imunológica]".

O *parser* identificou "[o aparecimento de anticorpos]" e "[o sangue]" como dois SNs distintos, gerando uma possível extração errada: Hiponímia (o sangue, janela imunológica). Caso o *parser* identificasse os SNs como um só, uma relação mais precisa poderia ser extraída: Hiponímia (o aparecimento de anticorpos em o sangue, janela imunológica). Para corrigir essa falha, seria preciso de um *chunker* que agrupasse os SNs nesses casos. Outra solução seria prover uma etapa de pré-processamento que unisse *chunks* que podem dar origem a esses casos.

Outro erro encontrado é o de correferência, que ocorre quando o sintagma nominal faz referência a outro SN citado anteriormente na

sentença. Um exemplo pode ser visto no trecho a seguir, no qual o SN faz referência a "corpo":

> "tornar dócil [um corpo] não é [coisa simples], pois ele, normalmente , está submetido a [seu chefe natural], chamado [personalidade]".

Uma extração adequada para essa sentença seria Hiponímia (personalidade, chefe natural do corpo). Uma abordagem para solucionar esse problema seria a utilização de métodos de resolução de correferência.

Outro erro encontrado se refere à falta de contexto. Ele ocorre quando o termo é extraído corretamente, mas ele só faz sentido quando está inserido em um determinado contexto. Segue um exemplo:

> "... [a segunda fase] , chamada de [análise]..."

A regra está correta ao extrair a relação Hiponímia (a segunda fase, análise), mas como não sabemos a que entidade a palavra "fase" faz referência, a extração perde o significado se analisada fora do seu contexto.

Outro erro encontrado está presente na expressão que explora relações formadas por listas de SNs. Tal expressão considera que todos os SNs seguidos por "e", "ou" e "," fazem parte da mesma lista, mas em determinados casos esses conectores podem apenas ligar duas sentenças, não tendo a função de criar lista de sintagmas nominais. Seguem alguns exemplos:

> "[um gênero de vírus] conhecido como [flavivírus] , [a enfermidade] apresenta ..." "[a bactéria] chamada [*Rickettsia mooseri*] e [os sintomas] são praticamente..."

A relação Hiponímia (a enfermidade, um gênero de vírus) é extraída indevidamente, assim como Hiponímia (os sintomas, a bactéria). Apesar de, em ambas as sentenças, o padrão ser aplicado corretamente ao

primeiro SN, o segundo sintagma nominal é considerado indevidamente como parte da lista.

Já quando analisamos as relações apontadas por ambos os avaliadores como pertencendo ao grupo da nota 1, o erro mais comum encontrado é a ocorrência de palavras desnecessárias para o significado da relação dentro de um dos sintagmas nominais. A seguir podem ser vistos exemplos desse fenômeno:

> "[a ação de os vírus] conhecidos como [Influenza A]" "[essas lesões] , chamadas de [isquemia]".

As relações extraídas nesse caso são Hiponímia (a ação de os vírus, Influenza A) e Hiponímia (essas lesões, isquemia). Caso as relações extraídas fossem respectivamente Hiponímia (influenza A, vírus) e Hiponímia (isquemia, lesões), as relações obteriam, acredita-se, uma classificação melhor. Para solucionar esse tipo de problema, Freitas e Quental criaram uma etapa de pós-processamento que aplica filtros para remover palavras dos sintagmas nominais que não agreguem significado à relação. Uma etapa semelhante poderia ser utilizada no trabalho aqui apresentado, com o objetivo de melhorar a precisão. Para isso, no entanto, é essencial dispor de uma lista de palavras que frequentemente não agregam valor semântico, por exemplo, preposições e pronomes.

4.5 Discussão dos resultados

Ao longo das seções 4.2 a 4.4 foram relatados os resultados dos testes realizados. A presente seção discute esses resultados no cenário dos trabalhos correlatos. A precisão das extrações, neste estudo, será tomada como 63%, percentual representado pelas 69 extrações às quais foi atribuída nota máxima por ambos os avaliadores, entre as 218 extrações avaliadas (regras 6, 7 e 8), mostrado na Tabela 8.

A dificuldade de avaliar os resultados por comparação está no uso de regras diferentes pelos diferentes autores, assim como na escolha de *corpora* distintos, além de etapas distintas de pré-processamento ou pós-

processamento. A avaliação humana também tem caráter subjetivo, o que repercute nas avaliações.

A primeira comparação é realizada em relação ao publicado em Freitas e Quental (2007), utilizando, em uma das etapas, o *corpus* CORSA. Em uma das etapas avaliativas, as autoras afirmam obter 73,4% quando aplicam as regras "como / tais como", "e outros", "tipos de", "chamado" e "conhecido como" sobre o *corpus* CORSA.

Já de início destaca-se a inclusão do primeiro padrão, proveniente de Hearst, que seguramente pode trazer um viés de ampliação aos resultados exclusivos das regras 6, 7 e 8 analisadas. A precisão de 73%, expressivamente superior aos 63% do presente trabalho, deve ser relativizada devido à introdução de uma primeira etapa de avaliação, conduzida por Freitas e Quental. Nessa etapa, foi realizada uma análise manual sobre o resultado e foram removidas 726 relações consideradas sintaticamente erradas. Uma tal remoção configura uma filtragem que visa a melhorar a precisão. O resultado de 73,4% obtido por Freitas e Quental considera a etapa de avaliação realizada pela autora sobre extrações no *corpus* CORSA, após a primeira etapa mencionada.

Na conclusão de seu trabalho, Freitas e Quental consideram o resultado final de 75%, esse resultado foi calculado utilizando o *corpus* CETEN-Folha, sem a realização da primeira etapa, em que são removidas manualmente relações sintaticamente errôneas, mas já utilizando filtros propostos com base no estudo sobre o primeiro *corpus*. O primeiro filtro remove relações cujo argumento hiperonímico é substantivo com alto grau de generalidade ou falta de especificidade. Outros dois filtros buscam remover palavras que não agregam valor semântico. Com esse objetivo o primeiro filtro remove pronomes dêiticos, e o segundo remove alguns adjetivos. Observa-se, aqui, um grau de subjetividade na definição dos filtros, que pode ser tema de estudos futuros. Feitas essas observações, a diferença de 73,4% para 75% de precisão necessitaria um aprofundamento, seja com a ampliação de experimentos com diferentes *corpora*, seja com uma sistematização dos filtros empregados.

A título de informação, é interessante trazerem-se resultados obtidos por outros autores como Hearst (1998) e Morin e Jacquemin (2003), sempre levando em conta as diferenças entre *corpora*, processo

avaliativo, regras e também idioma. Analisando a Tabela 12, é possível constatar que, embora todas as diferenças do escopo de avaliação, idioma e *corpora*, que impedem qualquer comparação, a precisão de 63%, obtida no presente estudo, não se distancia da reportada nos estudos de Hearst (1998) nem do trabalho de Cederberg e Widdows (2003). Por outro lado, a precisão alcançada por Morin e Jacquemin (2003), bem superior às demais, reflete uma cuidadosa base de aplicação dos estudos de Hearst para a língua francesa, ampliada com relações específicas associadas ao idioma francês. A precisão relatada também foi produzida com teste sobre um só *corpus*.

Tabela 12 – Precisão reportada em estudos da área

	Corpus em Língua Portuguesa		*Corpus* em Língua Estrangeira		
	Nosso trabalho	Freitas e Quental (2007)	Morin e Jacquemin (2003)	Cederberg e Widdows (2003)	Hearst (1998)
Precisão	63%	73,4%	81%	64%	63%

Ainda assim, existem técnicas, algumas já mencionadas, que poderiam melhorar essa precisão, para chegar aos patamares dos demais trabalhos. Por exemplo, a filtragem de palavras que não agregam valor semântico, ou a inclusão de uma etapa de pré-processamento que una *chunks* em situações específicas (como já exposto). Todas essas técnicas de filtragem, entretanto, baseiam-se em observação e, por consequência, sofrem limitações.

5 Considerações finais

Uma das principais contribuições do presente trabalho é a agregação, num único estudo, de regras elencadas por diferentes autores, como as propostas por Freitas e Quental (2007), Hearst (1992) e Taba e Caseli (2014). Não menos importante é a contribuição da análise minuciosa dos resultados obtidos. Esses foram estudados segundo diferentes critérios, tais como: por regras, por autor, por nota e por avaliador. Ainda foram discutidos fatores que tornam subjetivo o processo de avaliação manual.

Durante o desenvolvimento deste trabalho ficou evidente a necessidade de criação de uma *Gold Standard* para extração de relações hiponímicas na língua portuguesa. Esse artefato contribuiria para o desenvolvimento das pesquisas na área, permitindo o cálculo de precisão e cobertura e comparações mais efetivas. A tarefa, entretanto, teria de contar com a condução de especialistas trabalhando também questões de escopo, contexto e referência, bem além da etiquetagem de relações, esforço que teria de ser amplamente registrado, para formalizar critérios e condutas a ser adotadas.

6 Referências

BANKO, M. *et al*. Open information extraction from the web. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE – IJCAI, 20, 2007, Hyderabad, India. *Proceedings*... Editor: Manuela M. Veloso. Hyderabad, India, January 6-12, 2007. p. 2670-2676.

BASÉGIO, T. L. *Uma abordagem semiautomática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil*. 2007. 124 p. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Pontifícia Universidade Católica, RS, 2007.

BATISTA, D. S. *et al*. Extração de relações semânticas de textos em português explorando a DBpédia e a Wikipédia. *Linguamática*, Braga, v. 5, n.1, p. 41-57, 2013.

BICK, E. *The parsing system PALAVRAS* – automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus: Aarhus University Press, 2000.

CEDERBERG, S.; WIDDOWS, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 7, 2003, Edmonton. *Proceedings*...

Edmonton: Association for Computational Linguistics, 2003. CONLL, v. 4 p. 111-118. DOI: <http://dx.doi.org/10.3115/1119176.1119191>

CORRO, L.; GEMULLA, R. ClausIE: clause-based open information extraction. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 22, 2013, Rio de Janeiro. *Proceedings*… Rio de Janeiro, 2013. p. 355-366.

DEGERATU, M.; HATZIVASSILOGLOU, V. An automatic method for constructing domain-specific ontology resources. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 4, 2004, Lisboa. *Proceedings*... Lisboa, Porutgal, 2004. p. 2001-2004.

FREITAS, M. C. *Elaboração automática de ontologias de domínio*. 2007. 142 p. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Letras, Pontifícia Universidade Católica, Rio de Janeiro, RJ, 2007.

FREITAS, M. C.; QUENTAL, V. Subsídios para a elaboração automática de taxonomias. In: WORKSHOP DE TECNOLOGIAS DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 5, 2007, Rio de Janeiro. *Anais*... 2007. p. 1585-1594.

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. Dependency-based open information extraction. In: JOINT WORKSHOP ON UNSUPERVISED AND SEMI-SUPERVISED LEARNING IN NLP, 9, 2012, Avignon. *Proceedings*... Avignon, France: Association for Computational Linguistics, 2012. p. 10-18.

GRUBER, T. *Ontolingua: a mechanism to support portable ontologies*. 1992. 61p. Technical Report - Knowledge Systems Laboratory, Stanford University, 1992.

HEARST, M. Automatic acquisition of hyponyms from large text *corpora*. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14, 1992, Nantes. *Proceedings*... v. 2. Nantes, France: Association for Computational Linguistics. p. 539-545. DOI: <http://dx.doi.org/10.3115/992133.992154>

HEARST, M. Automated discovery of WordNet relations. In: FELLBAUM, C. (Org.) *WordNet*: an electronic lexical database. Cambridge: MIT Press, 1998. p. 131-153.

JURAFSKY, D.; MARTIN, J. *Speech and language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. 2. ed. Upper Saddle River: Prentice Hall, 2009.

MAEDCHE, A.; STAAB, S. *Ontology learning for the semantic web*. Massachusetts: Kluwer Academic Publishers, 2002. DOI: <http://dx.doi.org/10.1007/978-1-4615-0925-7>

MORIN, E.; JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. *Computer and the humanities*, Dordrecht, v. 38, n. 4, p. 363-396, 2003.

SANTOS, N.; OLIVEIRA, C. Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais. In: Congresso da Sociedade Brasileira de Computação, 25, 2005, São Leopoldo. *Anais...* Workshop de Tecnologias da Informação e da Linguagem Humana, 3, São Leopoldo, RS, 2005. p. 2138-2147.

TABA, L.; CASELI, H. Automatic semantic relation extraction from Portuguese texts. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9, 2014, Reykjavic. *Proceedings...* Reykjavic, 2014. p. 2739-2746.