Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Em direção ao delineamento conceitual de *corpus* via indexação léxico-conceitual

Andressa C. I. Zacarias
Ariani Di Felippo
Thiago A. S. Pardo

NILC-TR-02-04

Julho, 2012

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Na Sumarização Automática Multidocumento (SAM), produzem-se sumários a partir de coleções de textos que tratam de um mesmo assunto. Uma das estratégias utilizadas para produzir esses sumários consiste na seleção das sentenças que contêm os conceitos lexicalizados mais relevantes da coleção (LI et al., 2010). Para tanto, as unidades lexicais dos textos-fonte são mapeadas aos conceitos de uma ontologia de domínio construída de forma manual e, diante da identificação dos conceitos mais representativos, as sentenças que os contêm são selecionadas para o sumário. Visto que a construção de ontologias é uma tarefa cara, investigou-se o delineamento conceitual de *corpora* multidocumento pela indexação de suas unidades lexicais a uma ontologia construída previamente. Para tanto, selecionou-se a coleção C1 do *corpus* multidocumento em português CSTNews (CARDOSO et al., 2011). A C1 é composta por 3 notícias da seção "mundo" que relatam a "queda de um avião no Congo". Dessa coleção, 38 palavras da classe dos nomes foram selecionadas para a indexação, pois os nomes veiculam o conteúdo principal dos textos. As palavras foram manualmente indexadas à WordNet de Princenton (WN.Pr) (FELLBAUM, 1998), ontologia construída para o inglês. Especificamente, a indexação de uma palavra *x* englobou as seguintes tarefas: (i) tradução de *x* para o inglês, (ii) seleção do conceito da WN.Pr correspondente a *x*; (iii) associação da frequência de ocorrência de *x* na coleção C1 ao conceito selecionado em (ii), (iv) herança dos hipônimos e hiperônimos do conceito selecionado em (ii), e (v) propagação da frequência de *x* aos hiperônimos e hipônimos. Após a indexação das 38 palavras, obteve-se uma hierarquia conceitual, cujos conceitos foram pontuados em função da frequência de ocorrência em C1 das palavras que os expressaram linguisticamente. Diante da hierarquia, investigaram-se estratégias de poda que permitissem delimitar a região da hierarquia que englobava os conceitos mais representativos de C1. Para identificar os conceitos mais representativos e podar os de menor importância, adotou-se um dos critérios de Raimer e Hah (1988, apud MANI, 2001), a saber: frequência relativa do conceito na ontologia. Para tanto, optou-se pela estratégia de poda no sentido *top-down*, com base na qual se partiu dos conceitos mais genéricos em direção aos conceitos mais específicos. A pode especificamente consistiu em: (i) identificar a média das frequências dos conceitos do nível em questão; (ii) identificar uma porcentagem da média; (iii) podar os conceitos que apresentavam frequência menor que a obtida em (ii). Na etapa (ii), testaram-se 5 diferentes porcentagens sobre a média da frequência (30%, 40%, 50%, 60% e 70%), o que gerou 5 subontologias distintas. Para identificar qual a subontologia de menor tamanho englobava o maior número de conceitos representativos do conteúdo de C1, verificou-se quantos dos 13 conceitos nominais que ocorreram no sumário informativo (isto é, sumário que veicula a informação principal de seus textos-fonte) da C1 estavam presentes nas subontologias. Dessa comparação, verificou-se que a estratégia baseada na especificação de 50% da média (da frequência) dos conceitos em cada nível da hierarquia conceitual não podou os conceitos mais frequentes de C1. Assim, acredita-se que uma medida estatística pertinente para a delimitação da região da ontologia que engloba os conceitos mais representativos da coleção esteja em torno de 50% da média da frequência dos conceitos. Como trabalho futuro, pretende-se indexar outras coleções do CSTNews à WN.Pr para verificar se as estatísticas se confirmam.

Este trabalho contou com o apoio financeiro do CNPq.

1. Introdução

No âmbito do Processamento Automático das Línguas Naturais (doravante, PLN), os sistemas computacionais que processam (interpretam e/ou geram) língua natural registrada em meio escrito (p.ex.: tradutores e sumarizadores automáticos) podem ser concebidos como "sistemas baseados em conhecimento". Sob esse enfoque, tais sistemas somente conseguem automatizar uma tarefa linguística (p.ex.: tradução ou sumarização) quando munidos de conhecimento **linguístico** (MITKOV, 2004).

Os manuais de PLN baseiam-se em uma hierarquia de tipos de conhecimento linguístico, elaborada com base em uma escala de abstração e complexidade, ou seja, quanto mais alto for o nível nessa escala, mais complexos serão a modelagem e o tratamento computacional do conhecimento (DIAS-DA-SILVA, 2006). No nível mais inferior da escala, está o conhecimento morfológico, seguido pelo sintático, semântico e pragmático-discursivo (Figura 1). Os tipos de conhecimento linguístico necessários a um sistema dependem da natureza da aplicação para a qual ele foi desenvolvido.

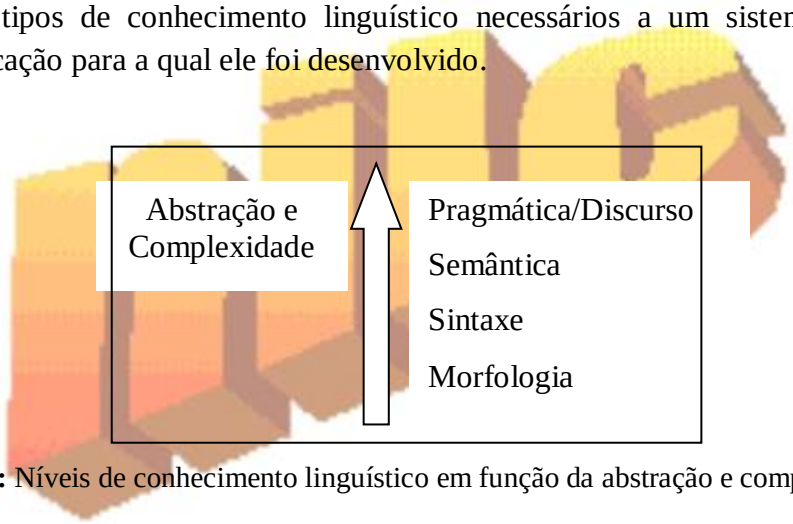

Figura 1: Níveis de conhecimento linguístico em função da abstração e complexidade.

Apesar de a arquitetura de um sistema de PLN variar de acordo com as especificidades da aplicação, o conhecimento linguístico necessário para a interpretação (e/ou geração) de um texto fica comumente armazenado em três bases de conhecimento: (i) gramatical, (ii) lexical e (iii) conceitual.

A base gramatical fornece a representação das regras sintáticas da língua, que podem ser vistas como condições de admissibilidade de estruturas sintáticas bem-formadas. A base conceitual ou ontologia fornece um inventário de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade" (isto é, o conhecimento de mundo compartilhado pelos membros de uma comunidade linguística). A base lexical fornece ao sistema uma coleção de unidades lexicais e seus principais traços morfológicos, sintáticos, semânticos e até pragmático-discursivos (MITKOV, 2004; DIAS-DA-SILVA, 2006). Algumas bases de conhecimento são tidas como "léxico-conceituais" ou "ontologias linguísticas" na medida em que armazenam apenas conceitos lexicalizados (em dada língua).

Atualmente, a maioria dos sistemas de PLN é composta por uma base gramatical e uma lexical, sendo que a lexical é responsável por armazenar, além das unidades lexicais, um conjunto mínimo de traços morfossintáticos que engloba, por exemplo, as informações de classe de palavra, gênero, número, etc. Consequentemente, os sistemas manipulam satisfatoriamente os conhecimentos linguísticos mais concretos, sendo capazes, por exemplo,

de: (i) reconhecer as palavras de um texto de entrada, (ii) identificar a classe das palavras e seus atributos morfossintáticos, (iii) construir a estrutura de constituintes das sentenças, (iv) identificar as funções sintáticas dos constituintes sentenciais e outras.

Alguns sistemas, quando munidos de uma base conceitual ou ontologia, também são capazes de manipular o conhecimento de nível léxico-semântico, sendo capazes de identificar, por exemplo, os conceitos subjacentes às unidades lexicais que ocorrem em um texto ou *corpus* (cf. 2.2.2) e as relações entre esses conceitos.

Na subárea do PLN denominada Sumarização Automática (SA), busca-se construir um sumário (resumo) a partir de um ou mais textos. Mani e Maybury (1999) sugerem que a sumarização automática, de um modo geral, envolva idealmente três etapas:

(i) análise dos textos-fonte, em que se produz uma representação completa de seu conteúdo;
(ii) transformação, em que o conteúdo completo do texto-fonte é condensado;
(iii) síntese, em que o conteúdo condensado é expresso em língua natural na forma de um sumário.

No caso da produção de extratos (isto é, sumários compostos por trechos inalterados do(s) texto(s)-fonte), a etapa de transformação engloba o processo de seleção de conteúdo, que consiste em reconhecer as unidades do texto-fonte (comumente sentenças) que contêm a ideia central do mesmo para compor o sumário (MANI, 2001).

Alguns sistemas de sumarização automática fazem uso de ontologias externas, ou seja, construídas previamente, para identificar as sentenças que expressam os principais tópicos do(s) texto(s)-fonte.

Para tanto, as unidades lexicais do(s) texto(s)-fonte são indexadas aos conceitos da ontologia e, partir de um ranqueamento das indexações, as sentenças que representam os conceitos mais bem pontuados são selecionadas para compor o sumário (REIMER, HAHN, 1988, *apud* MANI, 2001; HOVY, LIN, 1999; WU, LIU, 2003; HENNIG, 2008, LI *et al.*, 2010). Na maioria dos trabalhos, as ontologias são de domínio específico (p.ex.: produtos da marca Sony) e construídas previamente de forma manual ou semiautomática. No trabalho de Reimer e Hahn (1988, *apud* MANI, 2001), as indexações permitem a delimitação de uma subontologia, a qual engloba os conceitos mais representativos de um texto-fonte e, por isso, delimita conceitualmente o texto-fonte.

Diante da relevância de uma ontologia na interpretação dos textos, em especial, na seleção de conteúdo no processo de SA, propôs-se investigar o processo de delineamento conceitual de *corpora* em português do Brasil (PB) por meio da delimitação ou recorte de subontologias que englobam especificamente os conceitos mais representativos dos textos sob análise.

Para tanto, foram investigadas estratégias para indexar os conceitos provenientes de um *corpus* a uma ontologia e estratégias para delimitar a subontologia pertinente. Com isso, pretendeu-se gerar subsídios para que o delineamento conceitual de *corpora* via delimitação de subontologias pudesse ser automatizado.

Na próxima Seção, serão descritas de maneira detalhada as etapas de revisão de literatura necessárias para o desenvolvimento do projeto.

2. Revisão da Literatura

Os tópicos abordados na revisão bibliográfica foram: (i) conceitos gerais sobre sumarização automática, (ii) ontologia e (iii) indexação[1].

2.1 A sumarização automática: conceitos básicos

A sumarização é uma atividade bastante comum. Na modalidade escrita, tem-se, por exemplo, notícias de jornal e as sinopses de filmes. Os sumários produzidos a partir de textos são úteis porque podem ser indexadores, permitindo que o leitor descubra o assunto do texto-fonte correspondente, ou podem ser suficientemente informativos a ponto de permitirem que o leitor dispense a leitura do texto de origem (MARTINS *et al.*, 2001). Os sumários também são úteis em várias tarefas de PLN: (i) recuperação de informação, (ii) categorização de textos, etc.

Diante da utilidade dos sumários, do crescimento de informação disponível (principalmente, via *web*) e dos avanços na área de PLN, é de grande interesse a automação do processo de sumarização, foco da subárea do PLN denominada Sumarização Automática (SA). Nela, busca-se produzir automaticamente sumários a partir de um ou mais textos-fonte, sendo um "sumário" entendido como a versão mais curta de um texto ou mais textos.

De acordo com a função, os sumários podem ser informativos, indicativos ou críticos (MANI, MAYBURY, 1999). Os informativos contêm as informações principais de um texto-fonte de forma coerente e coesa ao ponto de dispensar a leitura do texto-fonte. Os indicativos apenas dizem do que o texto-fonte trata. Os críticos apresentam a informação principal do texto-fonte e avaliações sobre ele. Quanto à forma, os sistemas de SA podem produzir extratos ou *abstracts* (SPARCK JONES, 1993). Os extratos são compostos por trechos inalterados do texto-fonte. Os *abstracts* apresentam partes reescritas do texto-fonte. Quanto ao número de textos-fonte, a SA pode ser monodocumento, a partir de um único texto, e multidocumento, a partir de uma coleção de textos (MCKEOWN, RADEV, 1995). A SA monodocumento é uma tarefa bastante consolidada no PLN, havendo sumarizadores para diversas línguas, inclusive para o português do Brasil (PB) (UZÊDA *et al.*, 2010). O interesse pela SAM é recente e tem se fortalecido com o aumento do volume de informação similar ou repetida disponível na *web*.

Quanto à abordagem, Mani (2001) destaca que a SA pode ser superficial ou profunda em função da quantidade e do nível de conhecimento linguístico envolvidos na sumarização. Na superficial, utiliza-se pouco ou nenhum conhecimento linguístico para produzir sumários; o conhecimento utilizado é o empírico/estatístico. Por exemplo, uma abordagem que produz um sumário a partir da seleção e justaposição das sentenças do texto-fonte que apresentam as palavras mais frequentes do texto é classificada como "superficial". Sumarizadores superficiais costumam produzir extratos. A abordagem profunda caracteriza-se pela utilização de conhecimento linguístico morfológico, sintático, semântico e/ou pragmático-discursivo na SA.

Assim, os sumarizadores profundos podem gerar extratos e *abstracts*. As abordagens superficiais e profundas podem ser mescladas, originado abordagens híbridas. Idealmente, a SA é realizada em três etapas: (i) análise dos textos-fonte, em que se produz uma representação

[1] A investigação do delineamento conceitual (ou seja, da delimitação de subontologias), originalmente prevista neste projeto, integra a revisão da bibliografia do novo projeto apresentado na Parte II deste relatório.

completa de seu conteúdo; (ii) transformação, em que o conteúdo completo do texto-fonte é condensado e (iii) síntese, em que o conteúdo condensado é expresso em língua natural na forma de um sumário (MANI, 2001). Essas etapas devem ser guiadas pela taxa de compressão, ou seja, pelo tamanho desejado do sumário.

Atualmente, há dois *sumarizadores multidocumento*, ou seja, sistemas que produzem um sumário a partir de uma coleção de textos-fonte que abordam um mesmo tópico: o GIST SUMMarizer (PARDO, 2005), que é baseado em conhecimento linguístico superficial, e o CST SUMMarizer (JORGE, PARDO, 2010), que produz sumários com base na identificação de relações discursivas entre as sentenças dos textos de uma coleção.

Vê-se, assim, que, de um lado, tem-se um sistema baseado em conhecimento superficial e, de outro, um sistema baseado em conhecimento discursivo, considerado mais abstrato e complexo que o semântico. Não há, nesse cenário, trabalhos que se baseiam no tratamento léxico-conceitual do *corpus*. Um delineamento conceitual de *corpus* como proposto pode auxiliar a seleção de conteúdo nesses sistemas, que é realizada na etapa de transformação.

2.2 Ontologia

Na arquitetura de um sistema de PLN, pode estar prevista uma base de conhecimento conceitual. As bases conceituais contêm um "modelo do mundo" ou uma abstração da realidade, em que são descritos tipos de objetos, eventos, propriedades e relacionamentos entre esses tipos. Esse tipo de base desempenha um papel fundamental nos sistemas de PLN porque limita a "visão de mundo" simulada por eles (DIAS-DA-SILVA, 2006). Em outras palavras, uma base conceitual armazena o que se denomina "ontologia". O objeto "ontologia" é questão controversa em várias áreas, havendo, portanto, uma grande flutuação definicional e terminológica, como demonstram Guarino e Giaretta (1995). Uma das definições mais usuais é a de "uma especificação de uma conceitualização (ou seja, "uma visão simplificada do mundo") caracterizada por propriedades formais (explícitas) e propósitos específicos" (GRUBER, 1993, NIRENBURG, RASKIN, 2004). No âmbito do PLN, dois tipos de ontologia podem ser claramente identificados: *ontologias linguísticas* e as *ontologias conceituais* (VOSSEN, 1998, PALMER, 2001, FARRAR, BATEMAN, 2005). Essa distinção é feita com base no tipo de conceito que armazenam e no nível de formalização do conhecimento.

As *ontologias linguísticas* caracterizam-se por armazenar apenas conceitos lexicalizados (em uma determinada língua), isto é, conceitos expressos por uma ou mais palavras de uma língua, e as relações entre eles. Assim, esse tipo de recurso é um inventário dos sentidos de dada língua, ou seja, é um inventário somente daqueles conceitos compartilhados por uma comunidade linguística. Assim, uma ontologia linguística do holandês, por exemplo, não armazena o conceito [container][2], pois este não é lexicalizado nessa língua (VOSSEN, 1998). Quanto ao nível de formalização, os conceitos não são descritos por meio de um formalismo (p.ex.: lógica). Diz-se que as ontologias linguísticas são um tipo especial de ontologia porque armazenam conceitos lexicalizados (em dada língua) e não são objetos formais (MAGNINI, SPERANZA, 2002). A ontologia linguística mais difundidas no PLN é a WordNet de Princeton (WN.Pr) (FELLBAUM, 1998) (cf. 2.2.2c), desenvolvida para o inglês norte-americano.

[2] Os colchetes são usados neste texto para indicar que se trata de um conceito e não de uma expressão linguística.

As *ontologias conceituais* caracterizam-se pelo armazenamento de conceitos para os quais não há lexicalizações, ou seja, não há unidades lexicais que os representem, por exemplo: [coisa parcialmente temporal] e [partes do corpo humano] (VOSSEN, 1998, PALMER, 2001). Esses níveis são inseridos para que se alcance uma estruturação mais controlada dos conceitos dada a aplicação para a qual a ontologia foi feita. Além de apresentar níveis particulares para conceitos não lexicalizados, as ontologias conceituais podem negligenciar conceitos lexicalizados que não são relevantes para seus propósitos.

Quanto ao nível de formalização, essas ontologias descrevem os conceitos e as relações entre eles por meio de um formalismo. Atualmente, há vários formalismos baseados na lógica de primeira ordem e no XML (em inglês, *EXtensible Markup Language*) (ANTONUIOU, HARMELE, 2004). Segundo Palmer (2001), a ontologia conceitual mais difundida no PLN é a CYC (LENAT, GUHA, 1990).

Além dessa classificação, as ontologias podem ser categorizadas em função da especificidade do conhecimento nela organizado. Com base nesse critério, há as chamadas (i) *top-level ontology* ou *upper ontology*, que representa os conceitos mais genéricos, (ii) *middle-ontology*, que codifica conceitos intermediários, e (iii) *domain ontology* (ou *lower ontology*), que armazena conceitos específicos de um domínio do conhecimento (STUCKENSCHMIDT *et al.*, 2009). No projeto TermiNet (DI-FELIPPO, 2010), por exemplo, foi construída uma ontologia linguística em PB para o domínio da Educação a Distância no formato da WN.Pr, a WordNet.EaD. Na Figura 1, é apresentada parte da hierarquia dos conceitos expressos por nomes[3]. Nessa Figura, é possível observar os diferentes níveis de conhecimento conceitual.

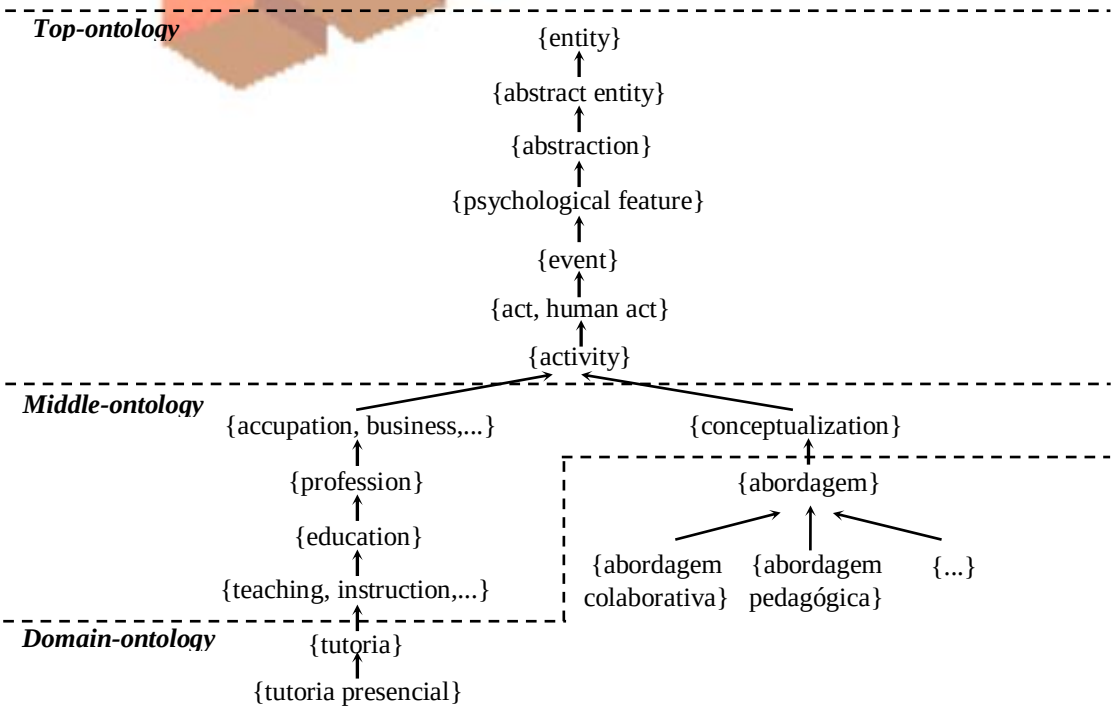

Figura 1: Diferentes níveis conceituais na WordNet.EaD.

[3] Os conceitos representados por rótulos em inglês foram herdados da WN.Pr.

2.3 Indexação léxico-conceitual

A indexação léxico-conceitual pode ser entendida como o processo de mapear unidades lexicais aos conceitos de uma ontologia. No cenário da SA monodocumento, Hovy e Lin (1999) propuseram mapear as sentenças (isto é, as unidades lexicais da sentença) de um texto-fonte aos conceitos de uma ontologia para, por meio da estrutura hierárquica da ontologia, identificar o conceito mais genérico que representa o conteúdo textual. Para tanto, os autores utilizaram a WN.Pr.

O mapeamento no trabalho de Hovy e Lin (1999) consiste em, dado um texto-fonte em inglês, selecionar as palavras de classe aberta e determinar a frequência de ocorrência das mesmas no texto. Na sequência, tais unidades são indexadas ou ligadas a conceitos da WN.Pr, os quais herdam a frequência das unidades referida no texto. Feito isso, a frequência de cada conceito ativado na ontologia é propagada aos conceitos hiperônimos (superordenados) e hipônimos (subordinados). Para identificar o conceito genérico mais específico dentre todos que foram ativados na WN.Pr, os autores utilizam uma medida estatística que calcula o nível conceitual mais adequado para representar o tópico do texto. Por meio de experimentos, os autores concluíram que o nível 6 da hierarquia conceitual dos nomes da WN.Pr armazena em geral os conceitos pertinentes.

Ressalta-se que os conceitos são representados na WN.Pr por conjuntos de formas sinônimos (*synsets*) (p.ex.: {car; auto; automobile; machine; motorcar}) (cf. 2.2.2.c) e, por isso, a probabilidade das unidades lexicais provenientes de um texto serem indexadas ou mapeadas a um conceito da ontologia é maior, ao contrário do que acontece com ontologias cujos conceitos são expressos por apenas um rótulo (p.ex.: as ontologias das Figuras 2 e 3).

No processo de sumarização monodocumento proposto por Wu e Liu (2003), a sumarização pode ser feita por meio de dois métodos. Um deles, que pode ser classificado como profundo, identifica os principais tópicos e subtópicos de um texto-fonte e, a partir deles, seleciona os parágrafos que contêm tais informações topicais para compor o sumário.

A identificação topical é feita pela comparação das unidades lexicais que ocorrem nos parágrafos aos conceitos da ontologia. Para a proposição dos métodos, os autores construíram um *corpus* composto por 51 artigos publicados originalmente *New York Times* ou no *The Wall Street Journal*, os quais foram coletados de uma base denominada ProQuest por meio da *query* (isto é, termo de busca) SONY.

Quanto à ontologia, os autores não fornecem detalhes sobre sua construção, eles apenas relatam que termos e sinônimos dos domínios relacionados à *Sony Corporation* (p.ex.: produção, informação financeira, competidores, etc.) compõem a ontologia. No total, a ontologia é composta por 142 conceitos/termos organizados hierarquicamente, na forma de uma árvore. No caso, diz-se que se trata de uma ontologia de domínio que armazena, por exemplo, (i) conceitos (p.ex.: Sony, Sony Music e Sony Pictures), e (ii) relações de subsunção (p.ex.: o conceito Sony subsume os conceitos mais específicos Sony Music e Sony Pictures). Por se tratar de uma árvore conceitual, diz-se que os conceitos são os nós ou folhas e as relações são os galhos.

A Figura 2 ilustra os conceitos mais genéricos[4] da referida ontologia.

[4] Os conceitos mais genéricos, dispostos nos níveis superiores de uma ontologia, constituem uma *top-ontology*.

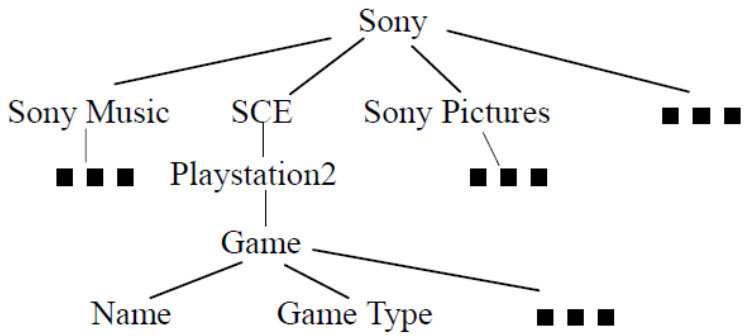

Figura 2: *Top-level ontology* do domínio *Sony Corporation* (WU, LIU, 2003).

Para que os parágrafos possam ser pontuados em função da informação topical que expressam, é preciso comparar as unidades lexicais que neles ocorrem aos conceitos da ontologia. No trabalho de Wu e Liu (2003), as unidades provenientes de um texto-fonte que não estão armazenadas na ontologia são descartadas. Caso a unidade lexical esteja presente na ontologia, é feita a indexação ou ligação da mesma na ontologia e o conceito indexado é pontuado. Quando se pontua um conceito na hierarquia, seus conceitos superiores são automaticamente pontuados. Por exemplo, na ontologia em questão, "Spider-man" é um nó-filho do nó-pai "movie", assim, se um parágrafo contiver o termo "Spider-man", ambos os conceitos, "Spider-man" e "movie", são pontuados na ontologia.

Por meio dessa estratégia de indexação e pontuação dos conceitos, o conceito mais genérico, o qual inicia a *top-ontology* (p.ex.: Sony), terá sempre a pontuação mais elevada, enquanto os conceitos do segundo nível, que representam subtópicos, terão pontuações diferentes. Com isso, após a pontuação dos conceitos por meio das indexações, apenas os conceitos mais bem pontuados localizados no segundo nível da hierarquia (p.ex: Sony Music, SCE, Sony Pictures, etc.) são selecionados para representar os subtópicos do texto/artigo.

Na sequência, os conceitos com maior pontuação são, então, selecionados como os principais tópicos do documento de origem e cada parágrafo é pontuado em função desses tópicos. Os parágrafos são selecionados até que o tamanho desejado do sumário seja alcançado. Ressalta-se que Wu e Liu (2003) consideram o parágrafo como unidade textual no processo de sumarização, e não sentenças, e utilizam apenas os rótulos dos conceitos para a indexação, e não sinônimos, por exemplo.

O método de sumarização monodocumento proposto por Hennig *et al.* (2008) também se baseia na utilização de ontologia para melhorar a interpretação das sentenças a fim de selecionar, ao final, as que contêm os conceitos mais importantes do texto. Para testar o método, os autores utilizaram uma ontologia genérica composta por 1036 conceitos, os quais são representados por rótulos simples, como *art*, *health*, *society*, etc. A cada conceito ou nó da ontologia inicial, foi acrescido um "saco de palavras" (do inglês, *bog-of-words*) proveniente de textos extraídos da *web*.

Por exemplo, dado o conceito [inteligência artificial] da ontologia inicial, o mesmo foi transformado em uma *query* e, a partir dos primeiros 20 textos retornados pelo motor de busca YAHOO, foram selecionadas, com base em várias medidas estatísticas, um conjunto de palavras (de classe aberta) para representar o conceito na ontologia.

Com base nessa ontologia modificada, o sistema mapeia cada sentença de um texto-fonte aos "sacos de palavras" (e, consequente, aos rótulos originais), indexando-a, ao final, aos rótulos originais da ontologia que mais adequadamente representam a sentença. De acordo com a similaridade entre a sentença e os "sacos de palavras", tais rótulos recebem pontuações distintas, sendo que o de pontuação mais alta indica o conceito que melhor representa a sentença.

Na Figura 3, vê-se que a sentença *"One person died today in Texas after being injured in a tornado"* foi indexada a três conceitos da ontologia, representados pelos rótulos (*tags*) *Weather*, *Breaking* e *News*. De acordo com a similaridade entre a sentença e os "sacos de palavras" associados aos conceitos, os rótulos receberam pontuações distintas, sendo *Weather* o de mais alta pontuação e, por conseguinte, o mais representativo do conteúdo da sentença.

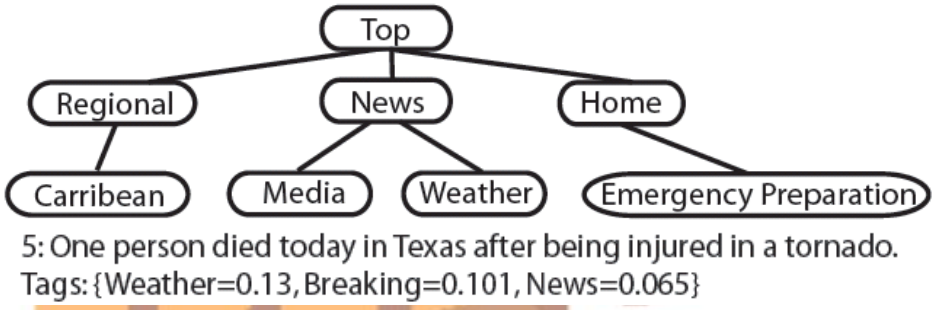

Figura 3: Exemplo de indexação em Henning *et al.* (2008).

No trabalho de Li *et al.* (2010), um método de sumarização multidocumento é proposto com base na inserção de conhecimento ontológico. Especificamente, os autores mapeiam ou indexam as sentenças de uma coleção ou *cluster* (isto é, conjunto de textos que versam sobre um mesmo assunto) aos conceitos de uma ontologia.

Dada a *query* de um usuário, a qual também é mapeada na ontologia, o sistema seleciona para compor o sumário, apenas as sentenças dos textos-fontes indexadas aos mesmos conceitos a que as unidades lexicais da *query* foram mapeadas e/ou aos conceitos mais específicos. Para tanto, os autores utilizaram uma ontologia linguística do domínio *desastre* construída manualmente por especialistas e cujos conceitos são expressos por rótulos únicos. Infelizmente, não há muitos detalhes em Li *et al.* (2010) de como efetivamente a indexação é feita.

Na Figura 4, vê-se, em (a), um exemplo em que as sentenças (elipses verdes) dos textos de uma coleção que versam sobre a passagem do furacão *Wilma* por Atlanta em 2005 foram indexadas aos conceitos da ontologia (elipses azuis) (no caso, *Transit*, *Bus*, *Rail* e *Airline*). Em (b), observa-se que, diante de uma *query* como "*get all the information related to transit in Miami-Dade County after Hurricane Wilma passed",* a qual teria sido indexada ao conceito *Transit*, apenas as sentenças indexadas a esse conceito e a seus subordinados, *Bus*, *Rail* e *Airline*, seriam selecionadas para compor o sumário.

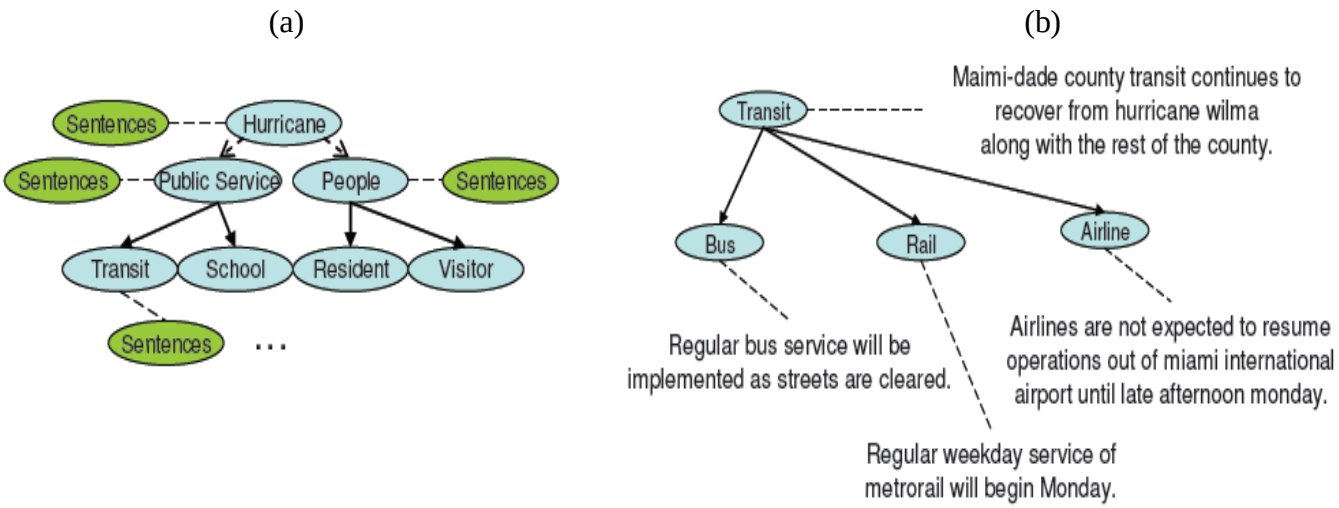

Figura 4: Exemplo de indexação e seleção de conteúdo em Li *et al.* (2010).

Por fim, Reimer e Hahn (1988, *apud* MANI, 2001) também utilizam o processo de indexação léxico-conceitual em sua proposta de sumarização monodocumento. Especificamente, os autores desenvolveram o TOPIC, um sumarizador para o alemão capaz de identificar os principais conceitos do texto e as relações que se estabelecem entre eles. Vale ressaltar, no entanto, que o TOPIC não chega propriamente a produzir sumários, apenas indica os trechos do texto-fonte em que os conceitos identificados estão expressos.

Para a sumarização de textos sobre o assunto *computadores*, o TOPIC inicialmente identifica o núcleo dos sintagmas nominais de um texto-fonte. Na sequência, indexa ou liga as unidades lexicais identificadas como núcleos sintagmáticos a uma estrutura ontológica em alemão do referido domínio que fora previamente construída por especialistas. O sistema, então, aumenta o peso do conceito na medida em que ele ocorre no texto e, por conseguinte, é indexado ou ativado na estrutura ontológica. No trabalho de Reimer e Hahn (1988, *apud* MANI, 2001), as indexações permitem que uma subontologia seja delimitada, a qual engloba os conceitos mais representativos de um texto-fonte. Pode-se dizer que essa subontologia delimita conceitualmente o texto-fonte.

3. Metodologia

A metodologia prevista para o projeto, seguiu as etapas descritas a seguir:

3.1 Seleção do *corpus*

Para a execução desta pesquisa, foi necessário um *corpus*. Por definição, um *corpus* é um conjunto de dados linguísticos sistematizados de acordo com determinados critérios, de maneira que possa ser processado por computador com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SINCLAIR, 2005). Por essa definição, um *corpus* é um artefato produzido para a pesquisa e que, por isso, a maioria de suas características é dependente dos objetivos da pesquisa.

Tendo em vista o trabalho ora descrito, o *corpus* tinha de apresentar as seguintes características: (i) *monolíngue*, especificamente do PB, já que este trabalho focaliza a investigação de estratégias de delineamento conceitual de *corpus* no cenário específico do processamento automático do PB, e (ii) *multidocumento*, posto que um *corpus* desse tipo fornece vários *clusters* ou coleções de textos que, ao versarem sobre um mesmo assunto ou tópico, tornam-se pertinentes como fontes para o delineamento conceitual.

Diante de tais características, buscou-se identificar na literatura um *corpus* que as satisfizesse. Nessa investigação, identificou-se o CSTNews (CARDOSO *et al.* 2011), único *corpus* multidocumento em PB. O CSTNews é composto por 50 coleções ou *clusters* em um total de 195 textos. Cada coleção trata de um domínio (p.ex.: esporte, política, dinheiro, etc.) e assunto específicos e contém em media 4 documentos. Os textos são do tipo informativo e gênero jornalístico, os quais foram coletados manualmente de jornais *on-line* (Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo) em um período de 2 meses durante o ano de 2007 (ALEIXO, PARDO, 2008).

O *corpus* CSTNews é assim denominado porque os textos das coleções estão alinhados por meio das relações estabelecidas pela teoria/modelo linguístico-computacional CST, de natureza semântico-discursiva (*Cross-document Structure Theory*) (RADEV, 2000). A CST permite estruturar o discurso por meio da conexão das unidades textuais (palavras, sintagmas, sentenças, etc.) provenientes de diferentes documentos que versam sobre um mesmo assunto (RADEV, 2000).

Tal conexão é rotulada por várias relações como: (i) *Identity* (isto é, relação entre dois segmentos textuais (S1 e S2) que possuem mesma forma e conteúdo), (ii) *Equivalence* (isto é, relação entre S1 e S2 que expressam o mesmo conteúdo), e (iii) *Subsumption* (isto é, relação entre S1 e S2, sendo que S1 expressa, além de todo o conteúdo de S2, informação adiciona), entre outras. O relacionamento de textos no nível semântico-discursivo é útil para diversas aplicações do PLN, como sumarização multidocumento, perguntas e respostas e extração de informação, entre outras. Além do alinhamento via CST, o CSTNews possui outros tipos de anotação (CARDOSO *et, al.*, 2011).

Para o fim desta pesquisa, foi feito um recorte no *corpus*, construindo-se um *subcorpus* do CSTNews. Esse recorte consistiu na seleção de 1 único *cluster* cujas unidades lexicais de seus textos constitutivos foram indexadas especificamente à ontologia selecionada. Tendo em vista o tempo de execução da pesquisa, optou-se por um *cluster* relativamente pequeno, o *cluster* 1 (C1), composto por 3 textos compilados dos jornais *on-line* Folha de São Paulo, Estadão e Jornal do Brasil. O C1 possui um total de 24 sentenças e 432 palavras (ALEIXO, PARDO, 2008). Nessa coleção, constam textos do domínio "mundo", sendo que os documentos relatam especificamente um desastre, no caso, a "queda de um avião no Congo". Assim, as unidades lexicais indexadas pertencem a ao domínio conceitual "desastre (de avião ou aéreo)".

3.2 Seleção das unidades lexicais

Para a indexação léxico-conceitual, ou seja, das unidades dos textos do *subcorpus* à ontologia, foi preciso delimitar quais unidades participariam desse processo. Para tanto, investigou-se dois critérios de seleção: (i) frequência e (ii) categoria sintática. Ao se adotar o critério da frequência, apenas as unidades mais frequentes no *cluster* seriam indexadas. Por consequência,

unidades que expressam diferentes tipos de conceitos estariam envolvidas no processo, posto que, entre as mais frequentes, estão os nomes e os verbos, os quais lexicalizam, na maioria das vezes, entidades e ações, respectivamente. Essa variedade de tipos conceituais tornaria as tarefas de identificação dos conceitos e indexação mais complexas.

Para que o processo de indexação fosse uma tarefa mais delimitada e contralada possível, optou-se pelo critério da categoria sintática, por meio do qual apenas as unidades da categoria dos nomes foram selecionadas no *cluster* e efetivamente indexadas. Essa restrição se deve a dois fatores: (i) os nomes estão entre as unidades mais frequentes do *cluster*, pois expressam, juntamente com os verbos, o conteúdo semântico principal dos textos, e (ii) os nomes lexicalizam conceitos que se organizam principalmente pela relação hierárquica da hiponímia, o que facilita a identificação dos conceitos e as indexações. Essa relação, aliás, é a única considerada no processo de indexação (cf. 2.2.3).

A relação de hiponímia é aquela que se estabelece entre, por exemplo, os conceitos lexicalizados [jeep] e [carro]. Nesse caso, [jeep] é hipônimo de [carro]. A relação inversa é a hiperonímia; assim, tem-se que [carro] é hiperônimo de [jeep]. Do ponto de vista conceitual, o conceito expresso por [carro] é mais genérico que o conceito expresso por [jeep] e, do ponto de vista extensional, a classe denotada por um hiperônimo inclui a classe denotada pelo hipônimo. Do ponto de vista lógico, a hiponímia é uma relação transitiva, pois, se A é um hipônimo de B e B é um hipônimo de C, então A é necessariamente hipônimo de C (CRUSE, 2004, 2006).

Para a seleção das unidades da categoria dos nomes, os textos da coleção previamente escolhida foram anotados no nível morfossintático. Para tanto, optou-se pela utilização de um etiquetador (em inglês, *tagger*), ou seja, ferramenta computacional responsável por associar às palavras de um texto ou sentença uma etiqueta que indica sua correta categoria sintática no contexto. O etiquetador selecionado foi o *LX-Tagger* (BRANCO, SILVA, 2004), disponível no portal do *LX-Center*[5]. Apesar de ter sido projetado para etiquetar textos em português europeu, o *LX-Tagger* foi selecionado devido à sua alta precisão (96,87%) e, sobretudo, sua interface *on-line* amigável. Em (1), ilustra-se a etiquetação de uma sentença do *subcorpus* pelo *LX-Tagger*.

(1) a_/LADV1 o/LADV2 menos/LADV3 17/DGT pessoas/PESSOA/CN#fp morreram/MORRER/V#pi-3p após/PREP a/DA#fs queda/QUEDA/CN#fs de/PREP um/UM#ms avião/AVIÃO/CN#ms de/PREP passageiros/PASSAGEIRO/CN#mp em_/PREP a/DA#fs República/PNM Democrática/PNM de_/PREP o/DA#ms Congo/PNM ./PNT

A partir da anotação, apenas as unidades etiquetadas como nomes foram automaticamente selecionadas do *cluster* C1, totalizando 42 nomes distintos. No caso da anotação realizada pelo *LX-Tagger*, a etiqueta utilizada para a seleção das unidades é CN (do inglês, *common* noun). Da sentença em (1), pertencente ao documento 1 (D1) do *cluster* C1, foram selecionadas, por exemplo, as unidades lexicais "pessoas", "queda", "avião" e "passageiros". Para a efetiva indexação, consideram-se apenas as formas canônicas das unidades, as quais foram recuperadas da própria etiquetação, na qual essas formas estão escritas em letras maiúsculas ("pessoa", "queda", "avião" e "passageiro"). Do conjunto de 42 unidades, os itens "junho", "março",

[5] Disponível em: http://lxcenter.di.fc.ul.pt/

“quinta-feira” e “sexta-feira” foram excluídos, pois não carregam conteúdo representativo do *cluster*. Assim, apenas 38 foram selecionadas a partir do *cluster* C1.

Após a identificação das unidades, verificou-se a frequência de ocorrência das mesmas no *cluster*. No Quadro 2, estão descritas as 38 unidades finais selecionadas do *cluster* C1 e suas respectivas frequências entre parênteses.

acidente (5)	distância (2)	montanha (1)	queda (1)	tripulante (1)
aeronave (1)	estrada (1)	nacionalidade (2)	quilômetro (3)	vítima (2)
aeroporto (4)	fabricação (3)	país (2)	setor (1)	
aterrissagem (2)	floresta (3)	passageiro (5)	sobrevivente (2)	
avião (11)	fonte (2)	permissão (1)	tarde (2)	
carga (2)	leste (2)	pessoa (3)	tempestade (1)	
chama (1)	localidade (2)	pista (3)	tempo (2)	
cidade (1)	membro (4)	porta-voz (7)	transporte (1)	
companhia (4)	mineral (2)	propriedade (2)	tripulação (4)	

Quadro 2: Unidades lexicais nominais e suas respectivas frequência no *cluster* C1.

3.3 Seleção da ontologia

Após a seleção do *corpus* e da categoria sintática (consequentemente, das unidades lexicais), foi preciso construir e/ou selecionar a ontologia. Seguindo a maioria dos trabalhos da literatura, a primeira opção foi a de construir uma ontologia que satisfizesse os requisitos do projeto. No entanto, essa tarefa não é simples, principalmente porque tal ontologia deveria contemplar os diferentes domínios e assuntos presentes no CSTNews. Mesmo que se optasse por construir uma ontologia apenas para o C1, tal decisão não nos pareceu pertinente, pois, a cada *cluster* distinto a ser processado, seria necessário construir uma ontologia específica.

Assim, passou-se a investigar a segunda opção, a de selecionar uma ontologia já existente. Atualmente, é possível encontrar várias ontologias disponíveis em PB, principalmente através do *OntoLP – Portal de Ontologias*[6]. Tais recursos, no entanto, são ontologias de domínio (*domain ontologies*) e representam conceitos, por exemplo, de áreas como nanotecnologia, música, ecologia, etc. Tendo em vista que os textos do CSTNews são de domínios variados, tais ontologias também não são adequadas.

Além desses recursos, salienta-se o projeto de construção da WordNet.Br (WN.Br) (DIAS-DA-SILVA *et al.*, 2008), que segue o formato da WN.Pr. A base atual da WN.Br contém aproximadamente 44.000 unidades lexicais e 18.500 *synsets*. Tais unidades e *synsets* estão distribuídos em 11.000 verbos (4.000 *synsets*), 17.000 nomes (8.000 *synsets*), 15.000 adjetivos (6.000 *synsets*) e 1.000 advérbios (500 *synsets*). De todas as relações prevista na WN.Pr, a WN.Br codifica, até o momento, apenas a relação de antonímia, que conecta principalmente conceitos expressos por adjetivos. Assim, a WN.Br também não é adequada à pesquisa em questão, já que o foco deste trabalho reside nos nomes e na relação de hiponímia.

Diante de falta de disponibilidade de ontologias em PB que pudessem ser utilizadas na pesquisa, optou-se pela utilização da WN.Pr devido aos seguintes fatores:(i) *acessibilidade*,

[6] http://www.inf.pucrs.br/ontolp/downloads.php

posto que seu arquivo-fonte está disponibilizado integralmente via *web*[7]; (ii) *pertinência linguística*, posto que a organização do conhecimento léxico-conceitual pauta-se em pressupostos da Psicolinguística e os dados foram manualmente compilados de vários dicionários monolíngues do inglês norte-americano, e (iii) *abrangência*, já que armazena, em sua versão atual (3.0), um total de 155.287 unidades lexicais, distribuídas em 117.798 substantivos, 11.529 verbos, 21.479 adjetivos e 4.481 advérbios.

Especificamente, a WN.Pr é o resultado das pesquisas iniciadas em meados da década de 1980 pelos pesquisadores do Laboratório de Ciência Cognitiva da Universidade de Princeton (EUA). Impulsionados por pressupostos psicolinguísticos sobre a organização do léxico mental, tais pesquisadores decidiram construir uma base lexical cujas unidades lexicais fossem organizadas em função do seu significado. Essa iniciativa deu origem, no início da década de 90, à WN.Pr.

Na WN.Pr, as unidades lexicais (palavras ou expressões) do inglês norte-americano estão divididas em quatro categorias sintáticas: nome, verbo, adjetivo e advérbio. As unidades de cada categoria estão codificadas em *synsets* (do inglês, *synonym sets*), ou seja, em conjuntos de formas sinônimas ou quase-sinônimas (p.ex.: {car; auto; automobile; machine; motorcar}). Cada *synset* é, por definição, construído de modo a representar um único conceito lexicalizado por suas unidades constituintes. Assim, não é preciso explicitar o valor semântico de cada conjunto de sinônimos por meio de um rótulo conceitual. Os *synsets* estão inter-relacionados pela relação léxico-semântica da antonímia[8] e pelas relações semântico-conceituais da hiperonímia/ hiponímia, holonímia/ meronímia, acarretamento e causa.

A Figura 5 ilustra a noção de synset e algumas relações lógico-conceituais. Nessa figura, cujo exemplo foi extraído da WN.Pr, observa-se que o synset {car; auto; automobile; machine; motorcar} está relacionado a:

(i) conceito mais geral ou *synset* hiperônimo {motor vehicle; automotive vehicle};
(ii) conceito mais específico ou *synset* hipônimo {cruiser; squad car; patrol car; police car; prowl car} e {cab; taxi; hack; taxicab};
(iii) partes que o compõem ou *synsets* merônimos (p.ex.: {bumper}, {car door}, {car mirror} e {car window}).

A WN.Pr também registra outras informações, ditas adicionais, a saber:

(i) para cada unidade lexical, há uma frase-exemplo para ilustrar o seu contexto de uso, p.ex.: para "car", no *synset* {car; auto; automobile; machine; motorcar}, há a frase-exemplo "he needs a car to get to work" ("ele necessita de um carro para ir trabalhar");
(ii) para cada *synset*, há uma glosa que especifica informalmente o conceito por ele lexicalizado, p.ex.: para o *synset* {car; auto; automobile; machine; motorcar}, há a glosa "a motor vehicle with four wheels; usually propelled by an internal combustion engine" ("um veículo com quatro rodas; usualmente impulsionado por um motor de combustão interno");
(iii) para cada *synset*, há também a especificação do tipo semântico expresso pelo conceito a ele subjacente; p.ex.: o *synset* {bicycle; bike; wheel; cycle} é do tipo semântico <noun.artifact>.

[7] http://wordnet.princeton.edu/man/wnstats.7WN.
[8] A antonímia é uma relação entre unidades lexicais, ou seja, formas linguísticas. A relação de antonímia entre *synsets* (ou conceitos) indica, na verdade, uma oposição conceitual e não uma antonímia propriamente.

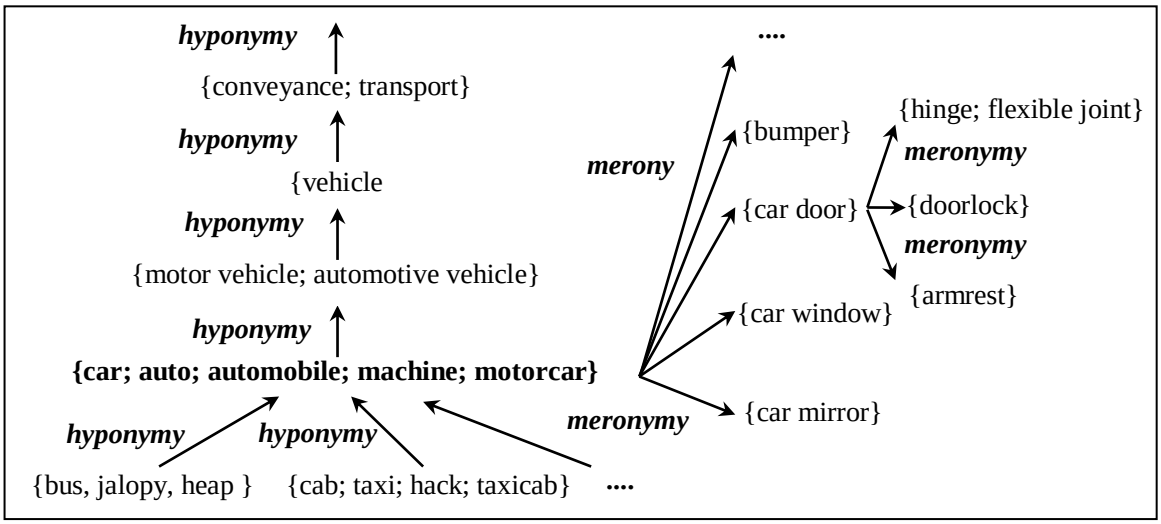

Figura 5: Ilustração do conceito de *synset* e das relações léxico-conceituais na WN.Pr (v.3.1).

Como mencionado, na WN.Pr, as unidades lexicais estão organizadas em quatro categorias sintáticas. Cada uma delas constitui uma base lexical própria, em que os *synsets* estão organizados por relações semântico-conceituais específicas, responsáveis pela estruturação interna da base. O Quadro 3, baseado em Fellbaum (1998), resume o conjunto principal de relações em função das categorias sintáticas.

Relações	**Categorias sintáticas**	**Exemplos**
Antonímia (oposição conceitual)	Adj, Adv N, V	*mulher* é antônimo de *homem*[9] *claro* é antônimo de *escuro* *rapidamente* é antônimo de *lentamente* *descer* é antônimo de *subir*
Hiponímia/ Hiperonímia (subordinação)	N	*veículo* é hiperônimo de *carro* *carro* é hipônimo de *veículo*
Meronímia/ Holonímia (parte-todo)	N	*carro* é holônimo de *roda* *roda* é merônimo de *carro*
Troponímia (modo)	V	*sussurrar* é tropônimo de *falar*
Acarretamento	V	*correr* acarreta *deslocar-se*
Causa	V	*matar* causa *morrer*
Legenda: N= nome; V= verbo; Adj=adjetivo; Adv=advérbio		

Quadro 3: As relações semânticas da WN.Pr em função das categorias sintáticas.

3.3.3. Indexação das unidades lexicais à WN.Pr

Como mencionado, a indexação léxico-conceitual neste projeto consiste em ligar as unidades provenientes de um *cluster* à WN.Pr. Diante da seleção da WN.Pr, a indexação léxico-conceitual passou a englobar a tradução das unidades lexicais extraídas do C1 do CSTNews

[9] Na WN.Pr, o *synset* {man, adult male} é considerado antônimo (ou seja, oposto conceitual) do *synset* {woman, adult female}.

para o língua inglesa. Especificamente, a indexação manual foi feita por meio dos seguintes passos metodológicos:

a) tradução das unidades extraídas para o inglês;
b) busca pelos *synsets* da WN.Pr que contêm o termo em inglês;
c) identificação do conceito/*synset* adequado;
d) associação da frequência da unidade ao conceito/*synset* selecionado em (c);
e) seleção de todos os hiperônimos do *synset* selecionado em (c);
f) propagação da frequência do *synset* ativado na ontologia aos hiperônimos;
g) seleção dos hipônimos imediatos do *synset* selecionado em (c);
h) propagação da frequência do *synset* ativado na ontologia aos hipônimos;
i) unificação das hierarquias parciais e das frequências dos *synsets* constitutivos.

Para ilustração, toma-se como ponto de partida a primeira indexação feita no projeto, a da unidade lexical "acidente". Para a tradução, em especial, foram utilizados dois dicionários bilíngues português-inglês: (i) a versão *on-line* do *Michaelis: moderno dicionário inglês* (WEISZFLOG, 2000), disponível através do portal UOL[10] e (ii) o dicionário *on-line WordReference*[11]. Além desses recursos, a tradução também contou, quando necessário, com outros dois serviços *on-line* de tradução: o *Linguee*[12] e o *Google Translator*[13]. No caso da unidade "acidente", ambos os dicionários bilíngues sugerem "accident" como equivalência.

De acordo com a metodologia, o próximo consiste em buscar a unidade traduzida na WN.Pr. Nessa base, a unidade "accident" é elemento constitutivo de dois *synsets*, a saber: (i) {accident}, cuja glosa[14] é "*a mishap; especially one causing injury or death*", e (ii) {accident, fortuity, chance event}, definido informalmente como "*anything that happens by chance without an apparent cause*". Ao constituir dois *synsets*, a unidade "accident" lexicaliza em inglês dois conceitos. Dessa forma, é preciso identificar o conceito expresso nos textos do *cluster* C1. Com base nas glosas e nos hiperônimos de cada *synset*, foi possível identificar que {accident} é o *synset* adequado para representar o conceito em questão. Na sequência, associa-se a frequência da unidade lexical "acidente" no *cluster* C1 ao *synset* {accident}; no caso, {accident} é associado à frequência 5.

No próximo passo, todos os hiperônimos de {accident} são identificados e extraídos, assim como os hipônimos imediatos. No caso, os hiperônimos são {mishap, misadventure, mischance}, {misfortune, bad luck}, {trouble}, {happening, occurrence, occurrent, natural event}, {event}, {psychological feature}, {abstraction}, {abstract entity} e {entity}, e os hipônimos imediatos são {collision}, {crash, wreck}, {injury, accidental injury}, {shipwreck, wreck} e {fatal accident, casualty}.

As Figuras 6, 7, 8, 9, 10 e 11 ilustram, respectivamente, os passos (a), (b), (c), (d), (e-f) e (g-h), os quais foram realizados manualmente.

[10] http://michaelis.uol.com.br/
[11] http://www.wordreference.com/
[12] http://www.linguee.com.br
[13] http://translate.google.com/
[14] O termo *glosa* indica uma definição informal do conceito subjacente aos *synsets*.

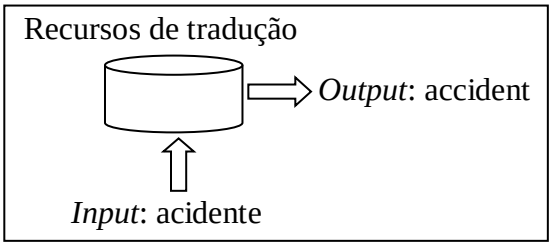

Figura 6: Tradução (a).

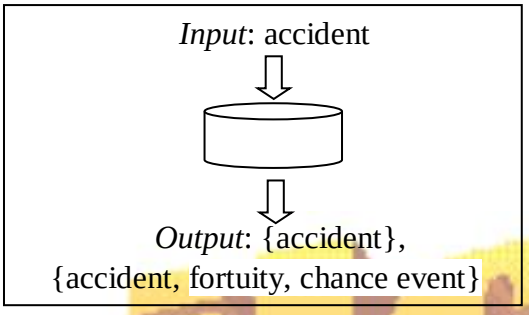

Figura 7: Busca na WN.Pr (b).

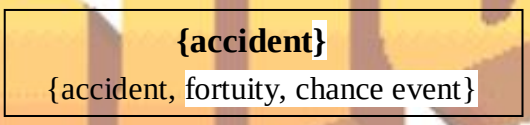

Figura 8: Identificação do conceito/ *synset* (c).

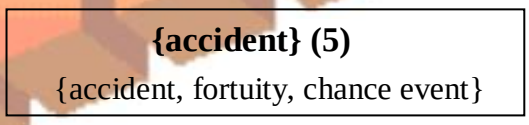

Figura 9: Associação da frequência (d).

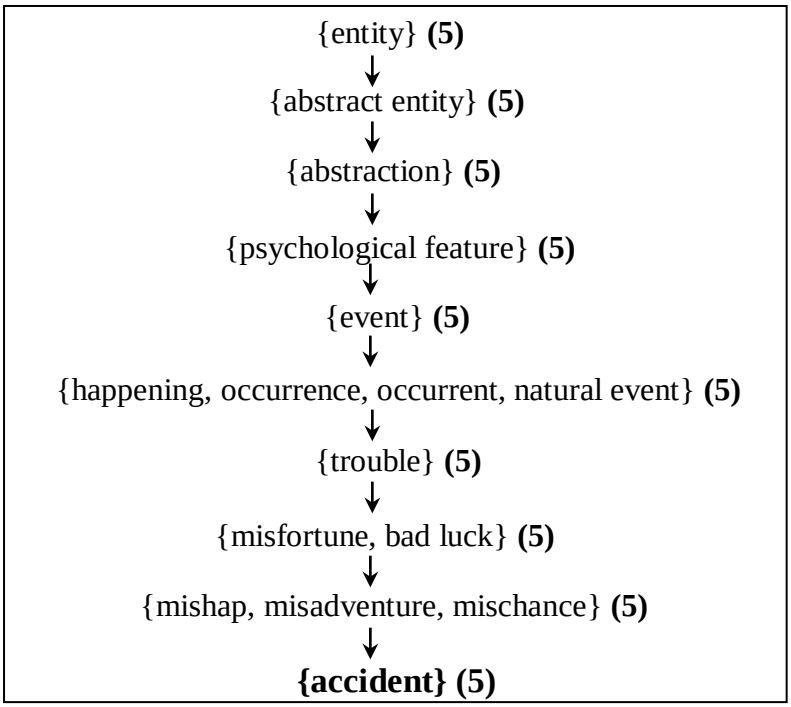

Figura 10: Seleção de todos os hiperônimos (e) e propagação da frequência (f).

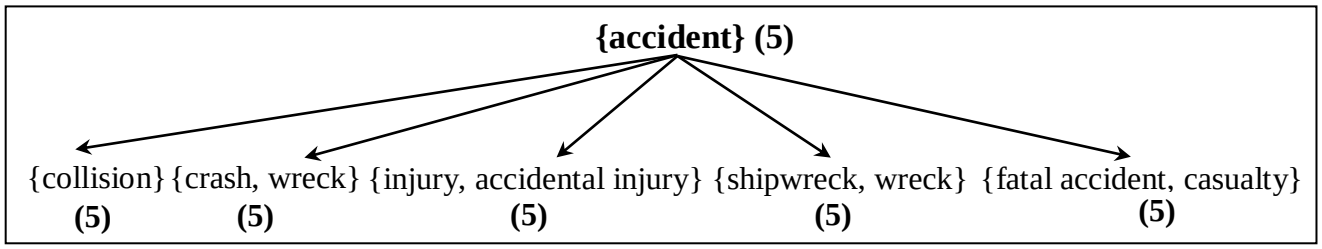

Figura 11: Seleção dos hipônimos imediatos (g) e propagação da frequência (h).

Ao realizar os passos de (a) e (h) para as 38 unidades extraídas do *cluster* C1, foram obtidas 38 hierarquias parciais da WN.Pr.

Atualmente, está sendo feita a tarefa descrita em (i), ou seja, a unificação dessas hierarquias parciais e das frequências de seus elementos constitutivos quando repetidos. Esse processo é o único que está sendo feito de forma semiautomática por meio da colaboração com os cientistas computacionais do NILC[15].

Dessa forma, pretende-se obter uma única hierarquia em que os conceitos estarão devidamente pontuados em função de sua ativação total na ontologia.

Na Figura 12b, tem-se a hierarquia obtida por meio da indexação da unidade lexical "queda".

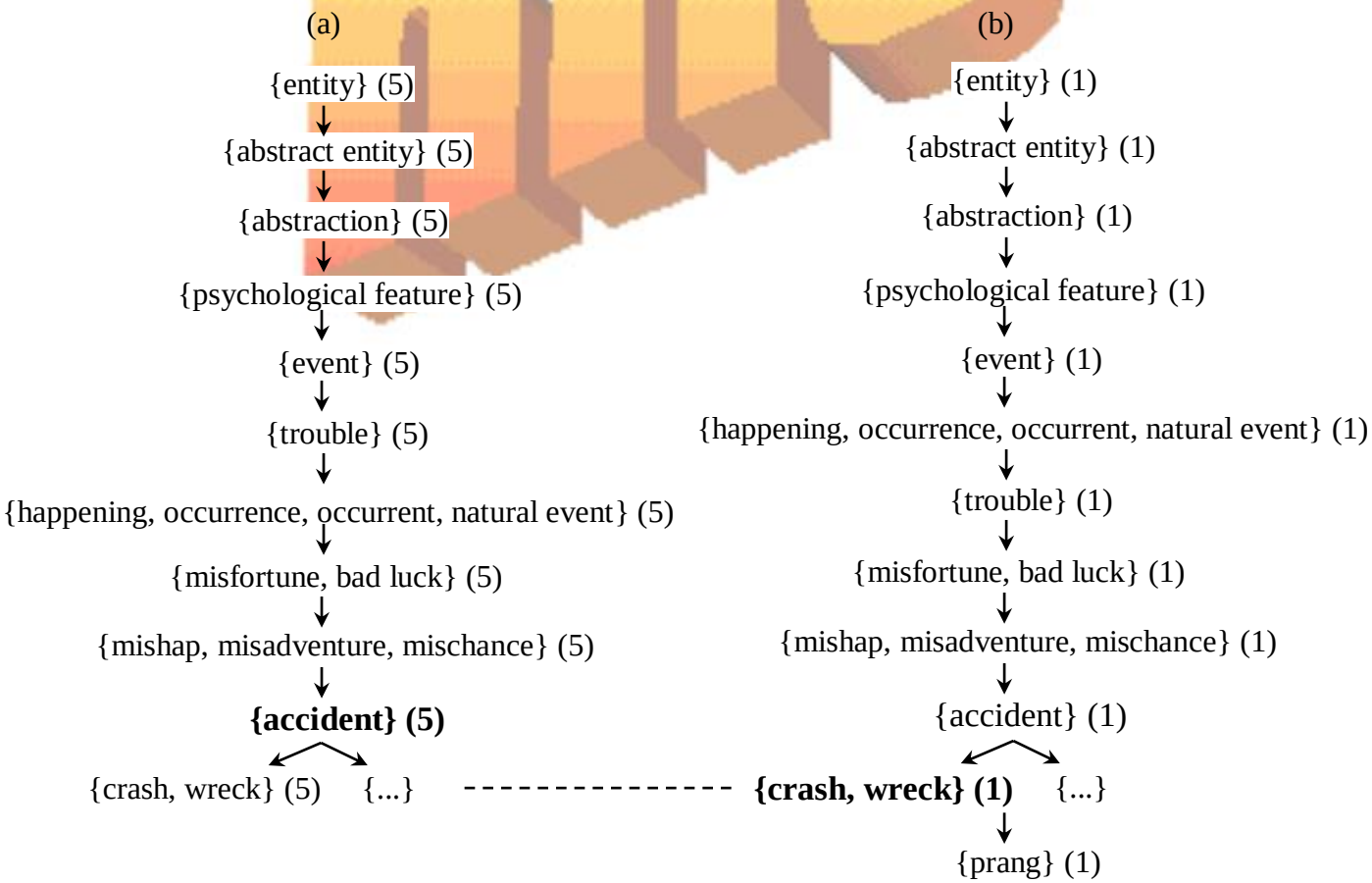

Figura 12: Exemplos de hierarquias parciais.

[15] Este projeto de iniciação tecnológica está sendo realizado no Núcleo Interinstitucional de linguística Computacional (NILC) (http://www.nilc.icmc.usp.br/nilc/index.html), do qual a orientadora e o coorientador são pesquisadores seniores.

Observa-se especificamente que, ao indexar a unidade “queda”, a hierarquia de hiperônimos e hipônimos obtida engloba a hierarquia herdada por meio da indexação prévia de “acidente” (Figura 12a). A partir dos conceitos mais específicos, o primeiro conceito/*synset* em comum nas hierarquias parciais é tomado como ponto de contato para a unificação das árvores conceituais. No caso, a unificação dessas árvores é feita por meio do conceito/*synset* {crash, wreck}. Como consequência dessa unificação, as frequências dos *synsets* constitutivos de cada hierarquia parcial também são unificadas

Na Figura 13, ilustra-se a hierarquia resultante da unificação das hierarquias parciais da Figura 12a-b e de suas respectivas frequências. Nessa Figura, vê-se que, com exceção de {crash, wreck}, os demais hipônimos de {accident} ficam com suas frequências originais (5), o mesmo acontece com {prang}, hipônimo de {crash, wreck}, que permanece com a frequência 1. Após a unificação das 38 hierarquias parciais, obter-se-á uma única hierarquia em os conceitos estarão devidamente pontuados em função de sua ativação total na ontologia.

Seguindo a metodologia proposta para este projeto, aplicou-se, à hierarquia unificada, estratégias que possibilitassem a delimitação do ramo da árvore conceitual que engloba os conceitos mais representativos do *cluster* C1. Em outras palavras, buscou-se delinear conceitualmente o *cluster* C1 por meio da delimitação de uma subontologia.

Para ilustração, consideram-se a hierarquia unificada da Figura 13 e a frequência de ativação dos conceitos/*synsets* como critério para a delimitação conceitual. Dessa forma, o ramo que delineia conceitualmente o *cluster* C1 é o da Figura 14. No entanto, esse processo não é simples como a ilustração pode sugerir, pois há diversos ramos da ontologia ativados com frequências iguais ou muito próximas. Assim, foi preciso investigar os critérios de delimitação de subontologia comumente utilizados na literatura para que, por meio deles, o recorte da subontologia mais adequada pudesse ser realizado.

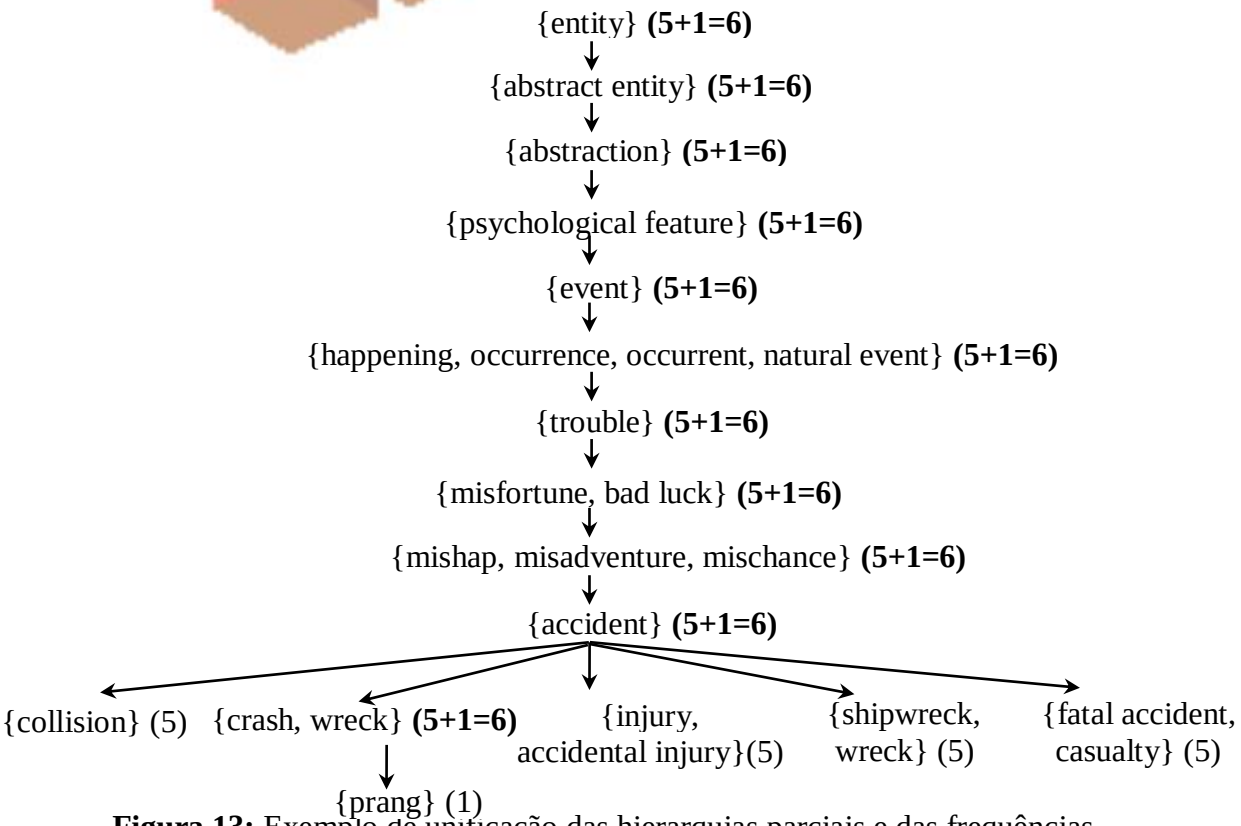

Figura 13: Exemplo de unificação das hierarquias parciais e das frequências.

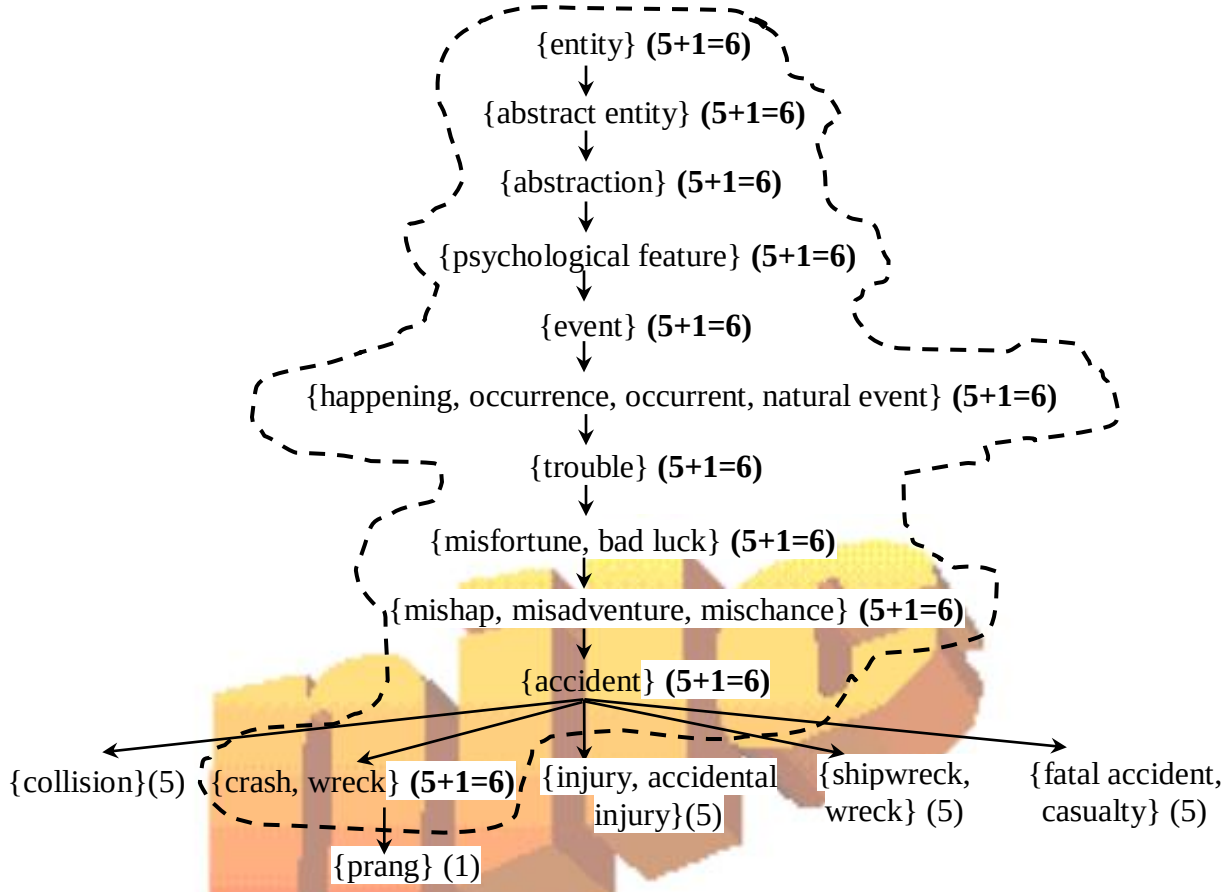

Figura 14: Ilustração do processo de delimitação de subontologias.

As hierarquias parciais herdadas da WN.Pr em função da indexação de cada uma das 38 unidades lexicais foram unificadas por meio da ferramenta *Cmap Tools*[16], que forneceu o ferramental para a representação gráfica (em formato de árvore conceitual) da hierarquia final.

4. Elaboração de regras heurísticas de delimitação de subontologias

Diante da hierarquia final, passou-se à fase de investigação, também manual, de estratégias de poda que permitissem delimitar a região da hierarquia ou árvore conceitual que englobasse os conceitos mais representativos da coleção C1.

Para tanto, optou-se pela estratégia geral de poda no sentido *top-down*, com base na qual se partiu dos conceitos (*synsets*) mais genéricos (hiperônimos) em direção aos conceitos mais específicos (hipônimos).

Especificamente para a exclusão dos conceitos, tomou-se como ponto de partida o trabalho de Raimer e Hah (1988, *apud* MANI, 2001). Nele, os autores apresentam o delineamento conceitual de um único texto em alemão por meio da indexação de suas unidades lexicais a uma ontologia do domínio da "computação" utilizando o sumarizador TOPIC. Para excluir ou não um conceito *x* da ontologia, os autores utilizaram os seguintes critérios: (i) frequência relativa do conceito *x* na ontologia e (ii) número de hipônimos de *x* que efetivamente ocorreram no texto em relação ao número total de hipônimos de *x*.

[16] http://ftp.ihmc.us/

Nesta primeira investigação, optou-se pelo critério (i), ou seja, pelo peso do conceito, especificado pela frequência.

Tendo em vista a estratégia *top-down*, iniciou-se a poda pelos níveis mais genéricos. ou seja, pelos hiperônimos que estavam no topo da hierarquia e que representavam os conceitos mais gerais do domínio. A hierarquia unificada foi montada e manipulada a partir de níveis, sendo que, no primeiro nível da hierarquia, tem-se o *synset* {entity} e, no segundo nível, os *synsets* co-hipônimos {physical entity} e {abstract entity} que não foram submetidos ao procedimento de poda, pois capturam os tipos básicos de conceitos. No entanto, a hierarquia foi subdividida em {physical entity} e {abstract entity} e, assim, a partir do terceiro nível, realizou-se a poda dos conceitos que consistiu em:

(i) identificar a média das frequências dos conceitos do nível, ou seja, para cada subontologia, a soma dos co-hipônimos dividida pelas soma de suas respectivas frequências consistiu em uma média;
(ii) identificar uma porcentagem da média – para cada média foram calculadas porcentagens de 30%, 40%, 50%, 60% e 70%;
(iii) podar os conceitos/*synsets* que apresentavam frequência menor que o valor obtido em (ii). A partir do teste das 5 porcentagens gerou-se 5 subontologias distintas a partir da hierarquia unificada.

Nas Tabelas 1 e 2 constam os cálculos realizados com seus respectivos resultados e as porcentagens utilizadas para a poda dos conceitos.

Abstract entity								
	Co-hipônimos	Frequência	Média	30%	40%	50%	60%	70%
Level 10	41	6	6,8	2,1	2,7	3,4	4,1	4,8
Level 09	41	9	4,6	1,4	1,8	2,3	2,7	3,2
Level 08	185	60	3,1	0,9	1,2	1,5	1,9	2,2
Level 07	38	14	2,7	0,8	1,1	1,4	1,6	1,9
Level 06	45	16	2,8	0,8	1,1	1,4	1,7	2,0
Level 05	68	29	2,3	0,7	0,9	1,2	1,4	1,6
Level 04	21	8	2,6	0,8	1,1	1,3	1,6	1,8
Level 03	36	9	4,0	1,2	1,6	2,0	2,4	2,8
Level 02	6	6	1,0	0,3	0,4	0,5	0,6	0,7

Tabela 1: Cálculo de média e porcentagens aplicadas para poda da subontologia {abstract entity}

Physical entity								
	Co-hipônimos	Frequência	Média	30%	40%	50%	60%	70%
Level 11	61	4	15,3	4,6	6,1	7,6	9,2	10,7
Level 10	58	7	8,3	2,5	3,3	4,1	5,0	5,8
Level 09	58	7	8,3	2,5	3,3	4,1	5,0	5,8
Level 08	194	142	1,4	0,4	0,5	0,7	0,8	1,0
Level 07	58	19	3,1	0,9	1,2	1,5	1,8	2,1
Level 06	605	89	6,8	2,0	2,7	3,4	4,1	4,8
Level 05	141	42	3,4	1,0	1,3	1,7	2,0	2,4
Level 04	113	30	3,8	1,1	1,5	1,9	2,3	2,6
Level 03	16	5	3,2	1,0	1,3	1,6	1,9	2,2
Level 02	11	1	11,0	3,3	4,4	5,5	6,6	7,7
Level 01	121	12	10,1	3,0	4,0	5,0	6,1	7,1

Tabela 2: Cálculo de média e porcentagens aplicadas para poda da subontologia {physical entity}

Na Figura 15, exemplifica-se o procedimento de poda no nível 4 da hierarquia dos conceitos do tipo {abstract entity}. Nessa Figura, os conceitos/*synsets* estão seguidos por suas respectivas frequências. Seguindo a metodologia, calculou-se a média das frequências dos conceitos nesse nível, no caso, 41/6=6,8.

Especificando-se, por exemplo, 30% da média, obtém-se o valor de 2,1. Dessa forma, todos os conceitos/*synsets* do nível 4 que apresentam frequência igual ou inferior a 2,05 (~2) foram podados, juntamente com seus respectivos hipônimos. No caso, os conceitos podados foram {communication} e {atribute), já que ambos possuem frequência 2. Especificando-se o valor de 70% da média, ou seja, 4,8, podam-se {communication}, {atribute} e também {relation}, pois este último possui frequência 4. Dessa forma, vê-se que quanto maior a porcentagem da média, mais conceitos são podados.

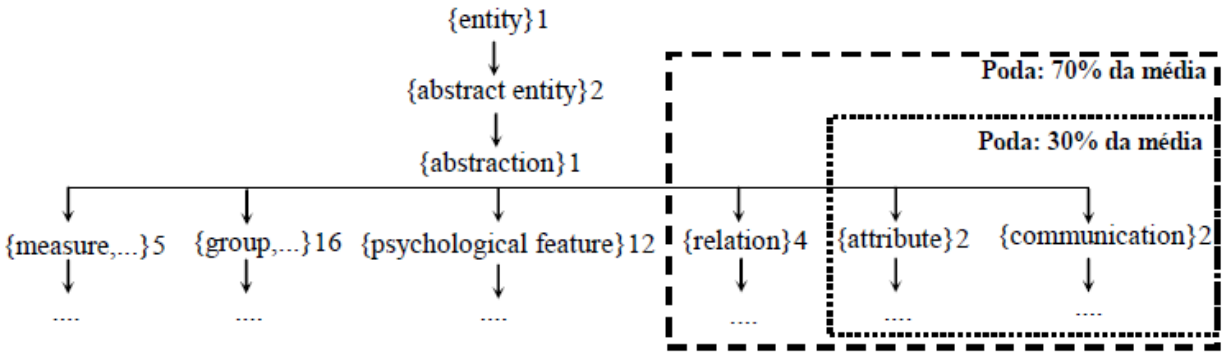

Figura 15: Exemplo de poda com base em diferentes critérios estatísticos

5. Avaliação das subontologias

Para avaliar a pertinência das subontologias geradas pela aplicação das diferentes porcentagens quanto ao delineamento conceitual da C1, verificou-se manualmente qual a menor das subontologias englobava os principais conceitos nominais presentes no sumário humano multidocumento da coleção C1, posto que este, por ser informativo, veicula idealmente os conceitos mais representativos de sua respectiva coleção.

Na Tabela 2, apresenta-se o resultado dessa verificação. Nela, as palavras anotadas com "x" foram excluídas da subontologia gerada em função das porcentagens de poda. No caso, a Tabela 2 evidencia que a subontologia gerada pela poda baseada em 50% da média (da frequência) é a menor, dentre todas, que engloba os conceitos mais frequentes da coleção C1. Diz-se isso porque as subontologias menores, geradas pelas podas baseadas em 60% e 70% da frequência dos conceitos, excluem o substantivos/conceito "membro", que está entre os mais frequentes da coleção.

Nome/ sumário	Frequência em C1	Nome podado				
		30%	40%	50%	60%	70%
avião	11					
passageiro	5					
membro	4				x	x
tripulação	4					
floresta	3					x
pessoa	3					x
carga	2					x
mineral	2	x	x	x	x	x
nacionalidade	2	x	x	x	x	x
tempo	2		x	x	x	x
vítima	2		x	x	x	x
montanha	1	x	x	x	x	x
Queda	1					

Tabela 2: Resultados da avaliação das estratégias de poda.

Assim, acredita-se que uma medida estatística pertinente para a delimitação da região da ontologia que engloba os conceitos mais representativos da coleção esteja em torno de 50% da média da frequência dos conceitos.

6. Considerações finais

Diante das atividades realizadas, os resultados obtidos do projeto são:

(i) Aquisição de um arcabouço teórico-metodológico sobre o processo de indexação léxico-conceitual no cenário da sumarização automática.

(ii) Indexação de uma coleção do *corpus* CSTNews à WN.Pr, no caso, do *cluster* C1.

(iii) Geração de regras heurísticas para a automatização do método(s)/ estratégia(s) especificada(s) em (b);

(iv) Aquisição de uma subontologia da WN.Pr para a coleção ou *cluster* C1 que compõe o *corpus* CSTNews.

Tendo em vista os objetivos iniciais do projeto, ou seja, identficar e aplicar as estratégias mais difundidas na literatura para a delimitação de subontologias a partir das indexações de unidades lexicais provenientes de *corpora*, enfatiza-se que os mesmos foram alcançados. Como trabalho futuro, pretende-se indexar outras coleções do CSTNews à WN.Pr para verificar se as estatísticas se confirmam pertinentes para o delineamento conceitual.

7. Agradecimentos

A aluna agradece à Coordenadoria de Iniciação Científica e Tecnológica da Pró-Reitoria de Pesquisa da UFSCar pela bolsa concedida no âmbito do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação – PIBITI/CNPq/UFSCar e aos pesquisadores do grupo de pesquisa NILC (Núcleo Interinstitucional de Linguística Computacional) pela participação ativa no projeto.

9. Referências bibliográficas

ANTONUIOU, G.; HARMELEN, F van. Web ontology language: OWL. In: STAAB, S., STUDER, R. (Eds.). **Handbook on ontologies.** International Handbooks on Information Systems. Berlin, Heidelberg: Springer-Verlag, 2004, p.67-92.

ALEIXO, P.; PARDO, T.A.S. CSTNews: um *corpus* de textos jornalísticos anotados segundo a Teoria Discursiva Multidocumento CST (*Cross-document Structure Theory*). Série de Relatórios Técnicos do ICMC, São Carlos-SP, n. 326, 12p., 2008.

BRANCO, A.; SILVA, J. Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4, 2004, Paris. **Proceeginds…** Paris, France, 507–510, 2004.

CARDOSO, P.C.F.; MAZIERO, E.G.; JORGE, M.L.C.; SENO, E.M.R.; DI FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. CSTNews: a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings…**Cuiabá/MT, Brazil, p. 88-105, 2011.

CRUSE, A. **Meaning in language: an introduction to semantics and pragmatics.** Oxford: Oxford University Press, 2004.

___. **A glossary of semantics and pragmatics**. United Kingdom: Edinburgh University Press, 2006.

DIAS-DA-SILVA, B.C. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, 2006.

DIAS-DA-SILVA, B.C.; DI FELIPPO, A.; NUNES, M.G.V. The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 6, 2008, Marrocos. **Proceedings…**Marrocos, Marrakech, p. 335-342, 2008.

DI-FELIPPO, A. The TermiNet Project: an Overview. In: NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, 1, 2010, Los Angeles, California. **Proceedings…**Los Angeles, p. 92–99, 2010.

FARRAR, S.; BATEMAN, J. Linguistic ontology baseline. **OntoSpace Internal Report I1-[OntoSpace]: D3. SFB/TR8**. Bremen: Collaborative Research Center for Spatial Cognition, 2005.

FELLBAUM, C (Ed.). **Wordnet**: an electronic lexical database. Ca, MA: MIT Press, 1998.

GUARINO, N.; GIARETTA, P. **Ontologies and Knowledge Bases**. Towards Terminological Clarification. http://www.loa.istc.cnr.it/Papers/KBKS95.pdf, 1995.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. **International Journal Human-Computer Studies**, v. 43, n. 5-6, p. 907-928, 1995.

HENNIG, L., UMBRATH, W., WETZKER, R. An ontology-based approach to Text Summarization. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING AND ONTOLOGY ENGINEERING (NLPOE 2008), 3, Toronto, 2008. **Proceedings…**Toronto, Canada, p. 291-294, 2008.

HOVY, E.H.; LIN, C.H. Automated Text Summarization in SUMMARIST. In: Maybury and Mani, I. (Eds.). **Advances in Automatic Text Summarization**. Cambridge: MIT Press, p. 2-14, 1999.

JORGE, M.L.C.; PARDO, T.A.S. Experiments with CST-based Multidocument Summarization. In: ACL WORKSHOP TEXTGRAPHS, 5, 2010, Uppsala, Sweden. **Proceedings…** Uppsala, p. 74-82, 2010.

LENAT, D. GUHA, R. **Building large knowledge based systems: representation and inference in the Cyc project.** Addison-Wesley Publishing, 1990.

LI, L., WANG, D., SHEN, C., LI, T. Ontology-enriched Multi-Document Summarization in disaster management. In: ACM Special Interest Group on Information Retrieval (SIGIR), Geneva, 2010. **Proceedings**… Geneva, Switzerland, p. 819-820, 2010.

MAGNINI, B., SPERANZA, M. Merging global and specialized linguistic ontologies. In: LREC, 3, 2002, Las Palmas. **Proceedings**… Las Palmas: University of Las Palmas, 2002.

MANI, I. **Automatic Summarization.** Amsterdam: John Benjamins Publishing Co., 2001.

___; MAYBURY, M.T. **Advances in automatic text summarization**. Cambridge, MA.: The MIT Press, 1999.

MARTINS, C. B. *et al.* Introdução à Sumarização Automática. **Rel. Técnico RT-DC 002/2001**, Departamento de Computação, UFSCar, São Carlos. Abril, 2001. 38p.

MCKEOWN, K.. RADEV, D.R. Generating summaries of multiple news articles. In: INTERNATIONAL ACM-SIGIR, 18, 1995, Seattle. **Proceedings…**Seattle, 1995, p. 74-82.

MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004.

NIRENBURG, S.;RASKIN, V.**Ontological semantics.** Cambridge, MA: The MIT Press, 2004.

PALMER, M. Multilingual resources, multilingual information management: current levels and future abilities. **Linguistica Computazionale**, v. XIV-XV, p. 1-33, 2001.

PARDO, T.A.S. GistSumm - GIST SUMMarizer: extensões e novas funcionalidades. **Série de Relatórios do NILC.** NILC-TR-05-05. São Carlos-SP, 8p., 2005.

SPARCK JONES, K. Discourse modeling for Automatic Summarisation. **Tech. Report No. 290**. University of Cambridge. UK, February, 1993.

SINCLAIR, J. Corpus and text: basic principles. In: Wynne, M. (Ed.). **Developing linguistic corpora: a guide to good practice.** Oxford: Oxbow Books, 2005. p.1-16. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em: 30 out. 2006.

STUCKENSCHMIDT, H,; PARENT, C.; SPACCAPIETRA, S. **Modular ontologies: concepts, theories and techniques for knowledge modularization.** (Lecture Notes in Computer Science / Theoretical Computer Science and General Issues, Vol. 5445). Springer-Verlag: Berlin, Heidelberg, 2009, 388p.

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. A comprehensive comparative evaluation of RST-based summarization methods. **ACM Transactions on Speech and Language Processing**, 6(4), 2010, p. 1-20.

VOSSEN, P. Introduction to EuroWordNet. **Computers and the Humanities**, Dordrecht: Kluwer Academic Publichers, v. 32, p. 73-89, 1998.

WEISZFLOG, W. **Michaelis: moderno dicionário inglês** (inglês-português/ português-inglês). Editora Melhoramentos, 2000. Disponível em <http://michaelis.uol.com.br/moderno/ingles/index.php>.

WU, C. W; LIU, C. L. Ontology-based text summarization for business news articles. In: ISCA INTERNATIONAL CONFERENCE COMPUTERS AND THEIR APPLICATIONS, 18, 2003, Honolulu. **Proceedings...** Honolulu, Hawaii, USA, 2003, p. 389–392, 2003.