Programa de Pós-Graduação em Lingüística

PPGL

INVESTIGAÇÃO DE ESTRATÉGIAS DE SELEÇÃO DE CONTEÚDO BASEADAS NA UNL (*UNIVERSAL NETWORKING LANGUAGE*)

MATHEUS RIGOBELO CHAUD

SÃO CARLOS
2015

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

INVESTIGAÇÃO DE ESTRATÉGIAS DE SELEÇÃO DE CONTEÚDO BASEADAS NA UNL (*UNIVERSAL NETWORKING LANGUAGE*)

MATHEUS RIGOBELO CHAUD

Bolsista: CAPES

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística.

Orientadora: Profa. Dra. Ariani Di Felippo

São Carlos – São Paulo – Brasil

2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C496ie

Chaud, Matheus Rigobelo.
Investigação de estratégias de seleção de conteúdo baseadas na UNL (*Universal Networking Language*) / Matheus Rigobelo Chaud. -- São Carlos : UFSCar, 2015.
169 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2015.

1. Linguística aplicada. 2. Sumarização automática. 3. Estratégias de seleção de conteúdo. 4. Interlíngua UNL (*Universal Networking Language*). I. Título.

CDD: 418 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Matheus Rigobelo Chaud, realizada em 03/03/2015:

Profa. Dra. Ariani Di Felippo
UFSCar

Prof. Dr. Thiago Alexandre Salgueiro Pardo
USP

Profa. Dra. Maria das Graças Volpe Nunes
USP

RESUMO

Na área de Processamento Automático das Línguas Naturais (PLN), há um destaque crescente para a Sumarização Automática Multidocumento Multilíngue (SAMM), cujo objetivo é processar uma coleção de documentos-fonte em mais de uma língua e gerar um sumário correspondente a essa coleção em uma das línguas-alvo. Na SAMM, a seleção das sentenças dos textos-fonte para composição do sumário pode ser feita com base em atributos linguísticos superficiais ou profundos. O objetivo deste projeto foi investigar se a utilização de conhecimento profundo, obtido a partir de uma representação conceitual dos textos-fonte, pode ser útil na seleção de conteúdo em textos do gênero jornalístico. Para isso, utilizou-se um sistema de representação formal – a UNL (*Universal Networking Language)*. Visando investigar estratégias de seleção de conteúdo baseadas nessa interlíngua, fez-se a representação em UNL de 3 coleções de textos, cada qual com 1 texto-fonte em português, 1 texto-fonte em inglês e 1 sumário humano de referência. Fez-se também o alinhamento das sentenças dos textos-fonte de cada coleção às sentenças de seus respectivos sumários humanos, objetivando identificar sobreposição total ou parcial de conteúdo entre essas sentenças. Esses dados permitiram a comparação entre estratégias de seleção de conteúdo baseadas em informações conceituais e um método de seleção tradicional baseado em um atributo superficial – a posição da sentença no texto-fonte. De acordo com os resultados obtidos, a seleção de conteúdo com base na posição no texto-fonte correlacionou-se mais adequadamente com a seleção realizada pelo sumarizador humano, comparado aos métodos conceituais investigados. Além disso, as sentenças iniciais dos textos-fonte, que, em textos jornalísticos, normalmente veiculam as informações mais relevantes, não necessariamente continham os conceitos mais frequentes da coleção; em diversas ocasiões, as sentenças com os conceitos mais frequentes estavam em posição intermediária ou final no texto. Esses resultados indicam que, ao menos nas coleções analisadas, outros critérios, além da frequência de conceitos, concorrem para determinar a relevância de uma sentença. Em outras palavras, na sumarização humana multidocumento, a seleção de conteúdo provavelmente não se resume a selecionar sentenças com os conceitos mais frequentes, tratando-se de um processo bem mais complexo.

Palavras-chave: Sumarização Automática Multidocumento Multilíngue. Processamento Automático de Línguas Naturais. Sistemas de Representação de Conhecimento. *Universal Networking Language*. UNL. Seleção de conteúdo.

ABSTRACT

The field of Natural Language Processing (NLP) has witnessed increased attention to Multilingual Multidocument Summarization (MMS), whose goal is to process a cluster of source documents in more than one language and generate a summary of this collection in one of the target languages. In MMS, the selection of sentences from source texts for summary generation may be based on either shallow or deep linguistic features. The purpose of this research was to investigate whether the use of deep knowledge, obtained from a conceptual representation of the source texts, could be useful for content selection in texts within the newspaper genre. In this study, we used a formal representation system – the UNL (Universal Networking Language). In order to investigate content selection strategies based on this interlingua, 3 clusters of texts were represented in UNL, each consisting of 1 text in Portuguese, 1 text in English and 1 human-written reference summary. Additionally, in each cluster, the sentences of the source texts were aligned to the sentences of their respective human summaries, in order to identify total or partial content overlap between these sentences. The data collected allowed a comparison between content selection strategies based on conceptual information and a traditional selection method based on a superficial feature - the position of the sentence in the source text. According to the results, content selection based on sentence position was more closely correlated with the selection made by the human summarizer, compared to the conceptual methods investigated. Furthermore, the sentences in the beginning of the source texts, which, in newspaper articles, usually convey the most relevant information, did not necessarily contain the most frequent concepts in the text collection; on several occasions, the sentences with the most frequent concepts were in the middle or at the end of the text. These results indicate that, at least in the clusters analyzed, other criteria besides concept frequency help determine the relevance of a sentence. In other words, content selection in human multidocument summarization may not be limited to the selection of the sentences with the most frequent concepts. In fact, it seems to be a much more complex process.

Keywords: Automatic Summarization. Multilingual Multidocument Summarization. Natural Language Processing. Knowledge Representation Systems. Universal Networking Language. UNL. Content selection.

RESUMO (ESPERANTO)

En la fako nomita Procezado de Naturaj Lingvoj (PNL) oni vidas pliigantan atenton al Plurlingva Plurdokumenta Aŭtomata Resumado (PPAR), kies celo estas prilabori aron da fontaj dokumentoj en pli ol unu lingvo kaj krei resumon de tiu kolekto en unu el la celaj lingvoj. En PPAR, la elekto de frazoj el fontaj tekstoj por generi resumon povas baziĝi sur neprofundaj aŭ profundaj lingvaj karakterizoj. Ĉi tiu esploro celis kontroli, ĉu la uzado de profunda scio, per koncepta reprezento de la fontaj tekstoj, povus esti utila por elekti taŭgan enhavon en tekstoj el la ĵurnala ĝenro. En ĉi tiu esploro, formala sistemo de reprezentado estis uzata – la Universala Reta Lingvo, UNL (*Universal Networking Language*). La ĉefa celo estis identigi strategiojn por elekto de enhavo surbaze de ĉi tiu interlingvo. Tri aroj da tekstoj estis reprezentitaj laŭ UNL. Ĉiu tekstaro konsistis el 1 teksto en la portugala lingvo, 1 teksto en la angla kaj 1 homfarita referenca resumo. Cetere, en ĉiu tekstaro, la frazoj el la fontaj tekstoj estis ligitaj al la frazoj el sia respektiva homfarita resumo, celante identigi parte aŭ tute kongruan enhavon inter la fontaj tekstoj kaj la resumo. La datumoj kolektitaj permesis komparon inter strategioj por elekto de enhavo bazitaj sur konceptaj informoj kaj unu strategio bazita sur malprofunda karakterizo – la pozicio de la frazo en la fonta teksto. Laŭ la rezultoj, elekto de enhavo baze de la pozicio de la frazo estis pli konforma al la elekto homfarita, kompare kun la konceptaj metodoj esploritaj. Krome, la frazoj en la komenco de la fontaj tekstoj, kiuj kutime enhavas la plej gravajn informojn en gazetartikoloj, ne nepre entenis la plej oftajn konceptojn en la kolekto de tekstoj; en pluraj okazoj, la frazoj kun la plej oftaj konceptoj estis en la mezo aŭ en la fino de la teksto. Tiuj rezultoj indikas ke, almenaŭ en la tekstaroj analizitaj, aliaj kriterioj krom la ofteco de konceptoj helpas determini la gravecon de frazoj. Alivorte, la elekto de enhavo en homa plurdokumenta resumado ne konsistas nur en elekto de frazoj kun la plej oftaj konceptoj. Ŝajne, ĝi estas multe pli kompleksa procezo.

Ŝlosilvortoj: Plurlingva Plurdokumenta Aŭtomata Resumado. Procezado de Naturaj Lingvoj. Sistemoj de Reprezentado de Scio. Universala Reta Lingvo. UNL. Elekto de enhavo.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE SIGLAS

AL – *Attribute Label* (Etiqueta de Atributo)
C (em *C1*, por exemplo) – *Cluster* ou coleção
CLEA – *Certificate of Language Engineering Aptitude*
CM2News – *Corpus* Multidocumento Bilíngue de Textos Jornalísticos
CM3News – *Corpus* Multidocumento Trilíngue de Textos Jornalísticos
CUP – *Certificate of Proficiency in UNL*
D (em *D1* e *D2*) – Documento
DUC – *Document Understanding Conference*
EN – Inglês
EOLSS – *Encyclopedia of Life Support System*
F(UWs) – Frequência das UWs
IDF – *Inverted Document Frequency* (Frequência Inversa nos Documentos)
NILC – Núcleo Interinstitucional de Linguística Computacional
PB – Português brasileiro
PLN – Processamento Automático de Línguas Naturais
RC – Representação de Conhecimento
RL – *Relation Label* (etiqueta de relação ou relação binária)
ROUGE – *Recall-Oriented Understudy of Gisting Evaluation*
SA – Sumarização Automática
SAM – Sumarização Automática Multidocumento
SAMM – Sumarização Automática Multidocumento Multilíngue
SD – Sentença do documento
SHMM – Sumarização Humana Multidocumento Multilíngue
SS – Sentença do sumário
Sum_ref – Sumário de referência
SUCINTO – *Summarization for Clever Information Access*
SUSTENTO – *Generation of Linguistic Knowledge for Multi-document Summarization*
TA – Tradução automática
TF – Frequência de ocorrência de um conceito
UFSCar – Universidade Federal de São Carlos
UNDL – *The Universal Networking Digital Language Foundation*
UNESCO – Organização das Nações Unidas para a Educação, a Ciência e a Cultura
UNL – *Universal Networking Language*
UW – *Universal Word* (palavra universal ou conceito)
WN.Pr – WordNet de Princeton

SUMÁRIO

Ao meu filho, Lucas, por mudar a forma como vejo o mundo,
me ensinando, a cada dia, o que é o amor, em faces que eu desconhecia.

AGRADECIMENTOS

Primeiramente a Deus, por sempre nos permitir novos começos.

À minha mãe, pelos exemplos de amor e dedicação; especialmente, por ter se esforçado tanto em nome de uma instituição abençoada – a família, preciosa fonte de aprendizado.

Ao meu pai, por ter me ensinado o valor do respeito e do trabalho, mostrando-me, pelo exemplo, que vale a pena seguir o caminho da honestidade.

À minha esposa, companheira, alma querida, que tanto me ensina e a quem agradeço a Deus todos os dias por ter ao meu lado, Viviane.

Aos amigos e familiares, pelo apoio e carinho que sempre manifestaram.

A todos os companheiros nessa breve jornada de Mestrado, tanto os na posição de aprendizes como de instrutores. Um agradecimento especial à minha orientadora, Profa. Ariani Di Felippo, por acreditar em um aluno vindo de outra área de conhecimento e por ter auxiliado, de maneira imprescindível, a dar forma e direcionamento a este projeto. Agradeço também ao Prof. Thiago Alexandre Salgueiro Pardo, cujos conselhos foram de imenso valor, com intervenções precisas, sábias e gentis. Meus agradecimentos à Profa. Lucia Helena Machado Rino pelas sugestões e, especialmente, por ter aceitado participar de minha Banca de Qualificação mesmo após sua aposentadoria, com correções de grande valia. Agradeço ainda à Profa. Vera Lucia Teixeira da Silva por todo o incentivo e pela maneira extremamente agradável com que conduziu duas disciplinas que tive a felicidade de cursar; as reflexões suscitadas nessas ocasiões mudaram minha forma de enxergar o processo de ensino-aprendizagem e fizeram uma diferença imensurável em minha prática pedagógica.

Por fim, agradeço à CAPES pelo apoio financeiro concedido sob a forma de bolsa de mestrado.

1 INTRODUÇÃO

1.1 Contextualização

Na *Sumarização Automática Multidocumento Multilíngue* (SAMM), busca-se de um modo geral processar uma coleção (em inglês, *cluster*) de textos-fonte sobre um mesmo assunto em ao menos duas línguas (Lx e Ly) e produzir o sumário correspondente a essa coleção em uma das línguas-fonte (Lx ou Ly) (p.ex.: EVANS; McKEOWN; KLAVANS, 2005; ROARK; FISHER, 2005; ORĂSAN; CHIOREAN, 2008; WAN; LI; XIAO, 2010).

Dada a grande quantidade de informação textual disponível em várias línguas na *web,* as pesquisas sobre SAMM vêm ganhando destaque nos últimos anos. Essa centralidade no cenário do Processamento Automático das Línguas Naturais (PLN) se deve exatamente ao fato de a SAMM permitir o acesso rápido, na língua do usuário, à informação relevante veiculada por uma coleção de textos sobre dado assunto publicados em sua própria língua e em uma língua distinta daquela do usuário. Por exemplo, para um falante do português brasileiro (PB), talvez seja interessante ter acesso, em PB, às informações mais importantes que circularam em jornais americanos e brasileiros sobre o furacão Sandy.

Os métodos de SAMM geralmente englobam um processo de tradução automática (TA). Ressalta-se que, apesar de ser uma alternativa para o multilinguismo, textos gerados por TA apresentam várias deficiências linguísticas, tais como agramaticalidade e ilegibilidade (EVANS; McKEOWN; KLAVANS, 2005; ORĂSAN; CHIOREAN, 2008). Por isso, os métodos de SAMM precisam englobar estratégias para tratar essas deficiências. Uma vez que a multiplicidade das línguas é tratada por meio da TA, a sumarização em si passa a ser exclusivamente multidocumento, já que a seleção do conteúdo para compor o sumário é feita a partir de uma coleção de textos em uma única língua (MANI, 2001).

Na SAMM, a seleção das sentenças dos textos-fonte para compor os sumários muitas vezes é feita com base em alguns atributos linguísticos superficiais (p.ex.: posição que a sentença ocupa no texto-fonte) e em um atributo profundo particular (quantidade de conceitos lexicalizados representativos da coleção presentes na sentença).

Nesse cenário, tem-se investigado a representação dos textos-fonte em nível conceitual. Estratégias de SAMM baseadas nesse tipo de representação devem possibilitar a identificação e seleção do conteúdo relevante, configurando-se como uma alternativa à TA direta dos textos-fonte.

Um exemplo dessas investigações são os métodos descritos em Tosta (2014), cujo objetivo é a produção de sumários extrativos[1] em português. Partindo de uma coleção contendo um texto em português e um em inglês, Tosta faz a indexação dos nomes a um único conjunto de conceitos; na sequência, pontua e classifica as sentenças dos textos-fonte com base na frequência de ocorrência desses conceitos na coleção. A partir do ranque obtido, um dos métodos consiste em selecionar para o sumário apenas as sentenças em português com pontuação mais alta, até que a taxa de compressão desejada seja atingida (ou seja, até que o sumário alcance o tamanho desejado). No segundo método, selecionam-se as sentenças mais bem pontuadas independentemente de sua língua-fonte e, no caso de sentenças em inglês serem selecionadas, faz-se a tradução automática dessas para o português.

Quando comparados a métodos mais tradicionais que realizam a TA dos textos-fonte, os métodos que utilizam conhecimento léxico-conceitual geram extratos genéricos mais informativos e com índices mais elevados de gramaticalidade, clareza referencial, não redundância, coerência e foco (TOSTA, 2014).

Frente aos resultados promissores no que diz respeito ao uso de conhecimento conceitual, pretendeu-se, neste projeto, investigar estratégias de seleção de conteúdo com base na representação dos textos-fonte em um modelo formal (ou formalismo), mais especificamente a UNL - *Universal Networking Language* (UCHIDA et al., 1999). Para isso, buscou-se identificar técnicas de seleção de conteúdo utilizadas por humanos na Sumarização Humana Multidocumento Multilíngue (SHMM) que se projetassem na representação do conteúdo subjacente às sentenças e pudessem ser formalizadas e aplicadas à representação conceitual dos textos-fonte. Tais estratégias poderiam, então, servir de subsídio para sistemas de SAMM baseados na UNL.

Por representar o conteúdo dos textos em nível conceitual, independente de língua, uma formalização condensada do conteúdo da coleção funciona como interlíngua, a partir da qual é possível gerar os sumários em diferentes línguas. Dessa forma, a SAMM pode ser realizada sem uma etapa de tradução automática direta entre as línguas, mas por meio dessa interlíngua.

A utilização de interlíngua na Sumarização Automática (SA) já foi investigada nos cenários monodocumento monolíngue (isto é, processo no qual se produz um sumário em uma língua Lx a partir de um único texto-fonte em Lx) (p.ex.: MARTINS, C. B., 2002; MARTINS, C. B.; RINO, 2002a) e monodocumento multilíngue (isto é, processo no qual se

[1] Sumários compostos por trechos (p.ex.: sentenças) integralmente extraídos dos textos-fonte.

produz um sumário em uma língua Ly a partir de um único texto-fonte em Lx) (p.ex.: (LENCI et al., 2002). Até onde sabemos, no entanto, o uso de interlíngua na SAMM envolvendo o PB não havia sido investigado.

1.2 Objetivos

Diante do cenário apresentado, o objetivo deste projeto foi investigar a SAMM a partir da representação do significado subjacente às sentenças dos textos-fonte de dada coleção C, ou seja, a partir de conhecimento profundo no nível conceitual. Especificamente, objetivou-se elaborar estratégias para a seleção da informação relevante a compor o sumário que se apliquem às representações conceituais das sentenças dos textos de C.

Por meio do alinhamento das sentenças dos textos-fonte às sentenças de sumários produzidos manualmente e também pelo alinhamento de suas respectivas representações conceituais, buscou-se identificar técnicas de seleção de conteúdo recorrentes empregadas pelos sumarizadores humanos. O intuito foi identificar estratégias que se projetassem na representação conceitual e que pudessem ser formalizadas e disponibilizadas para a SAMM baseada em interlíngua.

Esses objetivos se inserem em um projeto maior denominado SUSTENTO[2] (FAPESP 2012/13246-5 / CNPq 483231/2012-6), cujo intuito é a geração de conhecimento linguístico que venha a fornecer subsídios para aprimorar os métodos existentes de sumarização e/ou propor novos métodos, especialmente os voltados para o processamento do português.

Os resultados obtidos ao longo do projeto SUSTENTO, incluindo os desta pesquisa, poderão ser utilizados ainda em outro projeto denominado SUCINTO[3] (FAPESP 2012/03071-3), que tem como objetivo criar recursos, ferramentas e sistemas de SA não só como contribuição científica, mas também visando disponibilizá-los para pesquisadores e usuários finais. Ambos os projetos vêm sendo desenvolvidos no NILC[4].

---

[2] Mais informações sobre o projeto SUSTENTO (*Generation of Linguistic Knowledge for Multi-document Summarization*) podem ser obtidas em sua página eletrônica: http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23.

[3] A página eletrônica do projeto SUCINTO (*Summarization for Clever Information Access*) está disponível em: http://www.icmc.usp.br/~taspardo/sucinto/.

[4] Maiores informações sobre o NILC (Núcleo Interinstitucional de Linguística Computacional) encontram-se disponíveis em: http://www.nilc.icmc.usp.br/nilc/.

1.3 Hipóteses

Para esta investigação, três hipóteses a respeito da seleção de conteúdo baseada em representações conceituais foram formuladas:

- **Hipótese 1:** a análise das representações conceituais das sentenças dos textos-fonte e das representações das sentenças de um sumário manual permite a elaboração de estratégias de seleção de conteúdo formalizáveis e aplicáveis às representações dos textos-fonte de uma coleção;
- **Hipótese 2:** o alinhamento das sentenças dos textos-fonte às sentenças de sumários humanos permite comparar as estratégias de seleção de conteúdo baseadas em representações conceituais e avaliar se elas se correlacionam com as estratégias utilizadas por humanos;
- **Hipótese 3:** estratégias de seleção de conteúdo baseadas em representações conceituais correlacionam-se mais adequadamente com a sumarização humana do que uma estratégia de seleção de conteúdo baseada em conhecimento superficial.

1.4 Metodologia

Para atingir os objetivos apresentados, foram realizadas as seguintes tarefas:

(a) Revisão da literatura (Tarefa 1): consistiu na leitura da bibliografia fundamental e demais referências pertinentes ao projeto que surgiram no decorrer da pesquisa. A bibliografia foi composta basicamente por trabalhos sobre SAMM e modelos de representação de conhecimento que pudessem funcionar como interlínguas no cenário da SAMM.

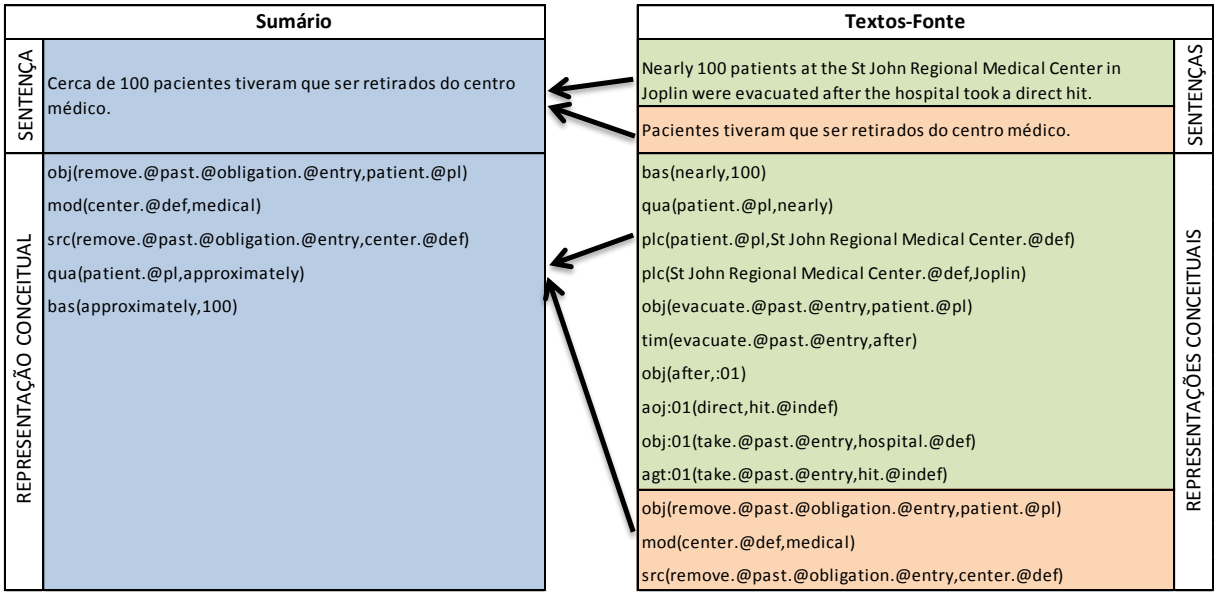
(b) Seleção do *corpus* (Tarefa 2): consistiu em selecionar um *corpus* adequado para servir de base à investigação em questão. No caso, necessitou-se de um *corpus* que fosse: (i) multidocumento, (ii) multilíngue e (iii) jornalístico. Deu-se predileção a um *corpus* bilíngue cujas coleções fossem compostas por (i) 1 texto em PB e 1 em inglês (EN) e (ii) 1 sumário multidocumento multilíngue humano ou manual em PB.

(c) Representação do *corpus* (Tarefa 3): consistiu basicamente na escolha do modelo de representação de conhecimento (ou formalismo) a partir da revisão da literatura (Tarefa 1) e

subsequente representação dos textos-fonte e também dos sumários manuais de acordo com o modelo escolhido. Essa representação foi feita manualmente, visto que não havia nenhuma ferramenta computacional automática disponível para realização da tarefa.

(d) Alinhamento dos textos-fonte aos sumários humanos (Tarefa 4): consistiu no alinhamento manual das sentenças dos textos-fonte de cada coleção às sentenças dos seus respectivos sumários humanos. Uma vez que essas sentenças foram representadas no formalismo escolhido durante a Tarefa 3, o alinhamento das sentenças permitiu o alinhamento indireto das representações conceituais. Assim, ao final da Tarefa 4, além do alinhamento sentencial, as representações conceituais das sentenças dos textos-fonte também foram alinhadas às representações conceituais das sentenças do sumário (Figura 1).

Figura 1 – Ilustração do processo de alinhamento



Fonte: Elaborado pelo autor

(e) Investigação de estratégias de seleção de conteúdo (Tarefa 5)

A Tarefa 5 consistiu em investigar estratégias de seleção de conteúdo com base na representação conceitual e no alinhamento realizados nas Tarefas 3 e 4. Objetivou-se identificar estratégias de seleção de conteúdo que pudessem ser formalizadas em regras aplicáveis às representações conceituais e que pudessem ser comparadas quanto ao seu potencial para selecionar o conteúdo considerado relevante pelos sumarizadores humanos.

1.5 Estrutura da dissertação

Esta dissertação organiza-se em 6 capítulos. Na Seção 2, apresenta-se a revisão da literatura, abrangendo noções sobre Sumarização Automática, sistemas de representação de conhecimento e interlínguas. Na Seção 3, relata-se a tarefa de seleção do *corpus* e sua representação de acordo com o sistema de representação de conhecimento investigado, ou seja, a UNL. A Seção 4 trata do alinhamento dos textos-fonte aos sumários humanos. Na Seção 5, descreve-se a investigação das estratégias de seleção de conteúdo, enquanto que nas Seções 6 e 7 apresentam-se a verificação das hipóteses e as considerações finais, respectivamente.

2 REVISÃO DA LITERATURA

2.1 Sumarização Automática

2.1.1 Noções básicas

O Processamento Automático de Línguas Naturais (PLN) geralmente é considerado uma subárea da Inteligência Artificial, relacionando-se diretamente com as áreas de Ciência da Computação e Linguística. O PLN tem como objetivo investigar como o computador pode ser utilizado para interpretar e gerar textos em língua natural com vistas a aplicações específicas, tais como corretores ortográficos, tradutores automáticos e sistemas de extração de informação, dentre outras (GRISHMAN, 1986; DIAS-DA-SILVA, 1996; NUNES; OLIVEIRA JR., 2000; DIAS-DA-SILVA et al., 2007; DI FELIPPO; DIAS-DA-SILVA, 2009; PARDO et al., 2010).

Um dos grandes desafios no PLN é a Sumarização Automática (SA), que atualmente figura entre os temas de pesquisa mais recorrentes nessa área. A SA tem como objetivo extrair o conteúdo relevante de uma fonte de informação, comumente textos, e apresentá-lo de uma maneira sucinta, levando-se em conta as necessidades do usuário. Geralmente, a SA envolve as etapas de (i) análise, (ii) transformação e (iii) síntese, ilustradas na Figura 2 (SPARCK JONES, 1998; MANI; MAYBURY, 1999; DIAS-DA-SILVA et al., 2007; PARDO, 2008; PANDIAN; KALPANA, 2013).

Figura 2 – Etapas da Sumarização Automática

Fonte: Adaptado de Sparck Jones (1998)

A análise geralmente visa interpretar os textos-fonte e extrair uma representação formal, ou seja, explícita e idealmente não ambígua dos mesmos.

A transformação é a etapa na qual, a partir da representação gerada na análise, condensa-se o conteúdo dos textos-fonte em uma representação interna do sumário. Essa etapa engloba a seleção do conteúdo que irá compor o sumário. Para tanto, os segmentos dos textos-fonte são comumente ranqueados em função de um critério de relevância e os de maior pontuação são selecionados, sendo que o critério de relevância depende da função e audiência do sumário.

A síntese visa à construção do sumário em língua natural a partir da representação interna gerada na transformação. Nessa fase, métodos de justaposição, ordenação, fusão e correferenciação dos segmentos selecionados podem ser utilizados. Essas etapas são guiadas, dentre outros fatores, pela taxa de compressão do sumário, definida por Mani (2001) como a razão entre o tamanho do sumário e o tamanho do texto-fonte.

Um sumário pode ser classificado como indicativo, informativo ou crítico. Nos sumários indicativos, não há preocupação em substituir os textos-fonte, mas apenas dizer do que eles tratam; tais sumários podem ser úteis como ponto de partida para a seleção de leituras mais aprofundadas que sejam de interesse do leitor. Portanto, o conteúdo de sumários indicativos não necessariamente conserva as informações mais relevantes dos documentos-fonte. Já os sumários informativos apresentam as informações mais importantes dos textos-fonte, podendo eliminar a necessidade da leitura dos mesmos. Os sumários críticos contêm informações dos textos-fonte e também uma avaliação sobre elas (MANI; MAYBURY, 1999; DIAS-DA-SILVA et al., 2007).

Quanto à audiência a que se destina, o sumário pode ser considerado genérico ou focado nos interesses dos usuários. Os sumários genéricos são produzidos sem vistas a um perfil específico de usuário. Já nos sumários focados nos interesses dos usuários, as informações são selecionadas em função de conhecimento prévio a respeito dos usuários ou com base em uma consulta (*query*, em inglês) por eles realizada. Exemplificando, caso o usuário não esteja familiarizado com o assunto de um texto-fonte, pode ser útil que o sumário desse texto traga também informações contextuais; por outro lado, caso o usuário seja um especialista no assunto, as informações contextuais deixam de ser relevantes, sendo mais adequado, por exemplo, agregar informações novas.

A sumarização pode ser feita com base em um único documento-fonte, sendo, nesse caso, chamada de sumarização monodocumento, ou com base em mais de um documento-fonte, denominada sumarização multidocumento (MANI, 2001).

Os sumários podem ainda ser divididos em extrativos e abstrativos. Na sumarização extrativa, todo o material que compõe o sumário é copiado do(s) texto(s)-fonte. Já na

sumarização abstrativa, pelo menos parte do material do sumário não consta no texto-fonte, sendo comum o uso de paráfrases e maior condensação de informações (MANI, 2001).

Apesar de trazer vantagens bastante significativas, como a facilidade de tratamento computacional e menor dependência de domínio, a sumarização extrativa apresenta também algumas desvantagens a serem consideradas. Pode ocorrer que sentenças sejam tiradas de seu contexto, prejudicando a coesão textual, ou ainda que sentenças contendo fragmentos pouco relevantes sejam levadas para o sumário, chamando a atenção para aspectos pontuais que normalmente não fariam parte do sumário (HATZIVASSILOGLOU et al., 2001; SOUZA et al., 2001).

Embora a sumarização extrativa por vezes funcione bem no cenário monodocumento, os resultados nem sempre são os ideais quando se trata de múltiplos documentos-fonte. Dificilmente algum dos documentos da coleção apresentará informações prontamente extraíveis que representem adequadamente o conteúdo de todos os documentos; por outro lado, se todas as sentenças com informações similares forem integradas ao sumário, a tendência é que haja repetição de informações e o sumário ultrapasse o tamanho desejável (McKEOWN et al., 1999).

Na sumarização multidocumento, especialmente, é importante que os sumarizadores consigam ir além da produção de extratos, para que eles sejam capazes de lidar eficientemente com as diferenças e as similaridades encontradas entre os documentos. Conforme observado por Mani (2001), a sumarização feita por humanos é abstrativa, e não extrativa: além de fragmentos dos textos originais, são feitas operações de reordenação, generalização e especialização de informação.

Com relação às línguas envolvidas no processo, pode-se falar em sumarização:

a) monolíngue: quando apenas uma língua é processada e o sumário é construído no mesmo idioma do(s) texto(s)-fonte;
b) multilíngue: quando parte-se de uma coleção de textos-fonte em mais de uma língua, sendo o sumário construído em uma dessas línguas;
c) *crosslanguage* (translíngue[5]): quando o sumário é produzido em um idioma diferente daquele(s) usado(s) como entrada em um ou mais textos-fonte (MANI, 2001).

[5] Embora não haja uma tradução corriqueiramente utilizada para o termo *crosslanguage* na literatura nacional, propõe-se aqui traduzi-lo como *translíngue*. O prefixo *trans* mostra-se adequado para traduzir satisfatoriamente esse movimento que é feito de uma língua em direção a outra. Reforça essa ideia a existência do termo *transexual*, por exemplo, que também dá a ideia de movimentação/mudança (no caso, de um sexo a outro).

A metodologia para desenvolvimento dos sumarizadores automáticos pode pautar-se no uso de informações superficiais ou profundas. Os sistemas baseados em conhecimento superficial fazem uso de pouco ou nenhum conhecimento linguístico, geralmente servindo-se de modelos matemáticos, empíricos ou estatísticos para identificação e extração dos segmentos textuais relevantes; já nos sistemas que priorizam o conhecimento profundo, prevalecem modelos linguísticos ou discursivos (RINO; SENO, 2006; DIAS-DA-SILVA et al., 2007).

Para avaliação de métodos/sistemas de SA, tem sido comum a realização de conferências internacionais, como a TAC[6] (*Text Analysis Conference*), o que ilustra bem a importância de tais atividades.

Os métodos/sistemas de SA podem ser avaliados de forma intrínseca ou extrínseca. Na avaliação intrínseca, verifica-se o desempenho dos métodos/sistemas analisando seus resultados, ou seja, os sumários. A avaliação extrínseca consiste na análise da utilidade dos sumários em tarefas previamente definidas, como a recuperação de informação (SPARCK JONES; GALLIERS, 1996).

A qualidade dos sumários automáticos é habitualmente avaliada por humanos, visto que o objetivo tende a ser analisar aspectos ligados à gramaticalidade, como ortografia e gramática, e à textualidade, como coesão e coerência (p.ex.: SAGGION; LAPALME, 2000; WHITE et al., 2000), os quais dificilmente podem ser avaliados de forma automática.

Na Sumarização Automática Multidocumento (SAM), a DUC (2007), por exemplo, apresenta vários atributos linguísticos para avaliação da qualidade dos sumários em termos de gramaticalidade, não redundância, clareza referencial, foco, estrutura e coerência. Na SAM abrangendo o português, Castro Jorge e Pardo (2011) fizeram uma avaliação humana pautada em um conjunto de critérios similares aos da DUC: informatividade, coerência, coesão, redundância e gramaticalidade. A informatividade pode ser avaliada quantificando-se a informação relevante dos textos-fonte também presente no sumário automático. Esse levantamento é feito pela comparação automática entre os sumários humanos (também chamados de "sumários de referência") e os sumários automáticos.

Utilizando-se o pacote de medidas da ROUGE (LIN; HOVY, 2003), calcula-se a informatividade pela coocorrência de n-gramas entre os sumários automáticos e os humanos. A informatividade é então expressa pelas medidas de precisão, cobertura e medida-f (LIN; HOVY, 2003; LIN, 2004). Devido à falta de consenso quanto à melhor maneira de se realizar

[6] http://www.nist.gov/tac/about/index.html

uma avaliação como essa, há vários autores que investigaram outras abordagens (p.ex.: SPARCK JONES, 1998; SAGGION et al., 2002; LOUIS, NENKOVA, 2013).

Em Saggion et al. (2002), por exemplo, há 3 propostas de avaliação baseadas na medição da similaridade entre os sumários: (i) similaridade do cosseno[7] (SALTON, 1989), (ii) sobreposição de unidades lexicais (unigramas ou bigramas) e (iii) sobreposição da maior subsequência de unidades lexicais.

No modelo conhecido como "pirâmide"[8] (NENKOVA et al., 2007; LOIUS; NENKOVA, 2013), há 3 formas de avaliação: (i) similaridade entre textos-fonte e sumários, ou seja, métodos que consideram que quanto maior a similaridade entre o sumário e seus textos-fonte, melhor o seu conteúdo; (ii) adição de pseudomodelos, em que utilizam-se não apenas sumários humanos de referência, mas também sumários automáticos escolhidos por humanos; e (iii) sumários automáticos como modelo, isto é, métricas nas quais considera-se que os sumários automáticos têm qualidade suficiente para servirem como sumários de referência, dispensando, portanto, qualquer esforço humano para a criação desses.

2.1.2 A Sumarização Automática Monolíngue

A SA envolvendo apenas uma língua pode ser tanto monodocumento quanto multidocumento. A primeira será abordada na seção 2.1.2.1, enquanto que a última na seção 2.1.2.2.

2.1.2.1 A Sumarização Automática Monodocumento

A SA monodocumento monolíngue tem sido objeto de pesquisas desde a década de 1950 (p.ex.: LUHN, 1958; EDMUNDSON, 1969; O'DONNELL, 1997; SALTON et al., 1997; MARCU, 2000; CONROY; O'LEARY, 2001; PARDO; RINO, 2002; PARDO et al., 2003; RINO et al., 2004; SVORE et al.., 2007; UZÊDA et al., 2010; CLARKE; LAPATA, 2010; LOUIS et al., 2010, etc.).

---

[7] A similaridade do cosseno é uma medida que permite uma normalização dos resultados de comparações entre dois documentos.

[8] O modelo de avaliação da "pirâmide" consiste em atribuir um valor ao sumário avaliando-se a similaridade entre suas *summarization content units* (SCUs) (unidades de conteúdo de sumarização): a SCU que estiver presente em todos os sumários de referência sendo avaliados recebe o maior peso, ocupando a última camada da pirâmide. Pode-se ainda prever o conteúdo ideal que um sumário deve apresentar, uma vez que no topo se localizam as unidades mais importantes.

Conforme descrito em Gupta e Lehal (2010), existem várias estratégias de seleção de conteúdo que podem ser utilizadas, de forma individual ou em conjunto, para a seleção das informações (geralmente sentenças) que irão compor o sumário. Essas estratégias são também empregadas em outras modalidades de SA, como a SAM.

Uma das abordagens é selecionar informações que se relacionam às palavras presentes no título/subtítulo dos textos-fonte, quando disponíveis. Essa estratégia consiste em identificar as palavras que compõem o título, subtítulo e tópicos e usar essas informações para selecionar sentenças contendo as ideias principais dos documentos. A ocorrência ou não de subtítulos e tópicos geralmente depende do gênero e do tamanho dos textos-fonte.

Outra estratégia, segundo esses autores, consiste em realizar a seleção do conteúdo relevante utilizando-se as palavras-chave dos documentos-fonte. As palavras-chave geralmente são os itens lexicais de classe aberta com maior ocorrência nos textos-fonte. Nessa estratégia, pressupõe-se que exista uma correlação entre as palavras mais frequentes de um texto e seu conteúdo principal.

Além do uso de palavras do título/subtítulo e palavras-chave, outro critério frequentemente utilizado para seleção de conteúdo é o tamanho (em número de palavras, por exemplo) das sentenças dos textos-fonte. Com base nesse critério, selecionam-se, preferencialmente, sentenças de tamanho médio para compor um sumário.

Gupta e Lehal (2010) destacam ainda a existência de expressões-chave ou indicativas de conteúdos que caracterizam os elementos da estrutura discursiva de determinados gêneros. Assim, uma das estratégias é selecionar sentenças contendo tais expressões. No gênero científico, por exemplo, a estrutura típica é composta por “Resumo”, “Introdução”, “Materiais e Métodos”, “Resultados”, “Discussão”, “Conclusão” e “Referências”. Essas palavras introduzem conteúdos específicos e indicam o tipo de informação que vem a seguir. Além disso, expressões como “este trabalho tem como objetivo...” também sinalizam informações específicas e fornecem pistas para a seleção de conteúdo.

A localização das sentenças no documento-fonte também pode ser usada como estratégia para a seleção de conteúdo. No gênero jornalístico, é possível gerar sumários selecionando-se as sentenças localizadas logo no início do texto-fonte, visto que essas expressam a informação principal do documento (LUHN, 1958; EDMUNDSON, 1969).

Essas e outras estratégias têm servido para subsidiar métodos tanto superficiais quanto profundos visando à SA monodocumento.

Como exemplo de um método superficial está um trabalho clássico de Baxendale (1958), em que um sumário científico é gerado selecionando-se as sentenças no início e no

final dos parágrafos do texto-fonte. Outro trabalho bastante conhecido na área de SA é o de Luhn (1958), que apresentou um método superficial que consiste em pontuar e ranquear as sentenças com base nas palavras de maior frequência no texto-fonte.

Os trabalhos de Wu e Liu (2003) e Hennig et al. (2008), por sua vez, referem-se a métodos profundos de SA Monodocumento, baseados em conhecimento léxico-conceitual. No método de Wu e Liu (2003), faz-se a identificação da informação topical por meio da comparação entre os termos que ocorrem no texto-fonte e os termos de uma ontologia[9]. Os conceitos dessa ontologia são pontuados, gerando-se uma relação com os principais tópicos e subtópicos do documento-fonte. Com base nessas informações topicais, cada parágrafo do texto-fonte recebe uma pontuação. Por fim, os parágrafos mais bem pontuados vão sendo selecionados até que o sumário atinja o tamanho desejado. De certa forma, pode-se considerar que o método de Wu e Liu (2003) é uma versão mais elaborada do método das palavras-chave, uma vez que objetiva identificar o conteúdo principal de um texto-fonte utilizando a frequência dos conceitos estruturados em uma ontologia.

Outro procedimento utilizado em métodos profundos de SA monodocumento consiste em fazer uso de uma modelagem discursiva do texto-fonte. A teoria *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987), por exemplo, permite modelar um texto-fonte pela elaboração de uma árvore retórica. Nessa árvore, os nós representam as unidades de conteúdo (sentenças, por exemplo) e as arestas representam as relações semântico-discursivas entre as unidades (p. ex.: *evidence*, *concession*, *background*, etc.). Em um texto do gênero jornalístico, a primeira sentença tende a ser a mais nuclear na árvore RST que representa esse texto, sendo, portanto, selecionada para compor o sumário. Uma vez que a RST permite a representação de conhecimento no nível semântico-discursivo, a localização de uma unidade de conteúdo em uma árvore RST é considerada um atributo sentencial profundo.

2.1.2.2 A Sumarização Automática Multidocumento

Dentro da SA, a SAM vem ganhando destaque em razão da crescente demanda por sistemas capazes de lidar com a quantidade cada vez maior de informação disponível na internet. Se a sumarização monodocumento não pode ser considerada uma tarefa simples, os desafios da SAM são ainda maiores. Tratando-se de múltiplos documentos, pode haver trechos com

[9] Na área de PLN, uma ontologia é definida como um recurso ou base de conhecimento que disponibiliza um inventário de conceitos, propriedades e relações entre conceitos cuja função é servir como "uma interpretação da realidade" (ou seja, o conhecimento de mundo partilhado pelos membros de uma determinada comunidade linguística) (GRUBER, 1995).

sobreposição parcial ou total de conteúdo, tornando ainda mais complexa a tarefa de seleção do conteúdo que fará parte do sumário. Além disso, é preciso lidar com incoerências entre documentos. Na SAM de textos jornalísticos, por exemplo, diferenças nas datas de publicação muitas vezes fazem com que as informações do texto mais antigo estejam desatualizadas, o que pode gerar informações contraditórias no sumário.

Dessa forma, a SAM tem de lidar com fenômenos multidocumento, como a presença de informações complementares, contraditórias ou redundantes. O mais comum desses fenômenos é a redundância, consequência da multiplicidade de documentos versando sobre um mesmo assunto. Normalmente, os sistemas/métodos de SAM monolíngue fazem a pontuação e ranqueamento das sentenças dos textos-fonte com base em um critério de relevância e, após o ranqueamento, as sentenças são selecionadas para compor o sumário desde que haja pouca similaridade com relação às sentenças já selecionadas, de modo a minimizar a redundância no sumário (MANI, 2001). Isso significa que os métodos de SAM costumam utilizar um fator de redundância, visando selecionar preferencialmente sentenças com baixo nível de redundância entre si (JURAFSKY; MARTIN, 2000).

Os critérios para pontuação das sentenças e seleção do conteúdo relevante muitas vezes são os mesmos usados na SA monodocumento, como a redundância e a localização da sentença no texto-fonte. A redundância também é importante na SAM, uma vez que uma informação que se repete ao longo de uma coleção de textos-fonte geralmente é relevante e, portanto, deve estar presente no sumário multidocumento (MANI, 2001).

Conforme o tipo de conhecimento linguístico empregado para seleção do conteúdo relevante, os métodos superficiais podem ser classificados em 3 categorias (GUPTA; LEHAL, 2010; KUMAR et al., 2012).

A primeira delas abrange os métodos baseados em atributos linguísticos (*feature-based methods*). Esses atributos podem ser combinados de diferentes formas (p. ex.: LIN; HOVY, 2002; SCHILDER; KONDADADI, 2008) e ter pesos diferentes, dependendo do gênero ou tipo dos textos-fonte a serem sumarizados (BOSSARD; RODRIGUES, 2011; SUANMALI et al., 2011).

Um atributo bastante comum é a frequência de ocorrência de palavras de classe aberta. Trata-se de uma análise relativamente simples, baseada em uma etapa de segmentação sentencial e no cálculo da frequência de ocorrência de cada palavra nos textos-fonte. Na etapa de transformação, as sentenças são pontuadas e ranqueadas de acordo com a soma das frequências das palavras que as compõem. Selecionam-se então as sentenças mais bem pontuadas, desde que não sejam redundantes entre si, até que se alcance a taxa de compressão

estipulada para o sumário. Obtém-se o sumário pela concatenação das sentenças na mesma sequência em que elas constam nos documentos-fonte.

Na segunda categoria estão os métodos/sistemas que utilizam os conceitos de *cluster* (grupo ou coleção) e centroide, como o de Radev et al. (2004). Na etapa de análise dos chamados *cluster-based methods*, faz-se um agrupamento das sentenças de uma coleção em conjuntos (*clusters*) de acordo com sua similaridade lexical. Portanto, esses *clusters* são conjuntos de sentenças semelhantes entre si. Representa-se cada um desses *clusters* em função de um centroide, isto é, um grupo de palavras (ou uma sentença) de maior relevância estatística. Em cada *cluster*, a sentença contendo o maior número de palavras do centroide é então selecionada.

Na terceira categoria de métodos superficiais encontram-se aqueles cuja etapa de análise envolve a modelagem dos textos-fonte sob a forma de grafos (*graph-based methods*) (p.ex.: SALTON et al., 1997; MIHALCEA; TARAU, 2005; WAN, 2008). Nesses métodos, as sentenças são representadas por nós e a similaridade entre elas é representada pelas arestas interligando esses nós. As sentenças com ligações mais fortes são selecionadas para gerar o sumário.

Os métodos profundos também podem ser classificados em 3 categorias, com base no tipo de conhecimento linguístico empregado (MANI, 2001).

A primeira categoria abrange os métodos que utilizam conhecimento sintático. Um exemplo desses métodos é o de Barzilay et al. (1999), em que, na etapa de análise, faz-se uma segmentação sentencial e uma análise da estrutura sintática com o uso de um *parser* (isto é, um analisador sintático). Após a análise sintática, realiza-se o agrupamento das estruturas predicado-argumento mais similares e selecionam-se as estruturas mais frequentes. A etapa de síntese consiste na reordenação das estruturas predicativas e geração de um sumário abstrativo.

Na segunda categoria estão os métodos pautados em conhecimento do tipo semântico-conceitual. Mani e Bloedorn (1997), por exemplo, apresentam um método em que é feita a modelagem de cada texto-fonte em um grafo. Nesse grafo, os nós representam as palavras e as arestas correspondem à similaridade distribucional entre as palavras. Dois nós com arestas semelhantes, por exemplo, indicam palavras sinônimas, expressando um conceito. As sentenças dos textos-fonte contendo palavras que representam os conceitos mais importantes da coleção são selecionadas para gerar o sumário. O método de Li et al. (2010) consiste em indexar as sentenças de uma coleção aos conceitos de uma ontologia. Dada uma consulta (*query*) de um usuário, também mapeada na ontologia, o sistema seleciona para o sumário

unicamente as sentenças indexadas aos mesmos conceitos aos quais os itens lexicais da consulta também foram mapeados (e/ou a conceitos mais específicos).

A terceira categoria contém os métodos baseados em conhecimento semântico-discursivo, que normalmente realizam uma modelagem dos textos-fonte de acordo com a teoria CST (do inglês, *Cross-document Structure Theory*) (RADEV, 2000). Um exemplo de analisador discursivo que permite fazer essa modelagem é o CSTParser, descrito em Maziero (2012). A teoria CST pode ser considerada um desdobramento da RST (*Rhetorical Structure Theory*) (MANN, THOMPSON, 1987), possibilitando interconectar sentenças ou outras unidades textuais de vários textos para formar uma estrutura discursiva (RADEV, 2000). As conexões intertextuais formadas são chamadas de relações CST. Elas permitem identificar fenômenos multidocumento, como similaridades e divergências de conteúdo, e até mesmo o estilo da escrita. Uma relação de equivalência (*equivalence*), por exemplo, significa que as unidades conectadas contêm informação redundante, sendo, portanto, importantes para o sumário. Além disso, pode-se considerar que sentenças com um maior número de relações CST possuem maior relevância, sendo, portanto, uma informação útil para seleção de conteúdo (MANI, 2001). Radev e McKeown (1998) fizeram a primeira proposta para utilização de relações discursivas na SAM, estabelecendo os fundamentos da CST.

Há ainda métodos híbridos de SAM, como o de Schiffman et al. (2002), que utiliza tanto informações superficiais (tamanho das sentenças e localização nos textos-fonte) quanto conhecimento léxico-conceitual. O conhecimento de nível profundo utilizado envolve determinar relações de sinonímia e hiponímia entre as palavras dos textos-fonte, visando identificar os conceitos mais relevantes da coleção. Essas relações são identificadas indexando as palavras à WordNet de Princeton, uma base léxico-conceitual desenvolvida para o inglês americano (FELLBAUM, 1998).

Especificamente para o português, é possível encontrar métodos/sistemas de SAM que utilizam as três abordagens – superficial, profunda e híbrida.

Dentre os trabalhos que seguem a abordagem superficial estão o de Pardo (2005), Akabane et al. (2011) e Ribaldo et al. (2011). No primeiro, Pardo (2005) relata o desenvolvimento do sumarizador GistSumm. Trata-se de um sistema que realiza a pontuação e ranqueamento das sentenças da coleção de acordo com a frequência com que elas ocorrem nos textos-fonte. A sentença que melhor representa o conteúdo principal da coleção, chamada de *gist sentence*, é a que obtém a pontuação mais alta, sendo automaticamente selecionada para o sumário. A seleção de sentenças adicionais para complementar o sumário pauta-se em dois critérios: (i) a sentença deve conter ao menos um radical em comum com a *gist sentence*

e (ii) sua pontuação deve ser maior do que a média da pontuação das demais sentenças. O método descrito por Akabane et al. (2011) é utilizado no sistema CNSumm, fazendo parte do grupo de métodos baseados em grafos. A estratégia utilizada consiste em modelar os textos-fonte em grafos e redes complexas. Também fazendo uso de grafos, Ribaldo et al. (2011) propõem um método de SAM que utiliza um mapa de relacionamentos em combinação com informações obtidas pela CST.

Para o português, o tipo de conhecimento geralmente usado tem sido o semântico-discursivo (CASTRO JORGE; PARDO, 2010; CARDOSO et al., 2011; CARDOSO, 2014). No sistema CSTSumm (CASTRO JORGE; PARDO, 2010), por exemplo, os textos-fonte são modelados na forma de grafos, onde os nós representam as sentenças e as arestas representam as relações CST. As sentenças com maior número de relações CST recebem uma pontuação mais alta, gerando um ranque sobre o qual atuarão operadores de seleção de conteúdo. Esses operadores buscam refletir as preferências do usuário (ex.: exibir ou não informação contextual), promovendo uma reordenação das sentenças no ranque de acordo com o que for importante para o usuário. Com base no ranque final, selecionam-se as sentenças para composição do sumário.

Em Cardoso et al. (2011) e Cardoso (2014), propõem-se métodos de seleção de conteúdo utilizando as teorias RST e CST, visando melhorar a informatividade dos sumários. Nos métodos investigados em Cardoso (2014), a combinação de informações dessas duas teorias levou à produção de sumários mais informativos.

Alguns exemplos de abordagem híbrida são os de Ribaldo et al. (2012), que combina técnicas estatísticas, grafos e relações CST em um sistema denominado RSumm, e Ribaldo (2013), que propõe, além dessas técnicas, a utilização de subtópicos. Nesse último, a identificação dos subtópicos presentes em textos versando sob um mesmo tópico é feita utilizando tanto conhecimento superficial quanto conhecimento profundo. Considera-se que subtópicos de textos diferentes contendo trechos semelhantes trazem informação relevante e, portanto, podem ser agrupados e utilizados para a construção do sumário.

Camargo (2013) também apresenta uma abordagem híbrida para identificação da informação relevante a compor o sumário, investigando atributos linguísticos superficiais e profundos. Tais atributos refletem estratégias de seleção de conteúdo utilizadas por humanos na sumarização multidocumento. As regras formais elaboradas com base na investigação desses atributos visam tornar a SAM mais linguisticamente motivada, uma vez que se apoiam em estratégias de sumarização multidocumento de fato empregadas por humanos, e não apenas em indícios.

2.1.3 A Sumarização Automática de Múltiplas Línguas

O volume de informação disponível em diversos idiomas, especialmente na *web*, tem despertado o interesse por sistemas de SA capazes de lidar com o multilinguismo. O objetivo é permitir que o usuário tenha acesso à informação mesmo que ela seja veiculada em uma língua que o usuário não domine. Os métodos/sistemas de SA que visam lidar com mais de um idioma podem ser classificados como (i) *cross-language*, (ii) multilíngues ou (iii) independentes de língua (ORĂSAN, 2009).

2.1.3.1 Métodos/sistemas cross-language

Nos métodos/sistemas *cross-language* (translíngues), o sumário é gerado em uma língua diferente da do(s) texto(s)-fonte. Esses métodos/sistemas, em geral, englobam um processo de TA que pode ser anterior ou posterior à sumarização. Assim, há os métodos *early translation* e os *late translation*, respectivamente (WAN; LI; XIAO, 2010).

Apesar dos desenvolvimentos recentes, textos traduzidos automaticamente ainda apresentam problemas linguísticos como agramaticalidade e falta de legibilidade. Devido a essa imprecisão dos sistemas de TA, os métodos *early translation* trazem uma pequena desvantagem com relação aos *late translation*, uma vez que nos primeiros existe a tradução total dos textos-fonte, enquanto que nos últimos são traduzidas apenas as sentenças que integrarão o sumário (WAN; LI; XIAO, 2010). Cabe observar ainda que textos traduzidos automaticamente, ao servirem de base para métodos de SA *early translation*, podem conter falhas de tradução que interfiram na aplicação das estratégias de SA.

O método de Wan, Li e Xiao (2010) é um exemplo de uma abordagem *late translation* monodocumento em que se faz a sumarização de uma coleção de textos em inglês para, na sequência, traduzir para o chinês apenas as sentenças selecionadas para compor o sumário. No método de Wan, Li e Xiao (2010), as sentenças dos textos-fonte em inglês de uma determinada coleção são pontuadas e ranqueadas em função da combinação de dois atributos: (i) relevância e (ii) "predição da qualidade da tradução automática" (do inglês, *machine translation quality prediction*). O atributo "relevância" busca identificar as sentenças que apresentam o conteúdo principal da coleção e o atributo "predição da qualidade da TA" busca identificar as sentenças que possuem estrutura mais adequada à tradução automática. Quanto ao índice "qualidade da TA", ressalta-se que, uma vez alto, a tradução da sentença em inglês

para o chinês tem maior probabilidade de ser gramatical, isto é, facilmente lida e entendida pelos falantes dessa língua.

No caso, a relevância de uma sentença é dada pela combinação dos pesos obtidos por dois métodos baseados em atributos linguísticos superficiais e por um método baseado no conceito de centroide (do inglês, *centroid-based method*). Um dos atributos superficiais é a posição da sentença no seu texto-fonte – critério comumente utilizado em outros métodos de sumarização (p. ex.: KATRAGADDA et al., 2009). O outro atributo superficial é a similaridade lexical em relação à primeira sentença do texto-fonte a que pertence: com base no fato de que a primeira sentença de um texto do gênero notícia jornalística veicula o tópico principal, as sentenças dos textos-fonte localizadas no início dos documentos recebem pontuação mais alta que as localizadas no meio ou fim. O mesmo é feito para as sentenças que possuem maior sobreposição de unidades lexicais com a primeira sentença do texto-fonte. Além desses dois métodos baseados em atributos linguísticos, as sentenças também são pontuadas em função do maior número de palavras em comum com o centroide, ou seja, conjunto de palavras estatisticamente importantes da coleção.

O cálculo da "qualidade da TA", por sua vez, é feito em função de atributos superficiais como (i) tamanho da sentença em número de palavras (no caso, sentenças de tamanho médio recebem pontuação maior que as demais), (ii) quantidade de nomes e adjetivos (sentenças com maior número de nomes e adjetivos recebem pontuação maior que as demais) e (iii) e quantidade de pronomes interrogativos iniciados com wh-, em inglês (p.ex.: *who, whom, whose, when*, etc.). Os atributos profundos são basicamente de nível sintático, a saber: (i) profundidade da árvore sintática, (ii) quantidade de sintagmas nominais e (iii) quantidade de sintagmas verbais.

Com base na combinação dos índices de relevância e "qualidade da TA", as sentenças são ranqueadas e as mais bem pontuadas nesse ranque são selecionadas para compor o sumário, até que se atinja a taxa de compressão. Como o método de Wan, Li e Xiao (2010) é do tipo *late translation*, somente as sentenças selecionadas são, na sequência, traduzidas para o chinês e utilizadas na geração do sumário durante a etapa de síntese.

Como exemplos de sistemas de SA *cross-language* multidocumento estão os trabalhos de Orăsan e Chiorean (2008) e Boudin et al. (2011). Em Orăsan e Chiorean (2008), um método superficial sumariza uma coleção de textos-fonte em romeno e o sumário resultante é então traduzido para o inglês. Assim como Wan, Li e Xiao (2010), esse também é um método *late translation*. Outra característica do método de Orăsan e Chiorean (2008) é o fato de ele ser *query-based*, ou seja, o sumário visa trazer informações que atendam a uma consulta do

usuário. No método de Boudin et al. (2011), parte-se de uma coleção de textos-fonte em inglês, do gênero jornalístico, e gera-se um sumário em francês. Entretanto, os textos-fonte em inglês são traduzidos automaticamente para o francês e somente depois sumarizados. Portanto, trata-se de uma abordagem *early translation*. O método de Bourdin et al. (2011) assemelha-se ao de Wan, Li e Xiao (2010) por realizar a seleção de conteúdo com base na qualidade de tradução e no nível de informatividade de cada sentença.

2.1.3.2 Métodos/sistemas multilíngues

Quando os métodos/sistemas de SA fazem a sumarização de uma coleção de textos em diferentes idiomas e geram o sumário em uma dessas línguas, fala-se em Sumarização Automática Multidocumento Multilíngue (SAMM).

Nos trabalhos de Evans et al. (2004) e Evans et al. (2005), por exemplo, a SAMM caracteriza-se pela tradução dos textos-fonte do *cluster* (*early translation*) e pela seleção de conteúdo com base em conhecimento linguístico superficial e profundo, sem estratégias de tratamento dos fenômenos multidocumento.

Em Evans et al. (2005), especificamente, os autores partem de *clusters* de textos jornalísticos traduzidos do árabe para o inglês. Na transformação, as sentenças dos textos traduzidos são pontuadas e ranqueadas em função de alguns atributos linguísticos superficiais que buscam capturar a relevância das mesmas. Para compor o sumário informativo e genérico, selecionam-se apenas as mais bem ranqueadas, até que a taxa de compressão seja atingida.

O atributo linguístico superficial empregado para determinar a relevância da sentença é a ocorrência de "*lead words*", ou seja, palavras que ocorrem nas primeiras sentenças dos textos e que indicam os tópicos dos mesmos. O critério profundo é a ocorrência dos conceitos mais representativos da coleção.

Com o objetivo de evitar problemas de gramaticalidade e/ou inteligibilidade dos sumários, Evans et al. (2005) identificam, em textos originais do inglês que abordam o mesmo assunto que os traduzidos, sentenças similares às sentenças traduzidas que foram selecionadas para compor o sumário. A similaridade entre as sentenças traduzidas e as originais é calculada pela ferramenta Simfinder, que se baseia no compartilhamento de (i) nomes próprios, (ii) itens lexicais morfologicamente relacionados, (iii) sinônimos, (iv) itens lexicais com mesmo hiperônimo e (v) núcleos sintagmáticos (HATZIVASSILOGLOU et al., 2001). A partir da identificação da similaridade entre uma sentença traduzida que foi

selecionada para compor o sumário e uma sentença original em inglês, a sentença original substitui a traduzida na geração ou síntese do sumário.

2.1.3.3 Métodos/sistemas independentes de língua

Algumas abordagens partem do pressuposto que, utilizando conhecimento superficial/estatístico, é possível formular generalizações sobre fenômenos recorrentes em um grande número de línguas. Esses métodos/sistemas que processam diferentes línguas sem fazer uso de conhecimento linguístico profundo são considerados independentes de língua, podendo ser mono ou multidocumento.

Uma dessas abordagens é apresentada em Radev et al. (2004), que descrevem a plataforma *online* MEAD[10]. Trata-se de um sistema que oferece métodos de SA independentes de língua e que podem ser utilizados isoladamente ou em combinação, além de disponibilizar métricas de avaliação intrínseca e extrínseca de sumários. Os métodos de SA na plataforma MEAD utilizam conceitos como centroide, localização da sentença no documento-fonte, tamanho da sentença, palavras-chave e similaridade lexical com relação ao título ou à primeira sentença do texto-fonte (RADEV et al., 2004).

2.1.3.4 A Sumarização Automática Multilíngue e o Português

Com relação à SAMM envolvendo o português, os únicos trabalhos identificados foram os de Tosta et al. (2013) e Tosta (2014).

Tosta et al. (2013) exploraram 2 métodos *baseline* de SAMM com vistas à sumarização de uma coleção multilíngue (inglês, espanhol e português) de 3 textos jornalísticos, sendo o sumário gerado em português. Para essa investigação, foi construído um *corpus* – provavelmente o primeiro *corpus* multilíngue não paralelo envolvendo o português – denominado CM3News (*Corpus Multidocumento Trilíngue de Textos Jornalísticos*)[11]. O CM3News abrange 10 coleções de textos jornalísticos em português, inglês e espanhol, contando com um total de 16.139 palavras. Os documentos que deram origem ao *corpus* foram coletados manualmente por acesso às versões eletrônicas dos jornais *A Folha de São Paulo*[12], *BBC News*[13] e *El pais*[14].

[10]http://www.summarization.com/mead/
[11] http://www.nilc.icmc.usp.br/arianidf/sustento/resources.html
[12]http://www.folha.uol.com.br/

Os métodos de Tosta et al. (2013) são considerados *baseline* pois utilizam tradução automática e empregam conhecimento superficial para o ranqueamento das sentenças. A tradução automática é realizada antes da seleção do conteúdo relevante, sendo, portanto, métodos *early-translation*. A seleção de conteúdo pautou-se em métodos tradicionais de SA, especificamente na posição da sentença no texto-fonte (Método 1) e na frequência de ocorrência das palavras de classe aberta (Método 2). Ambos os métodos são descritos em detalhe a seguir.

No primeiro método, as sentenças são ranqueadas conforme sua posição no texto-fonte. As sentenças localizadas no primeiro parágrafo de cada texto-fonte ocupam o topo do ranque; as sentenças do último parágrafo vão para o final do ranque; e as demais sentenças ocupam posição intermediária. A seleção de conteúdo relevante nesse método consiste nas seguintes etapas: (i) seleção da sentença mais bem pontuada no ranque e integração dessa ao sumário; (ii) seleção da sentença subsequente no ranque; (iii) cálculo da redundância entre as duas sentenças; (iv) integração da nova sentença ao sumário apenas se ela não apresentar deficiências resultantes de TA e se a similaridade entre ela e a sentença já adicionada ao sumário for baixa; (v) substituição da sentença selecionada não-redundante, mas que apresentar deficiência de tradução, por uma sentença semelhante localizada no texto-fonte em português; (vi) repetição das etapas para as demais sentenças até a obtenção da taxa de compressão de 70%.

O cálculo da similaridade, empregado tanto para verificação da redundância quanto para substituição de sentenças que passaram por TA e foram consideradas agramaticais, baseou-se na sobreposição de palavras de classes abertas. Mais especificamente, utilizou-se a métrica de sobreposição de palavras *(word overlap*) (JURAFSKY; MARTIN, 2000) para avaliação da similaridade entre as sentenças, calculada conforme (1).

(1)

$$Word\ Overlap\ (\mathrm{S1}, \mathrm{S2}) = \frac{\text{nº de palavras em comum (S1, S2)}}{\text{nº de palavras (S1)} + \text{nº de palavras (S2)}}$$

Dado um par de sentenças S1 e S2, a sobreposição de palavras é calculada dividindo-se o número de palavras em comum entre elas pela somatória do número de palavras de cada sentença. Nesse cálculo, os resultados podem variar de 0 a 0,5, sendo que valores tendendo a

[13]http://www.bbc.co.uk/news/
[14]http://elpais.com/

0 indicam um par de sentenças com pouca redundância, enquanto que valores tendendo a 0,5 correspondem a sentenças mais redundantes.

No segundo método, calcula-se uma pontuação para cada sentença dos textos-fonte com base na frequência de ocorrência das palavras de classe aberta que a constituem. As sentenças que possuem as palavras mais frequentes da coleção recebem pontuação mais alta e, portanto, ocupam o topo do ranque. Utiliza-se o sumarizador *GistSumm* (PARDO, 2005) para o cálculo da pontuação de cada sentença e a ordenação delas no ranque. Dessa etapa em diante, os mesmos procedimentos do Método 1 são utilizados para seleção manual do conteúdo, englobando tratamento de redundância e dos problemas decorrentes da TA. Os sumários são gerados manualmente, acrescentando-se as sentenças na mesma sequência em que elas são selecionadas.

Ambos os métodos passaram por uma avaliação intrínseca para verificação da qualidade dos sumários produzidos. Para tal, um especialista analisou os 5 parâmetros usados na DUC (*Document Understanding Conference*)[15]: gramaticalidade, não redundância, clareza referencial, foco temático e coerência. Verificou-se que o primeiro método, baseado na localização das sentenças e no tratamento de redundâncias e problemas de TA, obteve pontuações mais elevadas nos 5 critérios avaliados.

Em Tosta (2014), outros 2 métodos são investigados, porém baseados em conhecimento profundo do tipo léxico-conceitual.

No primeiro método, os nomes são indexados a um conjunto único de conceitos. Na sequência, as sentenças recebem uma pontuação de acordo com a frequência de ocorrência desses conceitos na coleção. Gera-se então um ranque com base no qual são selecionadas apenas sentenças oriundas de textos-fonte em português, partindo-se das sentenças mais bem pontuadas, até que uma determinada taxa de compressão seja obtida.

No segundo método, o procedimento é similar, porém as sentenças de maior pontuação no ranque são selecionadas para o sumário independentemente de sua língua-fonte. As sentenças em inglês escolhidas para compor o sumário são traduzidas automaticamente para o português.

Para geração e avaliação dos sumários, construiu-se um *corpus* denominado CM2News, composto por 20 coleções de notícias do gênero jornalístico, sendo que cada coleção é formada por 1 texto em inglês e 1 texto em português versando sobre o mesmo assunto. O *corpus* CM2News passou por anotação semântica, tarefa na qual os nomes comuns

[15] A partir de 2008, a DUC passou a ser parte de outra conferência – a *Text Analysis Conference* (TAC). No site http://duc.nist.gov/ encontram-se informações sobre a DUC no período de 2001 a 2007.

presentes nas coleções foram alinhados semiautomaticamente aos conceitos da WordNet de Princeton.

Os dois métodos foram avaliados intrinsecamente quanto à informatividade e à qualidade linguística dos sumários produzidos. A informatividade dos sumários foi avaliada automaticamente, comparando-os a sumários de referência por meio do pacote de medidas ROUGE. A avaliação da qualidade linguística consistiu em submeter os sumários ao julgamento de 15 especialistas em linguística computacional para que analisassem, manualmente, os seguintes aspectos: gramaticalidade, não redundância, clareza referencial, foco e estrutura/coerência. O primeiro método, cujas sentenças provêm de um mesmo texto-fonte na língua do usuário, obteve melhores resultados em ambas as avaliações. Comparados aos métodos mais simples em que se faz a TA integral dos textos-fonte, os dois métodos obtiveram melhor desempenho, o que indica que a utilização de conhecimento léxico-conceitual é uma estratégia que merece ser mais bem investigada no âmbito da SAMM.

2.1.4 Avaliação de métodos de Sumarização Automática

2.1.4.1 Avaliação intrínseca da qualidade linguística

O objetivo da avaliação intrínseca na SA é verificar a qualidade linguística e a informatividade de métodos e/ou sistemas de sumarização. Conforme a complexidade da tarefa de sumarização aumenta – por exemplo, com a utilização de um número maior de textos-fonte ou a inclusão de mais línguas – a tendência é que os problemas linguísticos nos sumários também aumentem. Alguns dos problemas comuns na SA monodocumento referem-se à coesão e coerência. Na SAM, além desses problemas, há também questões como a maior presença de contradições e redundância. A partir do momento em que se passa a lidar com mais de uma língua, somam-se ainda mais alguns desafios, como a agramaticalidade decorrente da TA (etapa comum em muitos sistemas de SAMM).

Na DUC (2007), os critérios apresentados para avaliação da qualidade de sumários produzidos automaticamente são: gramaticalidade, não redundância, clareza referencial, foco e estrutura/coerência. Cada um desses critérios é brevemente apresentado a seguir.

- Gramaticalidade: o sumário não deve conter erros como sentenças iniciadas com letras minúsculas, formatação incorreta, sentenças interrompidas ou outros tipos de incorreções gramaticais que prejudiquem a legibilidade do texto.
- Não redundância: o sumário não deve conter repetições desnecessárias.

- Clareza referencial: o papel de uma pessoa ou entidade mencionada deve ser suficientemente claro no sumário, possibilitando ao leitor saber a "quem" ou "o que" cada pronome e sintagma nominal se refere.
- Foco: as informações do sumário devem estar inter-relacionadas, permitindo ao leitor identificar um foco temático.
- Estrutura/coerência: um sumário não é apenas um aglomerado de informações sobre um mesmo assunto. Assim sendo, o sumário deve apresentar uma organização adequada e estrutura coerente.

2.1.4.2 Avaliação intrínseca de informatividade

O conjunto de medidas ROUGE (LIN; HOVY, 2003) permite calcular a informatividade de um sumário pela determinação da coocorrência de n-gramas. Na ROUGE, define-se um n-grama como uma palavra ou um conjunto de palavras que ocorrem sequencialmente. Um n-grama contém de 1 a 4 palavras e, conforme o número de n-gramas, a medida ROUGE recebe uma subdivisão: a ROUGE-1 avalia a coocorrência de unigramas; a ROUGE-2 mede a coocorrência de bigramas; e assim por diante.

As três medidas obtidas pela aplicação da ROUGE são a precisão (P), cobertura (C) e medida-f (em inglês, *precision*, *recall* e *f-measure*, respectivamente). A precisão e a cobertura são calculadas pelas fórmulas:

$$\mathbf{P=}\ \frac{\textbf{Número de n-gramas em comum com o sumário de referência}}{\textbf{Número de n-gramas do sumário automático}}$$

$$\mathbf{C=}\ \frac{\textbf{Número de n-gramas em comum com o sumário de referência}}{\textbf{Número de n-gramas do sumário de referência}}$$

Já a medida-f combina a precisão e a cobertura por meio da seguinte fórmula:

$$\textbf{Medida-f=}\frac{\textbf{(P x C)}}{\textbf{(P + C)}}\textbf{x2}$$

Os resultados da medida-f variam entre 0 e 1. Quanto mais o valor se aproxima de 1, mais informativo é o sumário; quanto mais próximo de 0, mais baixa a informatividade.

As medidas da ROUGE proporcionam uma maneira de se avaliar automaticamente, de maneira intrínseca, a informatividade de um sumário. Pela comparação entre sumários automáticos e sumários de referência, é possível obter-se uma medida de quanta informação o sumário automático preserva com relação aos documentos-fonte.

A seguir serão abordados os sistemas de representação de conhecimento para, na sequência, discutir a relação desses com a área de SA e mais especificamente a SAMM.

2.2 Sistemas de Representação de Conhecimento

O campo de estudo da Representação de Conhecimento (RC) é vasto e abrange desde aspectos filosóficos da epistemologia até questões mais práticas, como o tratamento de informações em grande escala. Entretanto, mesmo sob essas diferentes perspectivas, há uma questão em comum: o desafio de codificar o conhecimento humano de forma que ele possa ser utilizado (WELTY, 1995).

Embora as línguas naturais constituam o sistema de representação de conhecimento mais poderoso existente, a vagueza e ambiguidade inerentes a elas ainda representam um desafio considerável para seu tratamento computacional. Assim, a alternativa encontrada para representação de conhecimento no computador tem sido a utilização de linguagens formais, visto que elas proporcionam menos erros de interpretação decorrentes de ambiguidades sintáticas ou falta de clareza semântica (GRISHMAN, 1986).

Embora com objetivos diferentes, Jackendoff (1983, 1985) e Sowa (1984) desenvolveram algumas das teorias linguísticas que serviram de base para a representação de conhecimento semântico sobre o mundo. Sowa buscava encontrar soluções para viabilizar a criação de bancos de dados semânticos, enquanto que Jackendoff mostrava-se mais preocupado em desenvolver um modelo de representação conceitual para a compreensão e produção linguísticas.

Dentre os desenvolvimentos mais importantes em sistemas de representação de conhecimento, destacam-se as linguagens baseadas em frames, cuja ideia central é a utilização de objetos (conceitos) e a relação entre esses objetos. Conforme Welty (1995), as principais características dessas linguagens são:

- Orientação a objetos: as informações referentes a um conceito são armazenadas junto com o próprio conceito (ao contrário de sistemas baseados em regras, por exemplo, nos quais as informações podem estar dispersas).

- Generalização e especialização: as linguagens de representação de conhecimento agrupam os conceitos hierarquicamente, de modo que os conceitos nos níveis mais elevados representam atributos mais gerais, compartilhados pelos conceitos nos níveis inferiores.
- Raciocínio: trata-se da capacidade de inferir, por meio de um formalismo, que a existência de um determinado conhecimento implica a existência de outro conhecimento não previamente conhecido.
- Classificação: a partir de uma descrição abstrata de um conceito, a maioria das linguagens de RC consegue estabelecer se um conceito se encaixa na descrição dada.

Além das linguagens baseadas em frames, outros dois formalismos de grande importância para a RC são os sistemas de produção e os sistemas de banco de dados. Os sistemas de produção baseiam-se em regras condicionais do tipo se-então (*if-then*), porém mostram limitações ao lidar com problemas complexos. Dentre as razões para isso estão a falta de ordenação nas regras e a impossibilidade de limitar inferências somente a objetos de interesse. Os sistemas de produção enquadram-se nos sistemas baseados em frames, sendo que os últimos têm como vantagem permitir inferências, como no caso da classificação e da herança de propriedades, e fazer uso de estratégias de estruturação do conhecimento, como generalização e orientação a objetos. As regras de inferência constituem uma parcela importante do conhecimento que se tem sobre um determinado domínio. Os sistemas de banco de dados são capazes de representar apenas afirmações simples, porém, sem permitir inferências (WELTY, 1995).

Uma das principais motivações para o desenvolvimento de sistemas de representação de conhecimento é a sua utilização em processos computacionais. Quando esse é o objetivo primordial, a representação de conhecimento pode ser vista como um meio para computação eficiente (DAVIS et al, 1993). Dentre as características desejáveis para um bom sistema de representação de conhecimento estão a precisão, abrangência, facilidade de alteração e atualização, consistência, eficiência, legibilidade e clareza (MARTIN, 2002; ALANSARY; NAGI, 2013).

Na área de PLN, especificamente, Dias-da-Silva (2006) sugere que o desenvolvimento de sistemas envolve uma espécie de “engenharia do conhecimento linguístico”, beneficiando-se, portanto do arcabouço metodológico da Engenharia do Conhecimento. Assim, os modelos de representação de conhecimento podem ser úteis no processamento das línguas naturais.

Um dos estágios do processamento em sistemas de PLN é a extração do significado de expressões superficiais em línguas naturais e sua representação por meio de estruturas formais. Para isso, utilizam-se metalinguagens cujo componente vocabular permite representar os conceitos do mundo e cujo componente gramatical estabelece as possíveis relações entre esses conceitos. O que se faz nessa etapa, portanto, é identificar os conceitos e determinar quais são as relações existentes entre esses conceitos (SPECIA; RINO, 2002).

Todavia, é uma tarefa complicada "traduzir" o conhecimento expresso pelas línguas naturais para linguagens formais, que, comparativamente, são muito mais simples. Trata-se de um problema que começou a ser investigado no final da década de 80, porém, segundo Cardeñosa et al. (2008), ainda não foi bem solucionado.

Dentre as tentativas de se utilizar sistemas de representação de conhecimento na área de PLN e, mais especificamente, na SA, destaca-se a linguagem de representação IRep4 (LENCI et al., 2002). Nesse sistema, representa-se o conteúdo de cada sentença selecionada em uma estrutura hierárquica do tipo atributo-valor (KAY; FILLMORE, 1999). No caso, os elementos básicos são conceitos atômicos (p.ex.: "homem" e "carro") e predicativos (p.ex.: "comprou"). Em (2), descreve-se a representação do conteúdo da sentença "*O homem comprou um carro*" em função da IRep4.

(2)
```
PROP {
      Type = TYPE_PROP;
      Value = P_ARG1_compr_ARG2;
      CAT = V_SEN;
      Time_Rep = [PRESENT, PRES_USUAL];
      Arg1 = ITEM {
            Type = TYPE_ITEM;
            Value = C_homem;
            DET = def;};
      Arg2 = PROP {
            Type = TYPE_PROP;
            Value = P_carro_ARG1;
            CAT = N;
            DET = indef;};}
```

As expressões em IREP4 são formadas recursivamente utilizando-se os elementos PROP, para proposições, e ITEM, para termos. Esses elementos constituem estruturas formais capazes de expressar informações como definitude, estrutura argumental, tempo verbal e correferência. Além de permitir a representação de propriedades semânticas, as expressões em IREP4 também são capazes de registrar informações sintáticas. O atributo CAT, por exemplo,

especifica se uma determinada proposição aparece no texto-fonte como uma sentença ou como um sintagma nominal.

Outro sistema de representação com potencial para aplicações em SA é a linguagem UNL – *Universal Networking Language* (UCHIDA et al., 1999) – que, dada a sua relevância em particular para este projeto, será discutida em separado no item 2.2.1.

2.2.1 O Projeto UNL

O programa UNL foi iniciado em 1996 pelo Instituto de Estudos Avançados da Universidade das Nações Unidas, em Tóquio, sob o comando da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO). Em 2001 foi criada a *Universal Networking Digital Language* (UNDL[16]) *Foundation*, uma organização autônoma, não governamental, com sede na Suíça e que, desde então, tem sido responsável pela UNL. O projeto conta até agora com 16 línguas (UCHIDA et al., 1999; MARTINS, R. T., 2012; ALANSARY; NAGI, 2013).

A UNL pode ser vista como uma linguagem para computadores que visa possibilitar a comunicação em diferentes idiomas, codificando conhecimentos e informações provenientes de línguas naturais. Dentre as vantagens oferecidas por esse formalismo, estão sua natureza plurilinguística, envolvendo muito mais línguas do que outras abordagens similares disponíveis, e a possibilidade de se obter representações logicamente precisas, não ambíguas e teoricamente independentes de idiomas (MARTINS, R. T. et al., 2000; TIECHER, 2003).

Cardeñosa et al. (2008) observam que há uma lacuna entre as línguas naturais, bastante complexas, e as linguagens de representação de conhecimento, mais simples. Essa lacuna seria uma das causas das grandes dificuldades enfrentadas quando se tenta representar o conhecimento expresso por línguas naturais em linguagens formais. Segundo esses autores, a UNL funciona como uma representação intermediária entre as línguas naturais e as linguagens computacionais, preenchendo a lacuna entre ambas.

2.2.1.1 O sistema UNL

Na UNL, a informação expressa por cada sentença é representada por uma rede semântica composta por três tipos de unidades (NUNES et al., 2001; MARTINS, R. T. et al., 2002;

[16] http://www.undl.org/

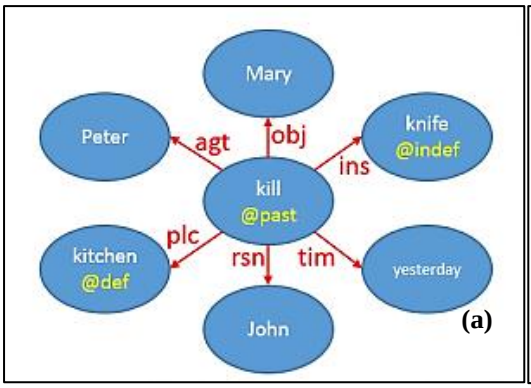
UNDL FOUNDATION, 2003; SPECIA; RINO, 2004; CARDEÑOSA et al., 2008; MARTINS, R. T.; AVETISYAN, 2009):

a) os conceitos universais ou palavras universais (UWs = *Universal Words*), que formam os nós dessas redes semânticas;
b) as relações binárias (RLs = *Relation Labels*), que interligam esses nós; e
c) os atributos (ALs = *Attribute Labels*), que servem para agregar informações às UWs.

As UWs são usadas para expressar o significado de qualquer conceito existente, estando associadas às palavras de categorias lexicais abertas – substantivos, verbos, adjetivos e advérbios. As relações binárias formam um conjunto pré-definido de casos semânticos e são expressas por símbolos de duas ou três letras – p. ex.: `agt` (agente) e `qua` (quantidade). Já os atributos permitem a representação de morfemas ligados e palavras de classes fechadas, tais como afixos (p. ex.: gênero, número, tempo, aspecto e modo), determinantes (artigos e demonstrativos), adposições[17], conjunções, advérbios de grau e verbos auxiliares e quase-auxiliares (UNDL FOUNDATION, 2003; MARTINS, R. T., 2012; ALANSARY; NAGI, 2013).

Exemplificando essas entidades, a sentença em inglês "*Peter killed Mary yesterday with a knife in the kitchen because of John*" ("Pedro matou Mary ontem com uma faca na cozinha por causa de John") poderia ser representada em UNL, de maneira simplificada, conforme ilustra a Figura 3.

Figura 3 – Exemplo de representação (a) gráfica e (b) textual em UNL




```
agt(kill.@past,Peter)
obj(kill.@past,Mary)
ins(kill.@past,knife.@indef)
tim(kill.@past,yesterday)
rsn(kill.@past,John)
plc(kill.@past,kitchen.@def)
```

(b)

Fonte: *UNDL Foundation* (2013a)

onde:

[17] As adposições englobam preposições, posposições e circumposições, sendo que as duas últimas inexistem no português.

- `Peter`, `kill`, `Mary`, `yesterday`, `knife`, `kitchen` e `John` são conceitos universais, ou *UWs*;
- `agt` (*agent*), `obj` (*patient*), `tim` (*time*), `ins` (*instrument*), `plc` (*place*) e `rsn` (*reason*) são relações; e
- `@past`, `@def` e `@indef` são atributos.

Cardeñosa et al. (2008) destacam algumas propriedades da UNL que viabilizam a sua utilização em tarefas que dependam de representação de conhecimento: i) a UNL não se restringe a um domínio específico; ii) ela pode ser utilizada para representar conhecimentos em qualquer língua; e iii) o número de relações é definido e capaz de descrever qualquer tipo de informação veiculada pelas línguas naturais.

Além do sistema de representação em si, há uma série de recursos baseados na UNL (ALANSARY; NAGI, 2013). Esses componentes são normalmente disponibilizados na UNLWeb[18] e podem ser classificados em três tipos:

a) recursos linguísticos: gramáticas, *corpora* e bases de dados lexicais (dicionários, bases de conhecimento e memórias);
b) ferramentas para representação manual em UNL; e
c) mecanismos (em inglês, *engines*) de representação automática.

O processo de representação de uma língua natural em UNL é chamado, em inglês, de *enconversion* ou *UNLization* (UNLização). Embora a representação resultante desse processo conserve as mesmas divisões entre sentenças do texto-fonte, o objetivo não é reproduzir a sintaxe ou escolhas lexicais originalmente empregadas. A meta é representar, sem ambiguidade, um dos significados possíveis do texto original – de preferência o significado mais provável (MARTINS, R. T., 2012).

2.2.1.2 Ferramentas e métodos de UNLização

Com relação à automaticidade do processo de representação, existem quatro métodos de UNLização (UNDL FOUNDATION, 2013b):

a) UNLização totalmente automática: o processo é realizado automaticamente por uma ferramenta, sem a interferência do usuário;

[18] http://www.unlweb.net/

b) UNLização automática auxiliada por humanos: o processo é executado por uma ferramenta mas existe algum grau de intervenção humana, antes, durante ou depois do processo;
c) UNLização humana assistida por máquina: o usuário é quem faz a UNLização, porém contando com o auxílio de ferramentas, como dicionários ou ontologias;
d) UNLização totalmente manual: o processo todo é realizado pelo usuário, sem a utilização de ferramentas.

A **precisão** do processo de UNLização pode ser profunda ou superficial: na UNLização profunda, a estrutura semântica final reproduz a estrutura sintática do documento original; já na UNLização superficial, a estrutura sintática original não é preservada. Como a própria nomenclatura sugere, na UNLização profunda, o alvo é a estrutura semântica profunda do documento original, enquanto que a UNLização superficial tem como foco a organização semântica superficial do documento (UNDL FOUNDATION, 2013b).

Exemplificando essa diferença, a sentença em inglês "*Mary saw Peter going to Paris*" teria a seguinte estrutura:

UNLização superficial: *Mary saw Peter* & *Peter was going to Paris*

UNLização profunda: *Mary saw* [*Peter going to Paris*]

Atualmente existem quatro ferramentas para UNLização: IAN, SEAN, UNL Editor e Enco (Quadro 1). As ferramentas IAN, UNL Editor e Enco são consideradas sistemas de UNLização profunda, enquanto que a ferramenta SEAN é considerada superficial.

Quadro 1 – Sistemas de UNLização disponíveis

Ferramenta	Utilização	UNLização	Método
IAN	Gratuita	Profunda	Automático ou Semiautomático
SEAN	Gratuita	Superficial	Automático
UNL Editor	Gratuita	Profunda	Manual
Enco	Paga	Profunda	Automático

Fonte: Adaptado de *UNDL Foundation* (2013b, 2013c)

2.3 Interlínguas e Sumarização Automática

2.3.1 Interlínguas

Uma interlíngua é uma língua artificial criada com a finalidade de se obter uma representação neutra do significado linguístico. A origem das interlínguas está, primordialmente, nos formalismos de representação do conhecimento e nas interlínguas usadas para TA.

Idealmente, uma interlíngua deve ser precisa, não ambígua e independente de outras línguas. Também é desejável que ela seja tão expressiva quanto as línguas naturais. Atingir esse objetivo, entretanto, não é uma tarefa simples: a construção de uma interlíngua que consiga balancear adequadamente esses requisitos e opere com muitas línguas em domínios abertos não foi bem sucedida até então. Em domínios restritos, envolvendo um número bastante limitado de línguas, alguns sistemas de TA baseados em interlíngua têm conseguido êxito (CARDEÑOSA et al., 2005).

2.3.2 Sistemas de Representação de Conhecimento como Interlínguas na SA

Na SAMM, uma das alternativas que vêm sendo estudadas para contornar a barreira linguística é a utilização de sistemas de representação conceitual que possam funcionar como interlínguas. Há, no entanto, possíveis desvantagens que devem ser levadas em consideração, como a dependência de domínio e a consequente baixa escalabilidade (LENCI et al., 2002).

Dentre as tentativas de tratar o multilinguismo na SA por meio do uso de interlíngua está o projeto MLIS-MUSI (LENCI et al., 2002). No sistema desenvolvido, as sentenças de um texto são ranqueadas com base em atributos linguísticos superficiais, como a ocorrência de expressões indicativas (do inglês, *cue phrases*) e a posição da sentença no texto-fonte. Ainda na etapa de transformação, o conteúdo das sentenças é mapeado ao formalismo IRep4, gerando-se uma representação independente dos idiomas dos textos-fontes. O IRep4 faz,

portanto, a conexão entre as etapas de análise e de síntese: as informações extraídas na etapa de análise das sentenças são convertidas para essa representação, que por sua vez alimenta os módulos de síntese. A representação conceitual das sentenças selecionadas por meio do IRep4 funciona como uma interlíngua, a partir da qual o sumário correspondente ao texto-fonte, originalmente na língua Lx, pode ser automaticamente gerado em diferentes línguas-alvo, Ly, Lw, etc.

Em Martins (2002) e Martins e Rino (2002b) é apresentado um método de SA monodocumento baseado no mapeamento do conteúdo de um único texto-fonte (do gênero legal ou científico) à linguagem UNL. Em (3), ilustra-se a representação UNL de "*Juliana went to college by car*".

(3)
```
agt(go.@past,Juliana)
plt(go.@past,college)
met(go.@past,car)
```

A partir da codificação em UNL de todas as sentenças de um texto-fonte x, aplicam-se regras que excluem informações irrelevantes que, por isso, não devem compor o sumário (informativo e genérico) do texto x. Tais regras foram inicialmente desenvolvidas a partir de análise de *corpora* de textos e, naturalmente, a delimitação das informações a serem excluídas depende dos gêneros e domínios dos textos-fonte. Como resultado da aplicação dessas regras, gera-se uma versão condensada do texto UNL.

No caso da sumarização de textos científicos, uma das regras elaboradas prevê a exclusão de relações rotuladas por `met`, posto que "modo" é informação complementar que não deve estar presente em sumários científicos. Ao aplicar essa regra à sentença em (3), por exemplo, descarta-se `met(go.@past,car)`. Consequentemente, a versão condensada de (3) é composta somente por `agt(go.@past,Juliana)` e `plt(go.@past,college)`, que equivale a "*Juliana went to college*". Tendo em vista que tais regras realizam a redução das sentenças dos textos-fonte, diz-se que a sumarização é intrassentencial. Ao final, a versão condensada da representação UNL é então convertida em língua natural.

Também trabalhando com a UNL, Sornlertlamvanich et al. (2001) propõem um método de sumarização envolvendo quatro etapas: i) cálculo de uma pontuação para cada sentença representada em UNL; ii) seleção das sentenças mais bem pontuadas para composição do sumário; iii) remoção das palavras ou frases redundantes nas sentenças selecionadas; iv) combinação de parte das sentenças selecionadas para melhoria do sumário.

A pontuação de uma sentença é calculada com base no peso de cada palavra que a compõe. O peso de cada palavra, por sua vez, é calculado levando-se em conta a frequência do termo (*term frequency*) e a frequência inversa nos documentos (*inverse document frequency*), conforme a seguir:

$$S(s) = \sum_{\forall UWi \in s} W(UWi) \quad (4)$$

$$W(UWi) = Tf(UWi) * Idf(UWi) \quad (5)$$

$$Idf(UWi) = \log\left(\frac{N(UWi)}{n(UWi)}\right) \quad (6)$$

onde:

S é a função de pontuação das sentenças;

s é a sentença sendo pontuada;

W é a função que calcula o peso de cada conceito;

UWi é o conceito;

Tf é a frequência de ocorrência do conceito;

Idf é a frequência inversa do conceito nos documentos;

N(UWi) é o número de documentos do *corpus;* e

n(UWi) é o número de documentos em que a UW ocorre.

A etapa de seleção de conteúdo proposta por Sornlertlamvanich et al. (2001) consiste em pontuar as sentenças conforme mencionado e selecionar as sentenças com melhor pontuação. Essas sentenças passam, então, por uma etapa de eliminação de informações redundantes, conforme exemplificado no Quadro 2. As palavras removidas são geralmente modificadores, que podem ser identificados pelas relações UNL em que aparecem. Por fim, é proposta uma etapa de combinação de sentenças: no caso de sentenças que compartilhem uma mesma UW e tenham menos de 15 palavras, elas poderão passar por um processo de junção, diminuindo assim a redundância intersentencial.

Quadro 2 – Eliminação de palavras redundantes usando a UNL

Sentença original	**Sentença após eliminação de redundâncias**	**Palavras removidas**
UNL represents the means to facilitate multilingual communication on the information network.	*UNL represents the means to facilitate multilingual communication on the network.*	*information*
The language exists only on the information network.	*The language exists on the network.*	*only, information*

Fonte: Adaptado de Sornlertlamvanich et al. (2001)

Pandian e Kalpana (2013) também apresentam uma proposta de SA envolvendo a UNL. Nesse sistema, a ideia é que o documento seja analisado, representado em UNL, sumarizado e transformado de UNL para língua natural. Quando finalizado, o sistema deverá gerar sumários em três níveis de complexidade diferentes, de acordo com o perfil intelectual do usuário. A sumarização, em si, consistirá em remover informações irrelevantes tais como sintagmas preposicionais e determinantes, com base na representação em UNL. Entretanto, não houve maior detalhamento de como isso será feito.

O que se pode notar na revisão da literatura é que, de um modo geral, a multiplicidade das línguas envolvidas no processo de SAMM costuma ser tratada por meio da TA direta dos textos-fonte, sendo que esta, por não ser totalmente precisa, gera problemas que precisam ser contornados. Além disso, observa-se que, em geral, pouco se sabe sobre as vantagens e desvantagens em se utilizar uma representação conceitual como a UNL para a SAMM. Um sistema de representação de conhecimento poderia ser útil para a elaboração de estratégias de seleção de conteúdo que de fato reflitam a sumarização multidocumento realizada por humanos? Até que ponto a observação das representações conceituais das sentenças de textos-fonte e de seus sumários permitiria compreender melhor a sumarização humana multidocumento? Uma análise como essa poderia levar a melhores estratégias de seleção do conteúdo relevante para elaboração do sumário? Questões como essa ainda permanecem em aberto e talvez, se melhor investigadas, possam auxiliar no desenvolvimento ou aprimoramento de sistemas de SAMM baseados em conhecimento profundo.

3 SELEÇÃO E REPRESENTAÇÃO DO *CORPUS*

3.1 Seleção do *corpus*

Um *corpus* pode ser definido, de maneira simplificada, como um conjunto de dados linguísticos coletados com o intuito de servir de evidência em pesquisas. Importantes ferramentas na área de PLN, os *corpora* devem refletir o uso efetivo da língua. Assim sendo, eles podem ser compostos, por exemplo, por publicações escritas ou até mesmo por transmissões de rádio ou conversas do dia-a-dia, a depender do propósito da investigação. Normalmente os *corpora* são construídos de forma que possam ser processados computacionalmente, visto que realizar análises em conjuntos grandes de dados impressos ou gravados em material audiovisual nem sempre é uma alternativa prática (McENERY, 2003; SARDINHA, 2004).

Um dos motivos para a utilização de *corpora* é o fato de eles permitirem a busca de padrões linguísticos. Neste projeto, em particular, utilizou-se um *corpus* para tentar identificar estratégias de seleção de conteúdo que se projetem na representação conceitual dos textos-fonte, fornecendo subsídios para a SAMM.

A escolha de um *corpus* está diretamente relacionada ao tema central da pesquisa. Para os propósitos desta investigação, definiu-se que o *corpus* seria multidocumento, multilíngue e composto por coleções contendo:

- 1 texto em PB;
- 1 texto em EN;
- 1 sumário multidocumento multilíngue em PB produzido por humanos.

A escolha desses idiomas deveu-se à necessidade de pesquisas na área de SAMM envolvendo o PB e à grande disponibilidade de textos em inglês na internet, além da relevância internacional deste último.

Quanto ao gênero dos textos, deu-se preferência ao jornalístico, dada a presença desses no cotidiano, a facilidade de se coletar textos em diferentes idiomas e a tradição de pesquisas com documentos desse gênero em SA.

Definiu-se também que o *corpus* deveria ser não paralelo, mas sim de conteúdo comparável. Isso significa que, além de multilíngue, o *corpus* deveria conter coleções de textos versando sobre um mesmo assunto, mas que não fossem traduções um do outro (McENERY, 2003; CASELI; NUNES, 2004). Esse requisito também decorreu do objetivo da

pesquisa, mais especificamente do intuito de analisar como ocorre a seleção de conteúdo multidocumento. A análise de um mesmo documento traduzido para várias línguas não permitiria uma investigação como essa.

Foi iniciada uma busca por *corpora* que já tivessem sido previamente UNLizados, no intuito de verificar sua adequação aos objetivos deste projeto. O resultado desse levantamento é mostrado no Quadro 3.

Embora alguns dos *corpora* levantados, como o *Org Information* e o *ITU*, disponham de textos em mais de um idioma, tratam-se de traduções de um mesmo texto-fonte, ou seja, são *corpora* paralelos. O mesmo ocorre com as obras *Crátilo*, de Platão, e *O Pequeno Príncipe*, de Exupèry.

Além do mais, dentre os *corpora* já UNLizados, somente 3 deles contam com sumários: *Theses*, *UNU* e *EOLSS*. No entanto, nesses 3 *corpora*, os sumários foram originados a partir de documentos em uma única língua. Infelizmente nenhum dos *corpora* atualmente disponíveis em UNL conta com sumários criados a partir de mais de um documento e mais de uma língua. Como o foco deste trabalho foi a investigação de estratégias de seleção de conteúdo na SAMM, nenhum desses *corpora* atendeu aos requisitos do projeto.

Considerando-se que não havia nenhum *corpus* já UNLizado que atendesse aos requisitos necessários para o desenvolvimento desta pesquisa, a alternativa que restou foi realizar a UNLização de um *corpus* que atendesse aos critérios estabelecidos.

Um *corpus* que atendeu satisfatoriamente a todos os requisitos foi o CM2News – *Corpus* Multidocumento Bilíngue de Textos Jornalísticos (TOSTA, 2014). O CM2News foi construído pelo recorte e extensão de outro *corpus* – o CM3News (TOSTA et al., 2012, 2013).

Os textos do CM2News foram coletados manualmente na internet, pelo acesso às versões *online* dos jornais *A Folha de São Paulo*[19] e *BBC News*[20]. Os critérios para escolha dos domínios foram a variedade e a atualidade dos temas; já para a seleção dos textos, foram considerados a originalidade e o tamanho, dando-se preferência a textos com tamanho similar. O resultado da aplicação desses critérios foi um *corpus* com notícias diversificadas, publicadas entre 2011 e 2013, abrangendo um total de 6 domínios: mundo, poder, saúde, ambiente, ciência e entretenimento.

[19] http://www.folha.uol.com.br/

[20] http://www.bbc.co.uk/news/

Quadro 3 – Relação de *corpora* em UNL

Corpus	**Referências**	**Idioma do(s) texto(s)-fonte**	**Há sumário disponível em UNL para o *corpus*?**	**Ano**
Theses	Martins, C.B. e Rino (2002b)	Português	Sim	Não informado
UNU	Martins, C.B. e Rino (2002b)	Inglês	Sim	Não informado
UN Charter	Martins, C.B. e Rino (2001) Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês	Não	1997
The Tower of Babel	Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês	Não	1997
The World Cup History	Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês	Não	1998
Press Release	*UNDL Foundation* (2012a)	Inglês	Não	1998
Great Barrier Reef	Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês	Não	1999
Love	Martins, R.T. (2008)	Inglês	Não	1999
Org Information	*UNDL Foundation* (2012a)	Inglês[21]	Não	2000
ITU	Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês[22]	Não	2001
UNL News	Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês	Não	2002
UNESCO	Martins, R.T. (2008) *UNDL Foundation* (2012a)	Inglês	Não	2003
Cratylus	Martins, R.T. (2007) *UNDL Foundation* (2012a)	Inglês	Não	2004
UNL Documents[23]	*UNDL Foundation* (2006) Martins, R.T. (2008)	Inglês	Não	2005
EOLSS	*UNDL Foundation* (2009, 2010)	Inglês	Sim	2005
Le Petit Prince	*UNDL Foundation* (2012b)	Francês	Não	2010
IGLU	*UNDL Foundation* (2012c)	Inglês	Não	2010
UNL Reference Corpus	*UNDL Foundation* (2012d)	Inglês	Não	2012

Fonte: Elaborado pelo autor

[21] Conta com traduções para: alemão, árabe, chinês, espanhol, hindu, indonésio, italiano, japonês, letão, português, russo e tailandês.

[22] Conta com traduções para: árabe, espanhol, hindu, indonésio e russo.

[23] *Corpus* também chamado de "*Other*" em *UNDL Foundation* (2012) e "*Current*" em Ronaldo T. Martins (2008).

Quadro 4 – Coleções do *corpus* CM2News

Coleção	Domínio	Assunto/ Tema	Documento	Língua	Publicação (data/hora)	No. de palavras
C1	Mundo	Ataques em Londres	D1_C1_folha	PT	11/08/2011 – 09:11	1.311
			D2_C1_bbc	EN	11/08/2011 – 11:10 (GMT)	
C2	Poder	Kit gay	D1_C2_folha	PT	25/05/2011 – 13:12	516
			D2_C2_bbc	EN	25/05/2011 – 21:07 (GMT)	
C3	Saúde	Intoxicação alimentar	D1_C3_folha	PT	30/05/2011 – 18:47	1.419
			D2_C3_bbc	EN	30/05/2011 – 5:43 (GMT)	
C4	Mundo	Massacre na Noruega	D1_C4_folha	PT	08/08/2011 – 14h20	911
			D2_C4_bbc	EN	02/08/2011 – 14:52 (GMT)	
C5	Ambiente	Novo código florestal	D1_C5_folha	PT	25/05/2011 – 00:43	1.217
			D2_C5_bbc	EN	25/05/2011 – 09:50 (GMT)	
C6	Mundo	Conflito na universidade da CA	D1_C6_folha	PT	20/11/2011 – 00:15	645
			D2_C6_bbc	EN	21/11/2011 – 23:26 (GMT)	
C7	Saúde	Proibição do fumo em NY	D1_C7_folha	PT	24/05/2011 – 13:38	887
			D2_C7_bbc	EN	24/05/2011 – 18:36 (HKT)	
C8	Mundo	Terremoto na Nova Zelândia	D1_C8_folha	PT	05/03/2011 – 05:01	948
			D2_C8_bbc	EN	03/03/2011 – 04:45 (GMT)	
C9	Mundo	Terremoto em Missouri	D1_C9_folha	PT	23/05/2011 – 08:04	1.169
			D2_C9_bbc	EN	23/05/2011 – 20:21 (GMT)	
C10	Mundo	Erupção vulcânica na Islândia	D1_C10_folha	PT	24/05/2011 – 12:13	1.476
			D2_C10_bbc	EN	24/05/2011 – 15:51 (GMT)	
C11	Ciência	Patentes genes humanos	D1_C11_bbc	PT	13/07/2013 – 16:34 (GMT)	963
			D2_C11_folha	EN	13/06/2013 – 23:50	
C12	Poder	Protestos: transporte	D1_C12_folha	PT	14/06/2013 – 07:25	808
			D2_C12_bbc	EN	14/06/2013 – 12:43 (GMT)	
C13	Mundo	Eleições do Irã	D1_C13_folha	PT	15/06/2013 – 17:57	1.266
			D2_C13_bbc	EN	16/06/2013 – 08:38 (GMT)	
C14	Saúde	Epidemia de dengue no MS	D1_C14_folha	PT	11/01/2013 – 19:03	534
			D2_C14_bbc	EN	21/01/2013 – 00:21 (GMT)	
C15	Saúde	Mastectomia preventiva	D1_C15_folha	PT	15/05/2013 – 03:01	1.367
			D1_C15_bbc	EN	14/05/2013 – 17:02 (GMT)	
C16	Ciência	Missão espacial chinesa	D1_C16_folha	PT	11/06/2013 – 21:06	793
			D2_C16_bbc	EN	11/06/2013 – 9:38 (GMT)	
C17	Poder	Protesto: copa das confederações	D1_C17_folha	PT	15/06/2013 – 14:53	918
			D2_C17_bbc	EN	16/06/2013 – 13:19 (GMT)	
C18	Ciência	Viagra feminino	D1_C18_folha	PT	16/06/2013 – 03:30	975
			D2_C18_bbc	EN	17/11/2009 – 9:35 (GMT)	
C19	Entrete-nimento	Lançamento: homem de aço	D1_C19_folha	PT	16/06/2013 – 13:24	898
			D2_C19_bbc	EN	11/06/2013 – 10:17 (GMT)	
C20	Mundo	Conflito na Turquia	D1_C20_folha	PT	17/06/2013 – 09h44	963
			D2_C20_bbc	EN	17/06/2013 – 13:00 (GMT)	
Total de palavras						**19.984**

Fonte: Tosta (2014)

No Quadro 4 são descritas as coleções do CM2News. Ao todo, o *corpus* é composto por 20 coleções, sendo que cada coleção conta com (i) um texto em português, (ii) um texto em inglês e (iii) um sumário humano, multilíngue, elaborado manualmente com base nesses dois textos (TOSTA, 2014).

O processo de elaboração dos sumários humanos, descrito em Tosta (2014), envolveu 13 linguistas computacionais integrantes da equipe de SA do NILC. O grupo foi orientado a produzir sumários abstrativos com tamanho equivalente a 30% (medido em número de palavras) do maior texto da coleção. Os participantes também foram orientados que o objetivo era produzir sumários informativos, isto é, trazendo as informações principais dos textos-fonte, e genéricos, ou seja, sem vistas a um tipo específico de leitor (MANI, 2001). Cada elemento do grupo recebeu 1 ou 2 coleções do *corpus* CM2News, cada qual consistindo de 1 texto-fonte em inglês e 1 em português. Como resultado da atividade, um sumário de referência foi elaborado para cada uma das 20 coleções do *corpus* CM2News.

A seção a seguir descreve o processo de representação conceitual dos textos do *corpus* CM2News.

3.2 Escolha do modelo de Representação de Conhecimento

Dentre as alternativas para representação do significado de textos multilíngues com vistas à SA, duas se destacam: a iRep4 (LENCI et al., 2002) e a UNL (UCHIDA et al., 1999). Ambas as representações foram apresentadas na Seção 2.2.

A UNL conta com um portal na internet[24] que disponibiliza uma série de informações sobre o sistema, desde textos introdutórios até treinamentos em diversos níveis (básico a avançado), todos eles gratuitos. A UNL conta ainda com recursos para o português, como um dicionário com cerca de 35.000 entradas (UNDL FOUNDATION, 2014a), e acesso aberto (mediante cadastro gratuito) às ferramentas e recursos já desenvolvidos. Por esses motivos, a UNL foi selecionada como a interlíngua a ser investigada para a SAMM neste projeto.

3.3 Ferramenta para UNLização

No início deste projeto, quatro ferramentas de UNLização (já apresentadas na seção 2.2.1.2) encontravam-se disponíveis:

[24] http://www.unlweb.net/unlweb/

- Enco
- SEAN
- IAN
- UNL Editor

A ferramenta Enco foi descartada, pois, das quatro, é a única cuja utilização não é gratuita. Para os propósitos desta investigação, pareceu mais interessante trabalhar com ferramentas de UNLização profunda, visando explorar ao máximo a estrutura semântica do documento. Logo, o analisador SEAN, por ser superficial, também foi descartado, restando as ferramentas IAN e UNL Editor.

O sistema IAN é classificado como uma ferramenta automática ou semiautomática, enquanto que o UNL Editor é totalmente manual. Assim, inicialmente houve preferência pelo sistema IAN.

Apesar da escolha inicial, surgiram dificuldades nas tentativas de utilizar a ferramenta IAN. Esse sistema deve ser parametrizado de acordo com a língua-fonte sendo representada. É necessário fornecer ao sistema: (i) o texto a ser UNLizado; (ii) um dicionário de análise e (iii) gramáticas de transformação e de desambiguação[25], sendo que o dicionário e as gramáticas devem estar de acordo com as especificações da UNL (UNDL FOUNDATION, 2013d).

No ambiente UNLarium[26], existem dicionários de análise disponíveis para vários idiomas, dentre eles o português e o inglês. Entretanto, quando consultado, o ambiente UNLarium não dispunha de gramáticas para o português ou para o inglês que pudessem ser utilizadas no sistema IAN visando à UNLização. Como consequência disso, não foi possível utilizar o sistema IAN, visto que ele não funciona sem tais gramáticas. Essas gramáticas encontravam-se em processo de revisão e não estariam disponíveis ao longo deste projeto[27].

Em outras palavras, durante esta pesquisa, não havia nenhuma ferramenta automática ou semiautomática para UNLização profunda em funcionamento à disposição. A única ferramenta de UNLização profunda que estava disponível era o UNL Editor – um sistema manual de representação em UNL.

Analisou-se ainda a possibilidade da elaboração de gramáticas para utilização da ferramenta IAN. Entretanto, o processo de criação de gramáticas em UNL mostrou-se

[25] No sistema UNL, uma gramática é um conjunto de regras utilizado para gerar representações em UNL a partir de textos em línguas naturais e para gerar textos em línguas naturais a partir da UNL (UNDL FOUNDATION, 2014b).
[26] http://www.unlweb.net/unlarium/
[27] Martins, R. T. (pesquisador na *UNDL Foundation*) (comunicação pessoal, 2013).

bastante complexo. A tarefa demandaria muito tempo, não só para um treinamento mais aprofundado em UNL, como também para o processo de criação das gramáticas em si.

Consideramos também o histórico de pesquisas envolvendo a UNL: já houve um projeto em que a elaboração de uma gramática de UNLização gerou resultados ainda distantes do desejado, mesmo envolvendo especialistas em UNL (MARTINS, R. T. et al., 2004).

Por esses motivos, inserir uma etapa de desenvolvimento de gramáticas UNL no projeto pareceria algo um tanto quanto ambicioso, havendo ainda a possibilidade de que essa empreitada não gerasse os resultados esperados e colocasse em risco o cronograma da pesquisa.

3.4 Escolha do método de UNLização

Durante a realização deste trabalho, não havia nenhum sistema de UNLização automática ou semiautomática completo (isto é, com gramáticas à disposição) que pudesse ser utilizado para UNLizar o *corpus* CM2News. Ademais, não havia nenhum *corpus* já UNLizado que atendesse aos requisitos necessários para o desenvolvimento desta pesquisa. Assim sendo, a alternativa que restou foi a UNLização manual do *corpus* CM2News.

Há relatos de UNLização manual bem sucedida de outros *corpora* (p. ex.: MARTINS, R. T., 2007, 2012). O risco maior decorrente dessa opção seria o de um possível atraso no cronograma de pesquisa, visto que a UNLização manual é uma tarefa um tanto quanto laboriosa.

Analisou-se ainda a possibilidade de envolver mais pessoas no projeto para auxiliar na etapa de UNLização. Entretanto, dado o curto período disponível para a pesquisa, em especial para a etapa de UNLização (inicialmente prevista em 3 meses), somente o período necessário para treinamento do pessoal já seria o bastante para gerar atrasos ainda maiores.

Considerando a experiência do próprio autor desta dissertação, foram necessários cerca de 4 meses para uma familiarização razoável com o projeto UNL, dispendidos com leituras sobre UNL e treinamentos para obtenção dos certificados CLEA250[28], CLEA500, CUP250[29] e CUP500, básicos para quem deseja uma noção mais abrangente sobre o funcionamento da UNL. Mesmo que houvesse um ganho de tempo posteriormente, durante a

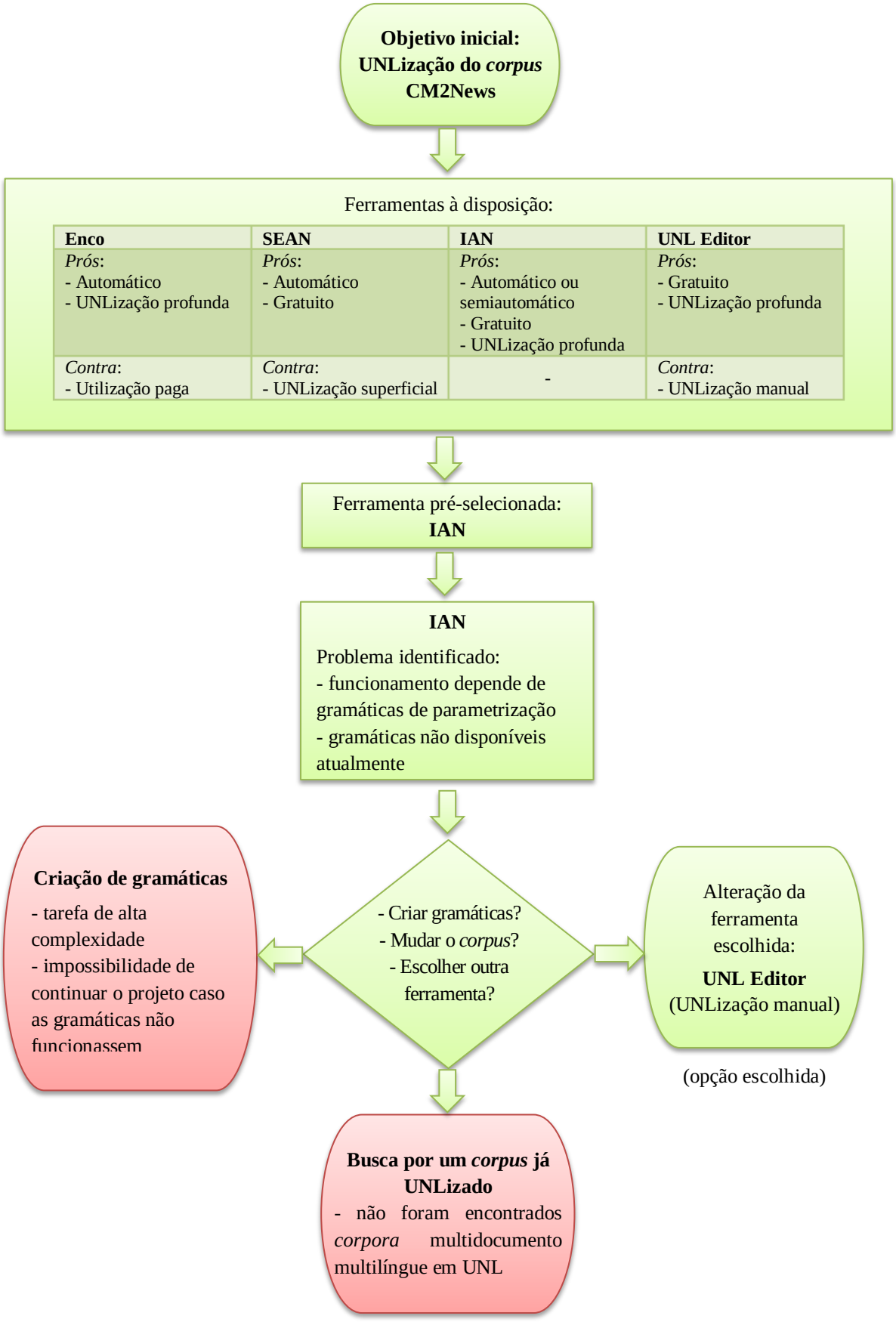
[28] CLEA = *Certificate of Language Engineering Aptitude*
[29] CUP = *Certificate of Proficiency in UNL*

etapa de UNLização, o tempo necessário para capacitar mais pessoas para realizar a tarefa não se mostrou compatível com nosso cronograma.

Assim sendo, a UNLização do *corpus* teve início sem o envolvimento de mais pessoas, sendo realizada integralmente pelo autor da dissertação. Decidiu-se que, caso necessário, a tarefa seria suspensa ao final do período de 3 meses, mesmo que não estivesse concluída. Ao final desse período, ainda que o *corpus* CM2News não tivesse sido completamente UNLizado, seria possível obter um fragmento do *corpus* suficiente para a realização de análises importantes e gerar discussões profícuas.

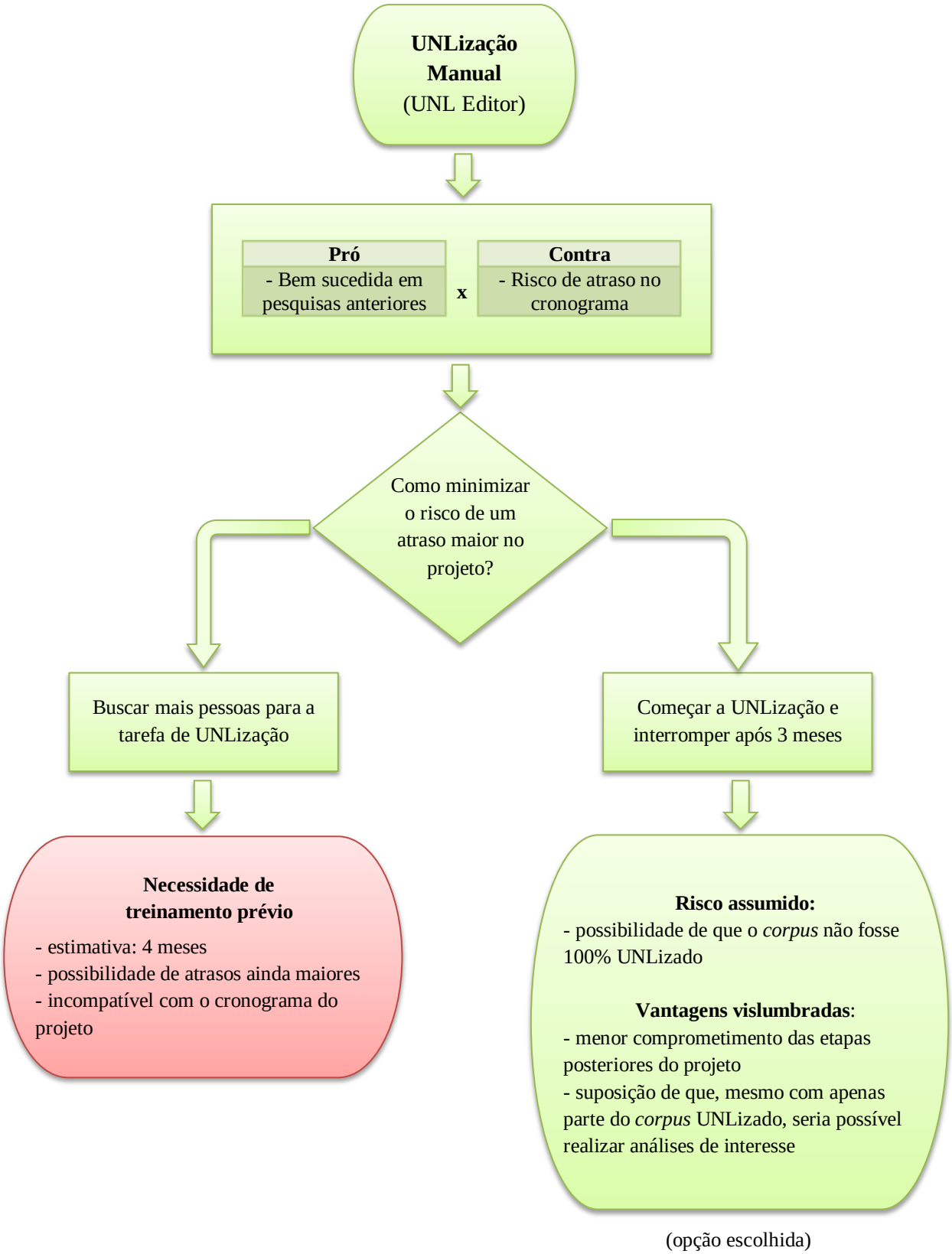
As Figuras 4 e 5 ilustram o processo de tomada de decisões relativas aos desafios iniciais do projeto.

Figura 4 – Escolha da ferramenta de UNLização



Fonte: Elaborado pelo autor

Figura 5 – Planejamento da UNLização manual



Fonte: Elaborado pelo autor

3.5 Treinamento para UNLização manual

Para a UNLização manual do *corpus* CM2News, foi necessário um estudo mais aprofundado sobre o sistema UNL. Assim, inseriu-se no projeto uma etapa de treinamento sobre UNL, que consistiu em uma revisão bibliográfica mais pontual e na realização de cursos sobre o sistema UNL.

Os cursos foram realizados no ambiente *online* VALERIE[30] (*VirtuAl LEaRnIng Environment*), criado pela *UNDL Foundation*. Os certificados oferecidos nesse ambiente proporcionam conhecimentos para que os usuários possam trabalhar com ferramentas e sistemas baseados em UNL, como o UNL Editor, por exemplo. Há vários níveis de certificação, cada qual abrangendo não só explicações teóricas, mas também exercícios práticos.

Há duas séries de certificados disponíveis: CLEA (*Certificate of Language Engineering Aptitude*) e CUP (*Certificate of Proficiency in UNL*).

A série CLEA abrange atualmente 4 certificados: CLEA250, CLEA500, CLEA750 e CLEA1000. Para cada um desses certificados, exige-se que o participante complete 25 níveis e faça um total de 250 exercícios – 10 em cada nível. A correção dos exercícios é realizada automaticamente pelo sistema e, para que se consiga avançar para o nível seguinte, deve-se fornecer a resposta correta para todos os exercícios apresentados.

O certificado CLEA250 relaciona-se ao mapeamento de conceitos em UNL para as línguas naturais. O CLEA500 trata do processo inverso – o mapeamento de línguas naturais para UNL. O certificado CLEA750 trata da elaboração de módulos gramaticais morfológicos e o CLEA1000 aborda a criação de módulos gramaticais sintáticos e semânticos.

A série CUP, por sua vez, é voltada especificamente para a representação dos fenômenos de língua natural em UNL. Essa série divide-se em dois certificados: o primeiro – CUP250 – concentra-se no componente lexical da UNL, e o segundo, CUP500, aborda as relações UNL. Analogamente à série CLEA, cada certificado da série CUP também abrange 25 níveis e 250 exercícios.

A representação do *corpus* CM2News em UNL é uma tarefa de mapeamento de documentos em língua natural para a UNL. Por esse motivo, o tipo de conhecimento desejável pareceu mais relacionado ao conteúdo abordado nos certificados CLEA250, CLEA500, CUP250 e CUP500. Dentre os treinamentos disponíveis, foram deixados de lado os que se

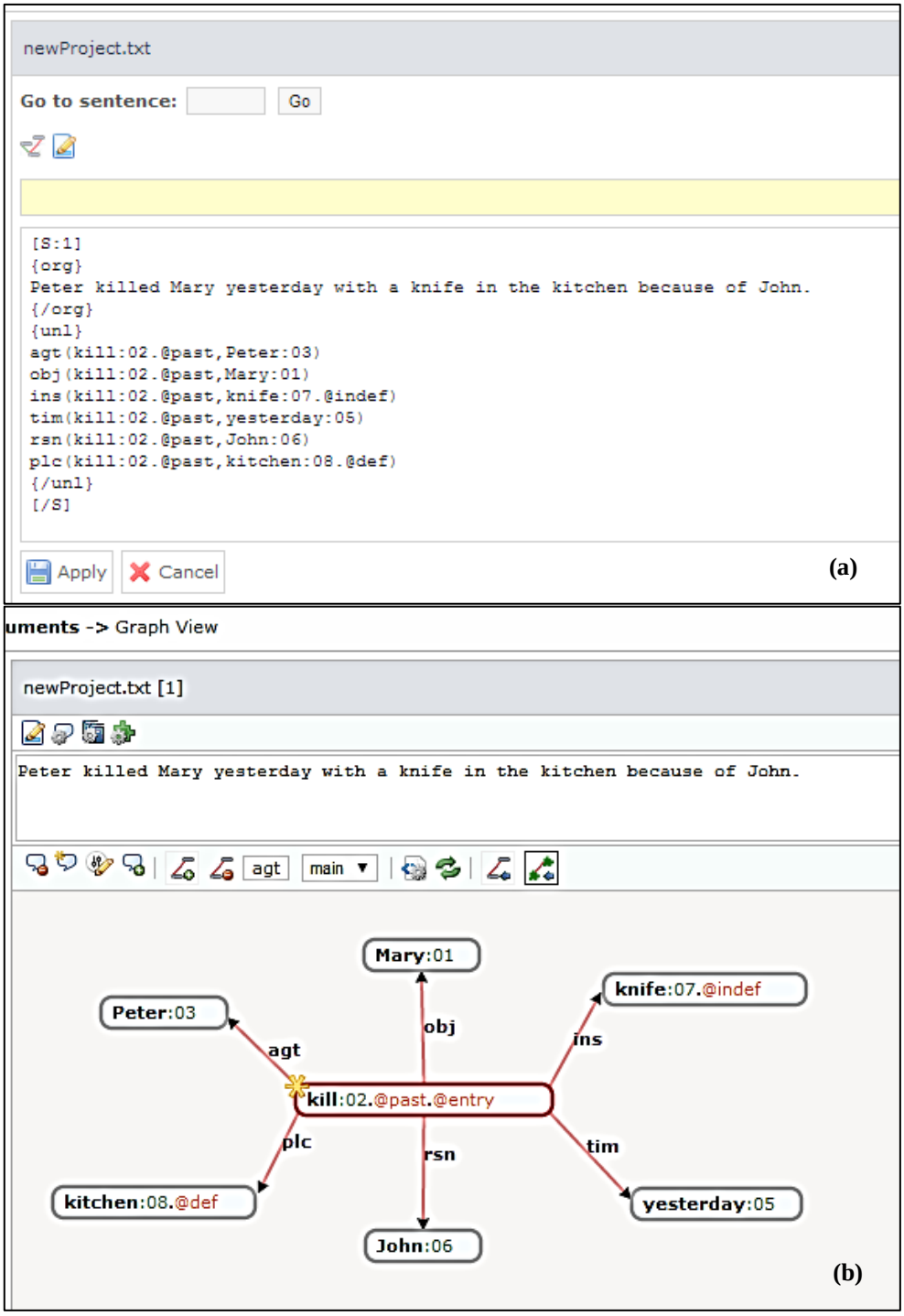
[30] http://www.unlweb.net/valerie/

concentravam na criação de gramáticas (CLEA750 e CLEA1000), visto que essa tarefa fugia ao escopo deste trabalho.

3.6 A ferramenta UNL Editor

O UNL Editor é uma ferramenta de anotação semântica criada pela *UNDL Foundation* em parceria com a *Bibliotheca Alexandrina*. Trata-se de um editor visual voltado para a análise de textos em línguas naturais e sua transformação na linguagem UNL. A interface do sistema permite a edição de dados tanto na forma textual quanto gráfica, fornecendo, assim, a possibilidade de visualização das redes semânticas que vão sendo criadas durante a UNLização (ALANSARY et al., 2011) (Figura 6).

Figura 6 – Interface da ferramenta UNL Editor: (a) edição textual e (b) edição gráfica

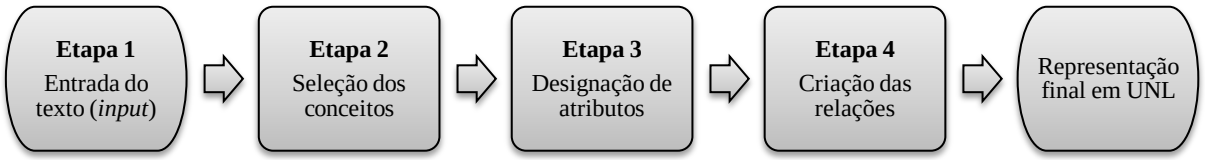


Fonte: *Print screen* da versão 1.1 do UNL Editor[31]

Para obtenção da representação em UNL, foram realizadas as seguintes etapas: (i) entrada do texto; (ii) identificação dos conceitos presentes nas sentenças; (iii) designação de atributos aos conceitos; e (iv) criação das relações semânticas entre os conceitos (Figura 7).

[31] http://dev.undlfoundation.org/unl_editor/login.jsp

Figura 7 – Etapas para representação no UNL Editor



Fonte: Adaptado de Alansary et al. (2011)

A Etapa 1 – entrada do texto – pode ser realizada manual ou automaticamente. Manualmente, pode-se optar por digitar o texto ou utilizar as operações de copiar e colar. Na opção automática, deve-se fornecer ao sistema um documento no formato txt. Como o *corpus* CM2News já estava disponível em formato txt, adotou-se a segunda opção.

Ao ser inserido no UNL Editor, cada texto passou por um processo automático de segmentação sentencial. Nos casos em que esse processo gerou sentenças divididas erroneamente, os resultados da segmentação foram corrigidos.

Com relação à identificação de cada sentença, cabe observar que tanto o formato utilizado quanto a numeração atribuída são gerados automaticamente pela ferramenta UNL Editor. Na etapa de segmentação automática, cada sentença recebe uma identificação sequencial na forma [S:n], onde "n" é um número, iniciando a partir do zero (S:0, S:1, S:2 e assim por diante). Entretanto, a segmentação sentencial automática ocasionalmente gera sentenças divididas de maneira inadequada. Sentenças com aspas, por exemplo, parecem ser um problema para o UNL Editor. Na coleção C2, texto-fonte em inglês, a sentença a seguir: *"I voted for her in the last elections," he said, "because I thought she would defend the rights of lesbian, gay and bisexual citizens."* foi dividida pelo UNL Editor da seguinte forma:

```
[S:11]
"I voted for her in the last elections," he said, "because I thought
she would defend the rights of lesbian, gay and bisexual citizens.

[S:12]
"
```

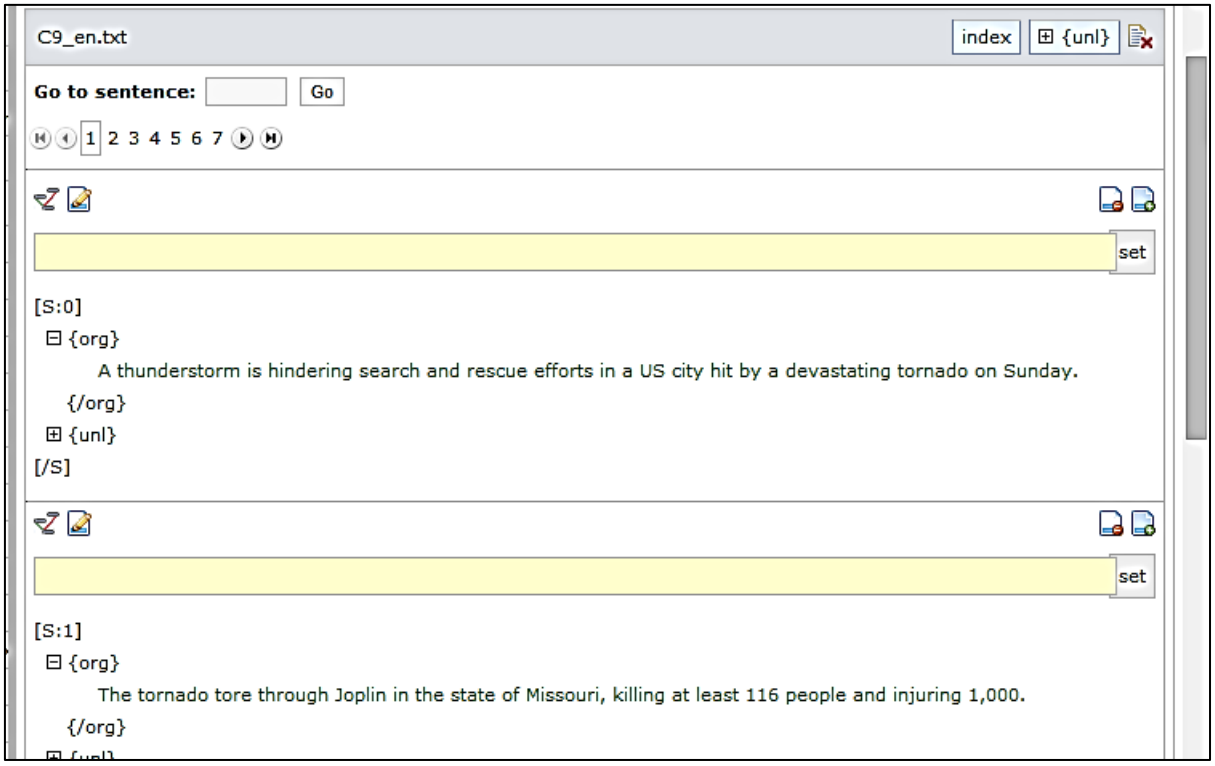
Em situações como essa, as sentenças foram manualmente editadas e corrigidas. No caso, a sentença [S:12] foi excluída e a sentença [S:11] foi complementada com as aspas que haviam sido realocadas, erroneamente, para a sentença subsequente. Entretanto, a sentença seguinte a essas duas, [S:13], continuou com a sua numeração original, visto que o programa não efetua

automaticamente a renumeração das sentenças no caso de exclusões. A consequência disso é que existem descontinuidades na numeração das sentenças: nesse caso em particular, há um salto da sentença [S:11] para a sentença [S:13].

Essas descontinuidades não parecem trazer nenhum problema em particular. Cabe apenas assinalar esse fato aqui, para que não cause estranheza a quem observar o *corpus* anotado, tabelas ou outros dados em que conste essa identificação sentencial.

O resultado dessa etapa é um texto dividido em sentenças pronto para receber a anotação no UNL Editor (Figura 8).

Figura 8 – Documento UNL após a segmentação sentencial (Etapa 1)



Fonte: *Print screen* da versão 1.1 do UNL Editor

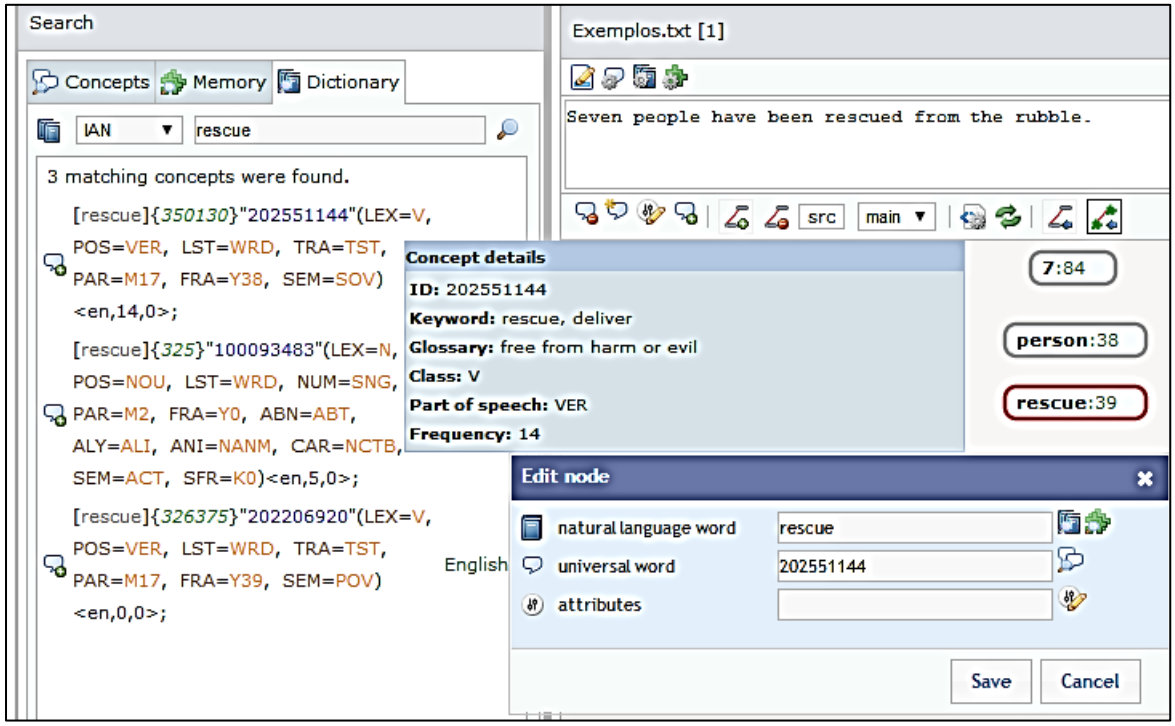
A seguir, descreve-se o processo de representação em UNL de uma das sentenças do *cluster* C9, mais especificamente do texto-fonte em inglês: *"Seven people have been rescued from the rubble."*

Os conceitos que fazem parte da UNL foram extraídos da WordNet 3.0[32] (FELLBAUM, 1998), sendo que cada conceito tem uma identificação numérica única (ID) no sistema. A identificação 202551144, por exemplo, corresponde ao conceito lexicalizado como "*rescue"* ou *"deliver*" em inglês e "livrar" ou "resgatar" em português. Para iniciar a

[32] Base relacional de dados semânticos da língua inglesa.

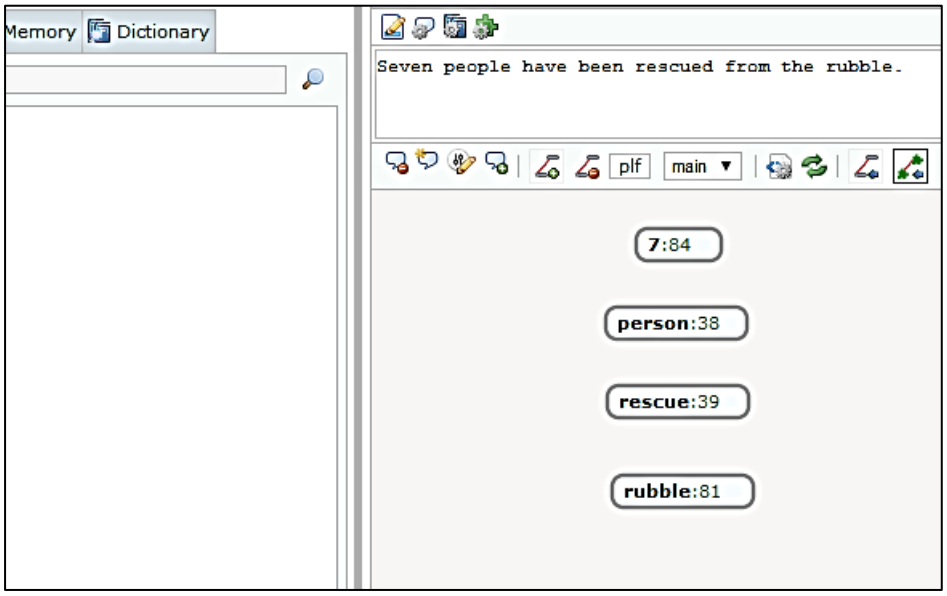
UNLização da sentença, foram selecionadas as IDs correspondentes aos conceitos nela identificados (Etapa 2). Isso é feito principalmente consultando-se os dicionários integrados ao UNL Editor (Figura 9). Obteve-se, então, uma série de conceitos, ainda sem os atributos e sem as relações que os interligam (Figura 10).

Figura 9 – Identificação dos conceitos nos dicionários UNL (Etapa 2)



Fonte: *Print screen* da versão 1.1 do UNL Editor

Figura 10 – Gráfico UNL resultante após a identificação dos conceitos



Fonte: *Print screen* da versão 1.1 do UNL Editor

Cada conceito recebe ainda uma identificação intrassentencial aleatória composta por dois dígitos (ex.: "**38**", em "person:38", na Figura 10). Em uma situação em que o conceito *person* fosse utilizado duas vezes na mesma sentença, as duas representações desse conceito receberiam identificações intrassentenciais distintas (ex.: "person:38" e "person:71"), o que é importante para que se possa saber a qual utilização do conceito *person* a representação se refere.

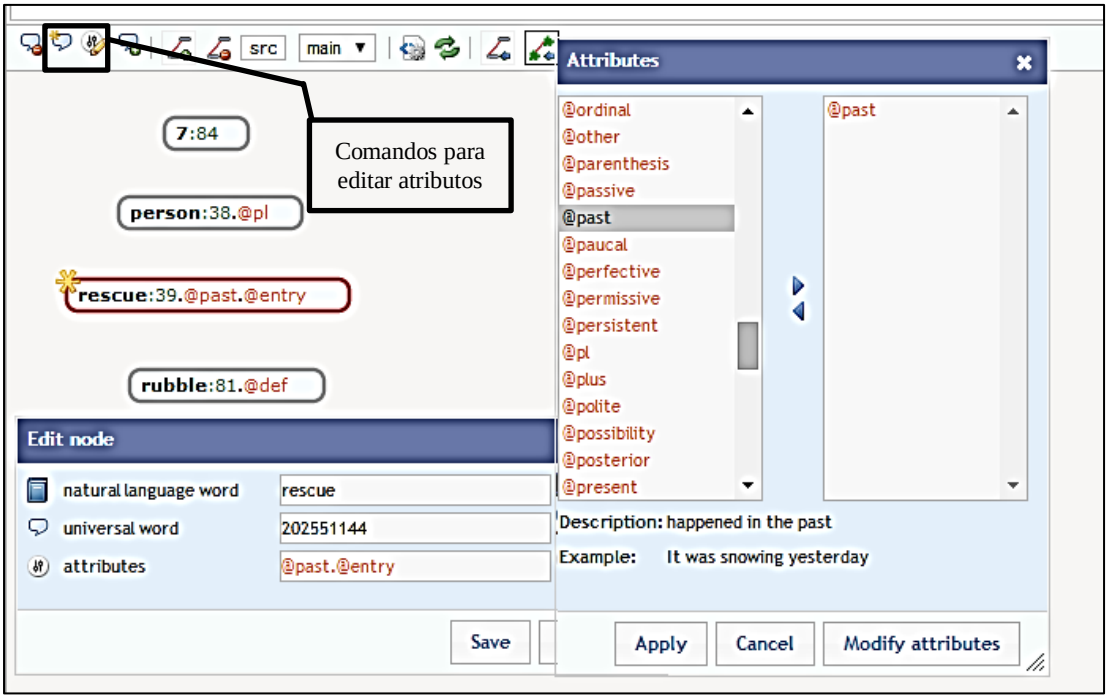
Na Etapa 3, esses conceitos foram complementados, com a designação de atributos a eles (Figuras 11 e 12). Além dos atributos que têm como função representar informações não exprimíveis por meio de UWs ou relações, existe um atributo, denominado "*entry*" (@entry), cuja função é servir de ponto de início para construção do grafo semântico. Esse atributo é necessário para todas as sentenças geradas no UNL Editor.

Por fim, na Etapa 4, foram criadas as relações semânticas entre os conceitos identificados (Figura 13), obtendo-se, assim, a representação final da sentença (Figuras 14 e 15). No exemplo dado, identificaram-se as seguintes relações:

- *qua (quantity)*: utilizada para expressar a quantidade de uma entidade. No caso, `7` é a quantidade de `person.pl` (pessoas);
- *obj (patient)*: um participante em uma ação ou processo sofrendo mudança de estado ou localidade. Essa relação ocorreu entre os conceitos `person.pl` e `rescue.past` (pessoas foram resgatadas);
- *src (initial state, place, origin or source)*: estado ou lugar inicial; origem ou fonte de uma entidade ou evento. No exemplo, `rubble` (escombros) indica o local no qual as pessoas foram resgatadas.

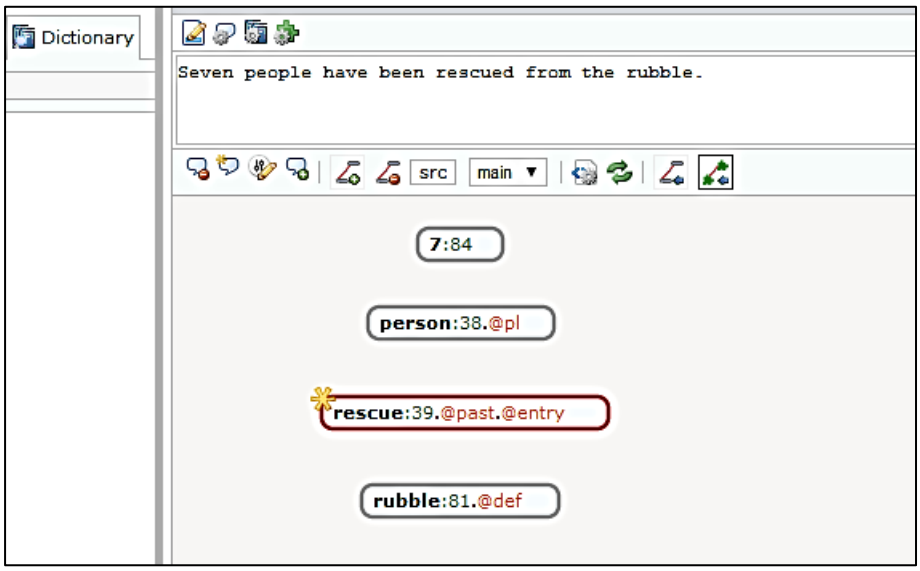
Ilustrações adicionais e procedimentos mais detalhados sobre como implementar cada uma dessas etapas constam em Alansary et al. (2011).

Figura 11 – Designação dos atributos no UNL Editor (Etapa 3)



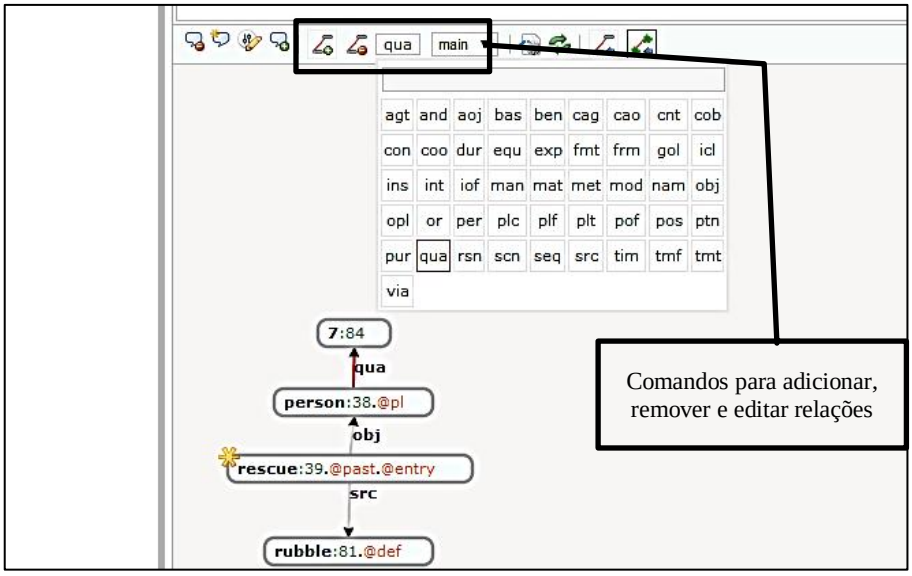
Fonte: *Print screen* da versão 1.1 do UNL Editor

Figura 12 – Representação gráfica UNL após a designação dos atributos



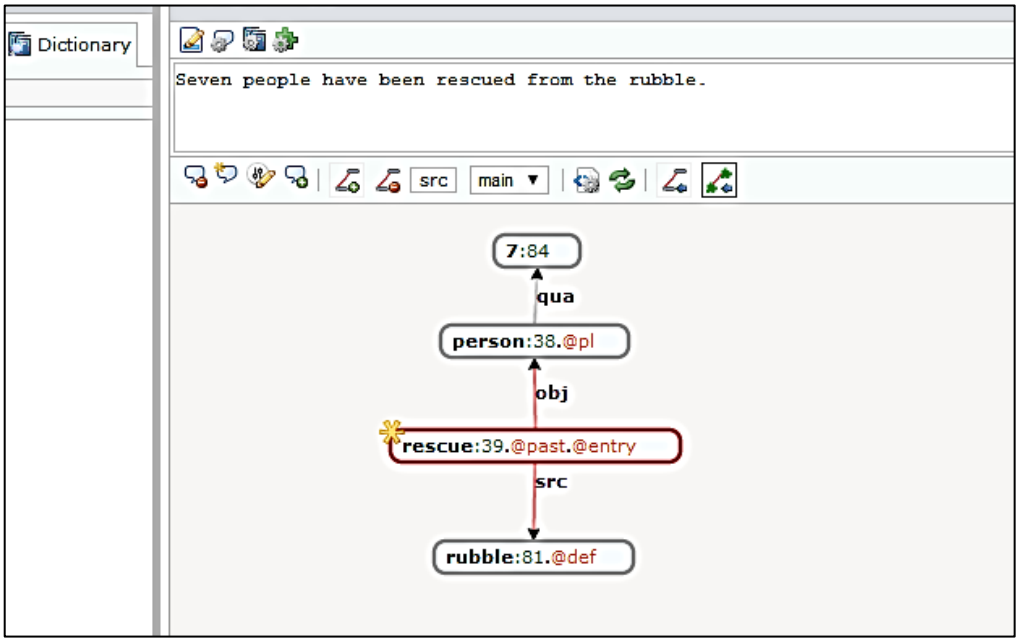
Fonte: *Print screen* da versão 1.1 do UNL Editor

Figura 13 – Criação das relações entre os conceitos no UNL Editor (Etapa 4)



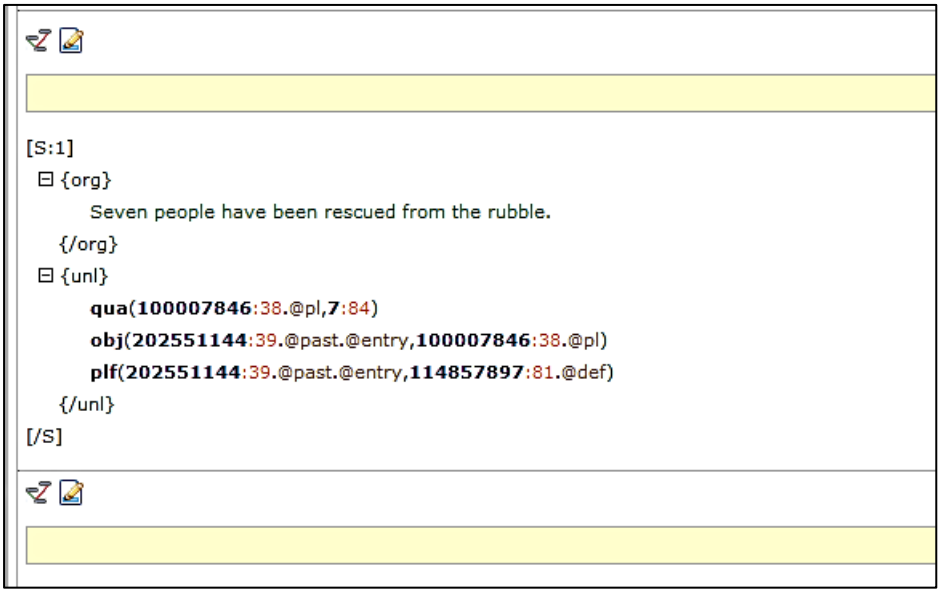
Fonte: *Print screen* da versão 1.1 do UNL Editor

Figura 14 – Representação final obtida – forma gráfica



Fonte: *Print screen* da versão 1.1 do UNL Editor

Figura 15 – Representação final obtida – forma textual



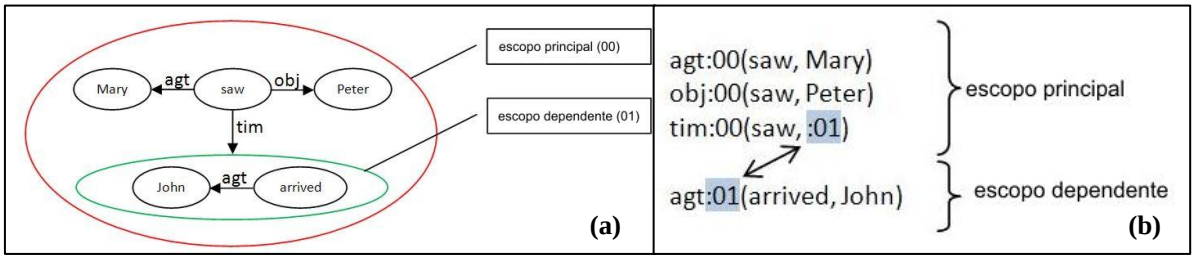
Fonte: *Print screen* da versão 1.1 do UNL Editor

Como parte da etapa de criação de relações semânticas, quando necessário, foram criadas as chamadas relações de escopo. Um escopo pode ser entendido como um grupo de relações entre os nós da rede semântica que funciona como uma entidade semântica única. Os escopos são subgrafos, ou hipernós, que correspondem aproximadamente à noção de orações subordinadas. A utilização de escopos é importante para evitar ambiguidade semântica em certos tipos de orações adverbiais, adjetivas e substantivas. (ALANSARY et al., 2011; UNDL FOUNDATION, 2013g). Essa, aliás, é uma das restrições do sistema de representação IRep4, comparado à UNL: no IRep4 não é possível representar escopos, o que, por conseguinte, acaba limitando o desempenho do sistema (LENCI et al., 2002).

Em cada sentença representada em UNL, esses subgrafos recebem uma identificação numérica de dois dígitos antecedida de dois pontos, começando por `:00`, que representa o nó principal e pode ser omitido.

A Figura 16 ilustra a representação da sentença "*Mary saw Peter when John arrived*" (UNDL FOUNDATION, 2013g). O segmento "*when John arrived*" serve de argumento em uma relação de tempo: "*when John arrived*" está em relação com o verbo "*saw*", representando o tempo em que essa ação ocorreu. Para que essa relação possa ser expressa adequadamente, deve-se recorrer ao uso de um subgrafo ou hipernó. Em (a), é possível visualizar o subgrafo (ou hipernó) formado por "*John arrived*" e, em (b), a representação escrita da sentença.

Figura 16 – Exemplo de uma representação em UNL envolvendo escopo



Fonte: *UNDL Foundation* (2013g)

3.7 Diretrizes para UNLização manual

A representação de documentos na linguagem UNL pode ser vista como um processo de anotação semântica. Nesse sentido, UNLizar um texto corresponde a identificar os conceitos, atributos e relações semânticas nele presentes. Para a representação de documentos em UNL, a principal referência existente são as próprias especificações UNL, que consistem em uma série de documentos que descrevem como as informações devem ser expressas. Além de determinarem como funcionam as UWs, as especificações UNL definem a lista de relações e atributos, bem como seus papéis na representação, provendo também um detalhamento da arquitetura do sistema (UNDL FOUNDATION, 2005, 2013e).

Como a representação em UNL não deve depender de conhecimentos implícitos, os textos-fonte, ao serem UNLizados, têm de passar por um processo de normalização. Assim como proposto na UNLização de *Le Petit Prince* (MARTINS, R. T., 2012), todas as valências semânticas foram saturadas, incluindo anáforas, elipses, pressuposições e implicaturas.

Sempre que possível, os pronomes foram substituídos por seus antecedentes no texto. Entretanto, frequentemente há situações em que isso não pode ser feito. Em casos assim, a recomendação nas especificações UNL (UNDL FOUNDATION, 2013f) é a de que se use a UW nula “00” e, havendo necessidade, que se faça uso de atributos. Isso geralmente acontece em situações envolvendo: (i) exóforas, isto é, ocorrências de expressões linguísticas cujos referentes são identificáveis apenas no contexto situacional; (ii) pronomes indefinidos, como “todos”, “nenhum” e “algo”, que fazem referência a categorias gerais de coisas ou pessoas; (iii) pronomes interrogativos, por exemplo “qual”, “quando” e “quem”, que se referem a constituintes não presentes na estrutura sintática; (iv) interjeições usadas separadamente; e (v) elipses não substituíveis por antecedentes. O Quadro 5 contém alguns exemplos de como se faz a representação em UNL em situações como essas.

Quadro 5 – Exemplos de representação em UNL envolvendo UW nula

Ocorrência	Exemplo	Representação em UNL
Pronomes pessoais sem antecedentes	1ª pessoa do singular	00.@1
	2ª pessoa do singular	00.@2
	3ª pessoa do singular	00.@3
	1ª pessoa do plural	00.@1.@pl
	2ª pessoa do plural	00.@2.@pl
	3ª pessoa do plural	00.@3.@pl
Pronomes indefinidos	"tudo"	00.@all
	"algo"	00.@any.@thing
	"todos"	00.@every.@person
Pronomes interrogativos[33]	"qual", "quando", "quem"	00.@wh
Interjeições isoladas	"Ei", na frase "Ei! Olhe!"	look.@attention

Fonte: Adaptado de *UNDL Foundation* (2013e)

Como diretrizes principais para o processo de UNLização, adotaram-se as especificações UNL2005 e UNL2010 (UNDL FOUNDATION, 2005, 2013e). As especificações 2010 ainda não são oficiais, constituindo uma extensão e correção das especificações 2005. Dentre as principais diferenças entre a UNL2005 e a UNL2010, destaca-se o conjunto de relações semânticas, composto por 46 relações em 2005 e 38 relações em 2010 (Quadro 6). Além de algumas relações terem sido excluídas e outras acrescentadas, a definição das relações que foram preservadas ocasionalmente apresenta dissimilaridades. Em outras palavras, o conjunto de relações e as indicações de como utilizá-las variam um pouco de uma especificação para outra.

O sistema usado para UNLização manual – o UNL Editor (v. 1.1) – ainda opera com o conjunto de relações definidas nas especificações de 2005. Assim, nem sempre foi possível basear-se nas especificações mais recentes, de 2010, visto que o UNL Editor não está plenamente integrado a essas novas especificações.

Iniciada a tarefa de UNLização, constatou-se que nem sempre as especificações UNL foram suficientes para se decidir como gerar a representação. Por vezes, as sentenças sendo UNLizadas não guardavam semelhança com nenhum dos exemplos apresentados nas especificações de 2010 ou mesmo as de 2005, sendo difícil optar por quais relações e atributos utilizar durante a UNLização.

[33] A diferença entre os pronomes interrogativos (ex.: "qual", "quem" e "onde") é especificada pela relação UNL em que ela aparece. Por exemplo: se "00.@wh" aparece em uma relação de *agente* (agt), a representação é interpretada como "quem"; se aparece em uma relação de *lugar* (plc), é interpretada como "onde".

Quadro 6 – Relações semânticas nas especificações UNL2005 e UNL2010

Relação	Definição	2005	2010
agt	agent	☑	☑
and	conjunction	☑	☑
ant	antonym		☑
aoj	attributive object	☑	☑
bas	basis	☑	
ben	beneficiary	☑	☑
cag	co-agent	☑	
cao	co-thing with attribute	☑	
cnt	content	☑	☑
cob	co-object	☑	
con	condition	☑	☑
coo	co-occurrence	☑	
dur	duration	☑	☑
equ	synonym	☑	☑
exp	experiencer		☑
fld	semantic field		☑
fmt	from-to	☑	
frm	origin	☑	
gol	goal	☑	☑
icl	inclusion	☑	☑
ins	instrument	☑	☑
int	intersection	☑	
iof	an instance of	☑	☑
lpl	logical place		☑
man	manner	☑	☑
mat	material		☑

Relação	Definição	2005	2010
met	method	☑	
mod	modifier	☑	☑
nam	name	☑	☑
obj	object / patient	☑	☑
opl	objective place	☑	☑
or	disjunction	☑	☑
per	unit to measure object	☑	☑
plc	place	☑	☑
plf	initial place	☑	
plt	final place	☑	
pof	part-of	☑	☑
pos	possessor	☑	☑
ptn	partner	☑	☑
pur	purpose	☑	☑
qua	quantity	☑	☑
res	theme		☑
rsn	reason	☑	☑
scn	scene	☑	
seq	sequential order	☑	☑
shd	sentence head	☑	
src	source	☑	☑
tim	time	☑	☑
tmf	time-from	☑	☑
tmt	time-to	☑	☑
to	destination	☑	
via	intermediate place	☑	☑

Fonte: UNDL Foundation (2005, 2013e)

Em vista disso, usou-se como referência adicional para UNLização um conjunto de anotações pessoais feitas com base nos treinamentos realizados no ambiente VALERIE, especialmente observações feitas durante a obtenção dos certificados CUP250 e CUP500.

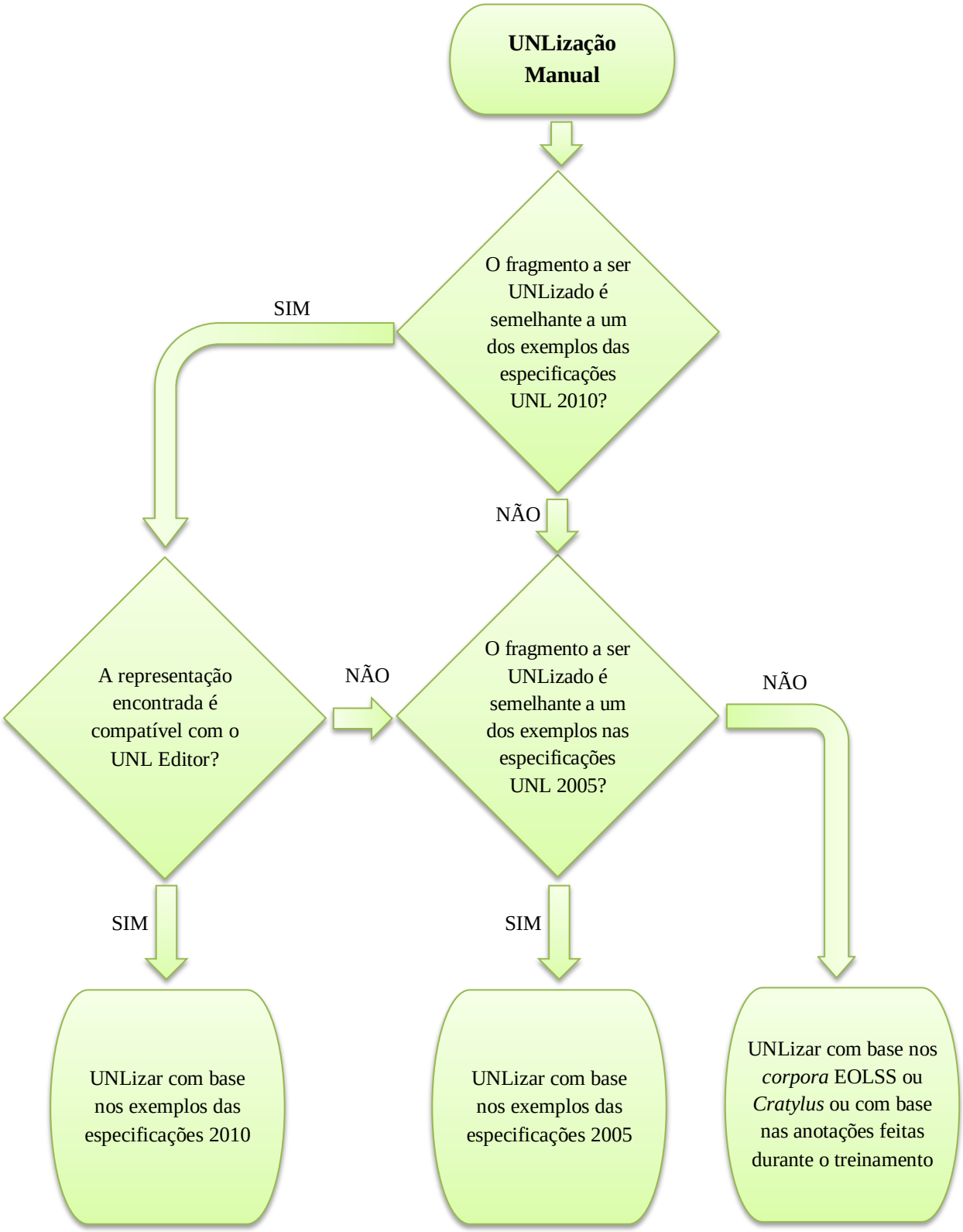
Ademais, foram também consultados dois *corpora* em UNL para auxiliar no processo de UNLização: o EOLSS (UNDL FOUNDATION, 2009, 2010) e o *Cratylus* (MARTINS, R. T. 2007; UNDL FOUNDATION, 2012a). Esses *corpora* foram selecionados por serem recentes e por terem sido construídos a partir do inglês, um dos idiomas utilizados nos documentos-fonte do *corpus* CM2News. Um *corpus* como o *Le Petit Prince* (MARTINS, R. T. 2012), por exemplo, apesar de recente, aparentemente não seria tão útil como referência para UNLização do *corpus* CM2News, visto que ele foi construído a partir de um texto-fonte em francês.

Por fim, o conjunto de diretrizes para UNLização consistiu em usar os recursos supramencionados da seguinte forma:

1) Caso o trecho a ser UNLizado fosse semelhante aos exemplos mencionados nas especificações de 2010 e caso a representação gerada fosse implementável na ferramenta UNL Editor, a representação foi feita com base nesses exemplos apresentados nas especificações de 2010.
2) Caso o segmento sendo UNLizado não fosse semelhante a nenhum dos exemplos citados nas especificações de 2010 ou caso a representação gerada fosse incompatível com o sistema UNL Editor, foram consultadas as especificações de 2005. Havendo, nessas últimas, algum exemplo semelhante ao trecho sendo UNLizado, a representação foi feita conforme esse exemplo.
3) Nos casos em que as estratégias (1) e (2) não tenham sido bem sucedidas, foram consultados os *corpora* EOLSS e *Cratylus* e as anotações pessoais feitas durante o treinamento, visando à busca de trechos similares àqueles sendo UNLizados.

A Figura 17 ilustra as diretrizes utilizadas para UNLização manual do *corpus* CM2News. Na sequência, os Quadros 7 a 11 exemplificam a aplicação dessas diretrizes a algumas sentenças do *corpus*.

Figura 17 – Diretrizes utilizadas para UNLização manual



Fonte: Elaborado pelo autor

Quadro 7 – Diretrizes utilizadas para UNLização manual – Exemplo 1

Estratégia utilizada	1) UNLizar com base em exemplo das especificações de 2010	
Fragmento a ser UNLizado	*Seven people*	
Sentença original / Contexto	*Seven people have been rescued from the rubble*.	
Especificações aplicáveis – UNL2010	**Relações**[34]	
	Relação	qua = *quantity* (quantidade)
	Definição da relação	Utilizada para expressar a quantidade de uma entidade
	Exemplo dado	*two books* = qua(book;2)
	Atributos	
	Atributo	@pl[35] (*plural*)
	Exemplo dado	*books* = book.@pl
Representação do fragmento, obtida com base nos exemplos mencionados na UNL2010	qua(person.@pl;7)	
A representação é implementável no UNL Editor?	Sim	
Representação final do fragmento	qua(person.@pl;7)	

Fonte: Elaborado pelo autor

[34] http://www.unlweb.net/wiki/Universal_Relations

[35] http://www.unlweb.net/wiki/Quantifier

Quadro 8 – Diretrizes utilizadas para UNLização manual – Exemplo 2

<table>
<tr><td>Estratégia utilizada</td><td colspan="2">2) UNLizar com base em exemplo das especificações de 2005</td></tr>
<tr><td>Fragmento a ser UNLizado</td><td colspan="2">In the incidents</td></tr>
<tr><td>Sentença original / Contexto</td><td colspan="2">In the incidents in central London, police arrested two boys aged 17 from Notting Hill and Belgravia on suspicion of burglary.</td></tr>
<tr><td rowspan="9">Especificações aplicáveis – UNL2010</td><td colspan="2">Relações</td></tr>
<tr><td><u>Relação</u></td><td>lpl = logical place (lugar lógico)</td></tr>
<tr><td><u>Definição da relação</u></td><td>Um lugar não físico onde uma entidade ou evento ocorre ou um estado existe</td></tr>
<tr><td rowspan="2"><u>Exemplos dados</u></td><td>John works in politics = lpl(works;politics)</td></tr>
<tr><td>John is in love = lpl(John;love)</td></tr>
<tr><td colspan="2">Atributos</td></tr>
<tr><td><u>Atributo</u></td><td><u>Exemplo dado</u></td></tr>
<tr><td>@pl (plural)</td><td>books = book.@pl</td></tr>
<tr><td>@def[36] (definite)</td><td>the book = book.@def</td></tr>
<tr><td>@past[37]</td><td>He spoke = speak.@past</td></tr>
<tr><td>Representação do fragmento, obtida com base nos exemplos mencionados na UNL2010</td><td colspan="2">lpl(arrest.@past,incident.@pl.@def)</td></tr>
<tr><td>A representação é implementável no UNL Editor?</td><td colspan="2">Não. A relação “lpl” ainda não existe no UNL Editor.</td></tr>
<tr><td>Solução encontrada</td><td colspan="2">Consulta às especificações de 2005</td></tr>
<tr><td rowspan="4">Especificações aplicáveis – UNL2005</td><td colspan="2">Relações</td></tr>
<tr><td><u>Relação</u></td><td>scn = scene (cenário)</td></tr>
<tr><td><u>Definição da relação</u></td><td>Indica um cenário onde um evento ocorre, ou um estado é verdadeiro, ou algo existe</td></tr>
<tr><td><u>Exemplo dado</u></td><td>... win (a prize) in a contest = scn (win,contest)</td></tr>
<tr><td>Representação final do fragmento, obtida com apoio dos exemplos mencionados na UNL2005</td><td colspan="2">scn(arrest.@past,incident.@pl.@def)</td></tr>
</table>

Fonte: Elaborado pelo autor

[36] http://www.unlweb.net/wiki/Specification

[37] http://www.unlweb.net/wiki/Time e http://www.unlweb.net/wiki/English_verbs

Quadro 9 – Diretrizes utilizadas para UNLização manual – Exemplo 3

Estratégia utilizada	2) UNLizar com base em exemplo das especificações de 2005	
Fragmento a ser UNLizado	*The main part*	
Sentença original / Contexto	*The main part of the building was razed on Thursday morning*.	
Resultado da consulta às especificações aplicáveis – UNL2010	O segmento a ser UNLizado não é semelhante a nenhum dos exemplos citados nas especificações de 2010	
Solução encontrada	Consulta às especificações de 2005	
Especificações aplicáveis – UNL2005	**Relações**	
	Relação	mod (*modification*)
	Definição da relação	Indica uma coisa que restringe uma coisa enfocada.
	Exemplo dado	*the main part* = mod(part,main)
	Atributos	
	Atributo	Exemplo dado
	@def (*definite*)	*the book* = book.@def
Representação final do fragmento, obtida com base nos exemplos mencionados na UNL2005	mod(part.@def,main)	

Fonte: Elaborado pelo autor

Quadro 10 – Diretrizes utilizadas para UNLização manual – Exemplo 4

Estratégia utilizada	3) UNLizar com base em exemplo de outros *corpora* em UNL ou em anotações feitas durante o treinamento	
Fragmento a ser UNLizado	*Roughly (...) damaged*	
Sentença original / Contexto	*Roughly 2,000 buildings were damaged in the region*.	
Resultado da consulta às especificações aplicáveis – UNL2010 e UNL2005	Mesmo após consulta às especificações de 2010 e de 2005, houve dúvidas sobre como proceder com a UNLização do segmento.	
Solução encontrada	Consulta aos *corpora* EOLSS e *Cratylus*	
Trecho similar encontrado no *corpus* EOLSS	Sentença original	*It **extends roughly** from 7-8 km (...)*
	Representação do trecho similar	man(extend,roughly)
Representação final do fragmento, obtida com base no trecho similar localizado no *corpus* EOLSS	man(damage.@past,roughly)	

Fonte: Elaborado pelo autor

Quadro 11 – Diretrizes utilizadas para UNLização manual – Exemplo 5

Estratégia utilizada	3) UNLizar com base em exemplo de outros *corpora* em UNL ou em anotações feitas durante o treinamento	
Fragmento a ser UNLizado	*(...) send (...) condolences (...)*	
Sentença original / Contexto	*Michelle and I send our deepest condolences to the families of all those who lost their lives in the tornadoes and severe weather that struck Joplin, Missouri, as well as communities across the Midwest today, the president said.*	
Resultado da consulta às especificações aplicáveis – UNL2010 e UNL2005	Mesmo após consulta às especificações de 2010 e de 2005, houve dúvidas sobre como proceder com a UNLização do segmento.	
Solução encontrada	Consulta às anotações pessoais feitas durante o treinamento	
Trecho similar encontrado nas anotações pessoais realizadas durante o treinamento	Sentença	*The duke **gave** this **teapot** to my aunt.*
	Representação do trecho similar	cnt(give,teapot)
Representação final do fragmento, obtida com base no trecho similar localizado nas anotações pessoais	cnt(send,condolence.@pl)	

Fonte: Elaborado pelo autor

O processo de UNLização visou, portanto, sempre que possível, UNLizar de acordo com as especificações mais recentes, ou seja, UNL2010. Quando as especificações de 2010 não trouxeram um nível de detalhamento que permitisse decidir, com segurança, como gerar a representação, as especificações de 2005 foram examinadas. Fez-se o mesmo nas situações em que o sistema UNL Editor não permitiu utilizar as especificações de 2010. Adicionalmente, foram valorizadas as experiências de UNLização anteriores, com consultas frequentes aos *corpora* EOLSS e *Cratylus* nos casos de dúvidas remanescentes.

3.8 Da representação computacional a uma representação legível por humanos

Conforme pôde ser observado anteriormente na Figura 15, a representação obtida no UNL Editor é composta por algarismos que correspondem às identificações numéricas (IDs) dos conceitos. Apesar de computacionalmente vantajosa, essa forma de representação não permite uma leitura da representação por humanos. Para efeito de comparação, a Figura 18 apresenta uma sentença do *corpus* experimental UC-B1[38] representada de duas formas diferentes: em (a), mostra-se a representação tal qual ela é disponibilizada no *corpus* UC-B1; em (b), mostra-se a representação da mesma sentença, dessa vez gerada pelo sistema UNL Editor. Como é

[38] http://www.unlweb.net/wiki/UCB1

possível notar, no primeiro caso, a representação dos conceitos é feita usando palavras em língua natural (inglês) e, no segundo caso, usando as identificações numéricas dos conceitos.

Figura 18 – Sentença do *corpus* UC-B1 representada em UNL

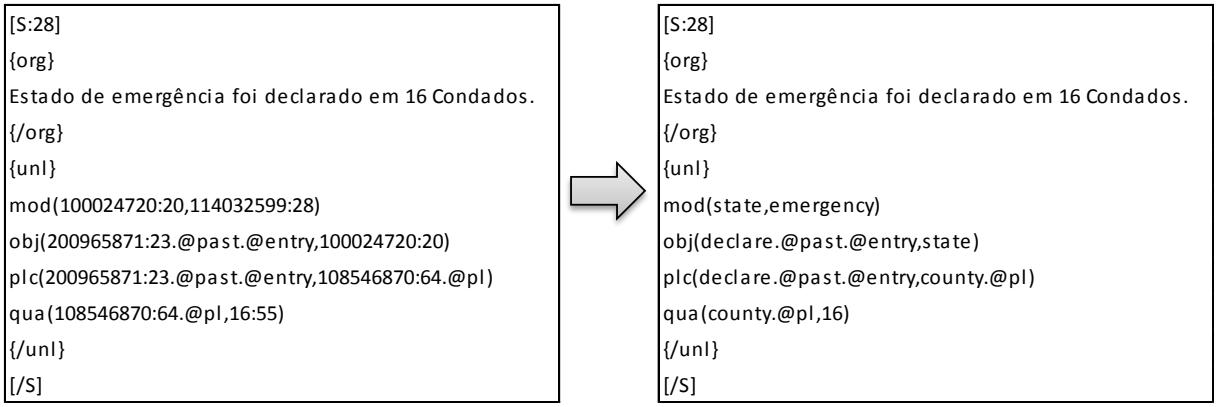
(a)	[S:12] {org:en} The Tortoise had already won the race. {/org} {unl} exp(win.@past.@perfective,tortoise.@def) obj(win.@past.@perfective,race.@def) tim(win.@past.@perfective,already) {/unl} [/S]
(b)	[S:12] {org} The Tortoise had already won the race. {/org} {unl} exp(201100145.@past.@perfective,101670092.@def) obj(201100145.@past.@perfective,107458453.@def) tim(201100145.@past.@perfective,400031798) {/unl} [/S]

Fonte: Elaborado pelo autor

Uma das etapas deste projeto previa o alinhamento entre as sentenças do sumário e suas sentenças correspondentes nos textos-fonte, além do alinhamento entre as representações dessas sentenças. Com isso, objetivava-se identificar estratégias de sumarização que se tornassem salientes na representação em UNL. Entretanto, trabalhar com a representação dos conceitos por IDs dificultaria consideravelmente a visualização e interpretação das representações em UNL. Assim sendo, converteu-se a representação exemplificada na Figura 18b, que é a representação gerada pelo UNL Editor, para uma representação como a mostrada na Figura 18a.

Tendo em vista que não há uma ferramenta para que essa tarefa seja feita de maneira automática, a transformação para uma representação que empregue conceitos em língua natural, no lugar das IDs numéricas, foi feita manualmente, com apoio dos dicionários disponíveis no sistema UNL Editor. A Figura 19 ilustra o resultado desse processo para uma das sentenças do *cluster* 9 do *corpus* CM2News.

Figura 19 – Transformação da representação computacional para uma representação legível por humanos



Fonte: Elaborado pelo autor

3.9 Resultados da UNLização manual

Ao final do período dedicado à tarefa de UNLização, foram representados em UNL um total de 9 textos do *corpus* CM2News. As características desses textos são mostradas no Quadro 12.

Quadro 12 – Textos do *corpus* CM2News representados em UNL

Coleção	Domínio	Assunto / Tema	Documento	Língua	Tipo de Documento	Nº de palavras
C1	Mundo	Ataques a Londres	C1-PT	Português	Texto-fonte	518
			C1-EN	Inglês	Texto-fonte	788
			C1-Sum-ref	Português	Sumário de referência	229
C2	Poder	Kit gay	C2-PT	Português	Texto-fonte	287
			C2-EN	Inglês	Texto-fonte	229
			C2-Sum-ref	Português	Sumário de referência	84
C9	Mundo	Terremoto em Missouri	C9-PT	Português	Texto-fonte	511
			C9-EN	Inglês	Texto-fonte	660
			C9-Sum-ref	Português	Sumário de referência	198
					Total	3.504

Fonte: Adaptado de Tosta (2014)

Para que a etapa de alinhamento entre sentenças pudesse ser realizada, foram UNLizados *clusters* inteiros, ou seja, conjuntos compostos por 1 texto em português, 1 texto em inglês e o sumário multidocumento correspondente a esses textos. A UNLização começou sequencialmente, com a representação dos *clusters* C1 e C2. Durante a UNLização da

segunda coleção, ficou evidente que a representação total do *corpus* não seria possível, considerando-se o tempo disponível. Tendo-se em vista que o *cluster* C1 está entre os três maiores *clusters* do *corpus* e o *cluster* C2 é o menor da coleção, optou-se por UNLizar mais um *cluster*, desta vez de tamanho médio, o que levou à UNLização do *cluster* C9. Assim, o resultado foi a UNLização de 3 *clusters*: um de tamanho grande (C1), um de tamanho médio (C9) e um de tamanho pequeno (C2).

Como já observado anteriormente, a expectativa inicial era UNLizar automaticamente o *corpus* todo, o que, entretanto, não se mostrou viável devido à ausência de ferramentas automáticas para a execução dessa tarefa. Ainda assim, a partir das representações em UNL desses 9 textos, foi possível discutir e levantar hipóteses sobre estratégias de seleção de conteúdo baseadas nessa interlíngua.

A etapa subsequente da investigação foi o alinhamento entre as sentenças dos sumários humanos e as sentenças dos textos-fonte e o alinhamento entre as suas respectivas representações conceituais.

4 ALINHAMENTO DOS TEXTOS-FONTE AOS SUMÁRIOS HUMANOS

> "*A riqueza inigualável de textos alinhados para um grande número de propósitos é clara para qualquer um ver.*"
> (KAY, 2000, p. xvii, tradução nossa)

Em PLN, a tarefa de alinhamento consiste em se relacionar elementos provenientes de textos diferentes, localizando pontos de correspondência. Pode-se realizar esse processo utilizando desde unidades menores, como o alinhamento por palavras, até unidades maiores, como sentenças, parágrafos, seções ou até mesmo documentos inteiros (CASELI et al., 2004; CASELI; NUNES, 2005; CAMARGO, 2013).

O alinhamento, ou indexação, tem se mostrado útil para uma série de aplicações na área de PLN, como a TA, sumarização, simplificação textual e sistemas de perguntas e respostas (SENO; NUNES, 2008).

Para a SA, especificamente, o alinhamento pode ser bastante profícuo: ao se alinhar um sumário aos seus textos-fonte, a procedência das informações presentes no sumário torna-se mais saliente. Essa comparação entre as informações do sumário e as dos textos-fonte pode revelar estratégias usadas para a elaboração do sumário, o que é um passo importante para a automatização do processo de sumarização.

A tarefa de alinhamento sentencial foi feita manualmente pelo autor da dissertação tomando-se por base a metodologia utilizada por Camargo (2013). Seguindo esses critérios, durante o alinhamento, buscou-se identificar sobreposição total ou parcial de conteúdo, e não apenas sobreposição lexical. O Quadro 13 ilustra duas sentenças alinhadas com base em sobreposição parcial de conteúdo – sentenças que não seriam alinhadas com base unicamente na sobreposição de palavras (*word overlap*). Os trechos em negrito destacam os segmentos em que há sobreposição de conteúdo. As regras gerais de alinhamento utilizadas são apresentadas no Quadro 14.

Quadro 13 – Exemplo de alinhamento com base na sobreposição de conteúdo

Sentença do sumário (SS)	Sentença do documento (SD)
Vários moradores e turistas nas regiões, inclusive brasileiros, foram retirados dos locais, enquanto outros estão **se preparando para a passagem do furacão**.	Na Jamaica, muitos **estocaram alimentos, água, lanternas e velas**.

Fonte: Camargo (2013)

Quadro 14 – Regras gerais para o alinhamento sentencial

Regra	Exemplo
Alinhar com base na sobreposição de conteúdo e não de forma Essa regra estabeleceu que o alinhamento fosse feito em função da sobreposição de conteúdo entre uma SS e uma ou mais SDs, e não em função da ocorrência de unidades lexicais comuns ou mesmo estruturas sintáticas semelhantes.	*Sentença do Sumário (SS):* A expectativa de lideranças da Câmara e do Conselho de Ética é que pouco mais de 10% dos 69 deputados denunciados no relatório parcial da CPI **abrirão mão de seus mandatos**.
	Sentença do Documento (SD): **As renúncias** têm que ser publicadas até terça-feira, quando o presidente do Conselho de Ética, deputado Ricardo Izar (PTB-SP), vai instaurar os processos de perda de mandato contra os 69 deputados acusados pela CPI dos Sanguessugas de envolvimento com a máfia das ambulâncias.
Alinhar com base na sobreposição da informação principal Essa regra estabeleceu que o alinhamento fosse feito em função do conteúdo principal veiculado pelas sentenças. Assim, uma SS foi alinhada a SDs quando houve sobreposição da ideia central, expressa pelo verbo principal.	*Sentença do Sumário:* Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, **descobriram** um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro, denominado Oph 162225-240515, o primeiro planemo duplo.
	Sentença do Documento: Os pesquisadores Ray Jayawardhana e Valentin D. Ivanov **informam** a descoberta na edição de quinta-feira do serviço *online* Science Express, mantido pela revista Science.
Alinhar com base na sobreposição de informação secundária Essa regra especificou que as sentenças fossem alinhadas diante da sobreposição de conteúdo ou informação secundária. Assim, uma SS foi alinhada a uma ou mais SDs não somente pelo conteúdo principal, mas também pelo compartilhamento de informação periférica.	*Sentença do Sumário:* Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.
	Sentença do Documento: Ambos os mundos têm massa semelhante à de outros exoplanetas já catalogados, mas não giram em torno de uma estrela - na verdade, **giram em torno um do outro**.
Alinhar todas as sobreposições de um mesmo conteúdo Essa regra estabeleceu que uma SS deveria ser alinhada sempre que uma SD com sobreposição de conteúdo fosse identificada, mesmo que a SS já tivesse sido alinhada por causa do compartilhamento desse mesmo conteúdo.	*Sentença do Sumário:* Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.
	Sentença do Documento: Astrônomos do Observatório Europeu Austral, localizado no Chile, anunciaram a descoberta de uma dupla de planetas errantes (sem estrela-mãe) que **giram ao redor deles mesmos** e que vagam livremente pelo espaço. *Sentença do Documento:* O fato extraordinário é que **ele não gira em volta de uma estrela, mas em torno de outro corpo frio** com o dobro de sua massa.

Fonte: Camargo (2013)

Nos exemplos de alinhamento apresentados no Quadro 14, é possível observar que sentenças que veiculavam informação em comum foram alinhadas, mesmo que algumas apresentassem baixa sobreposição lexical (*word overlap*).

Para cada sentença dos sumários manuais dos *clusters* C1, C2 e C9 (*clusters* que foram UNLizados), foram identificadas quais sentenças dos textos-fonte continham informações relacionadas. Cabe observar que a sumarização humana é abstrativa e envolve processos de reescrita, tais como fusão, substituição e generalização de informações. Assim, identificar, nos textos-fonte, as informações a partir das quais esses processos foram realizados nem sempre é uma tarefa simples. Por vezes, essa avaliação é subjetiva, apoiando-se na intuição do falante.

Além do alinhamento entre as sentenças do sumário e as sentenças dos textos-fonte, foi feito também o alinhamento entre as representações conceituais dessas mesmas sentenças. Assim, o alinhamento foi de dois tipos:

a) alinhamento sentencial: sentenças dos sumários ↔ sentenças correspondentes nos textos-fonte;
b) alinhamento conceitual: representação UNL das sentenças dos sumários ↔ representação UNL das sentenças correspondentes nos textos-fonte.

O esquema utilizado para o alinhamento é ilustrado no Quadro 15. A Figura 20 mostra um exemplo do alinhamento obtido seguindo essa estrutura.

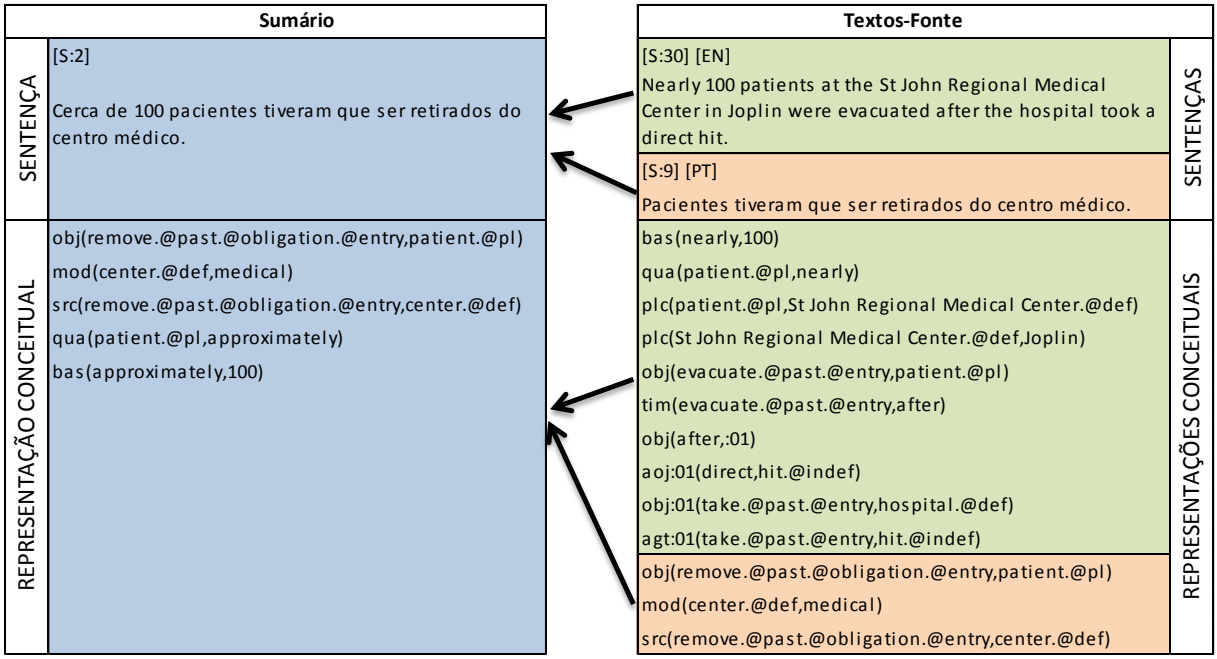
Quadro 15 – Esquema utilizado para os alinhamentos sentencial e conceitual

Sentença do Sumário	**Sentença(s) dos Textos-fonte**
[Identificação da sentença] *Sentença do sumário*	*[Identificação da sentença] [EN/PT]*[39] *Sentença(s) do(s) texto(s)-fonte*
Representação em UNL da sentença do sumário	*Representação em UNL da(s) sentença(s) do(s) texto(s)-fonte*

Fonte: Elaborado pelo autor

[39] "EN" para texto-fonte em inglês e "PT" para texto-fonte em português.

Figura 20 – Exemplo de alinhamentos sentencial e conceitual



Fonte: Elaborado pelo autor

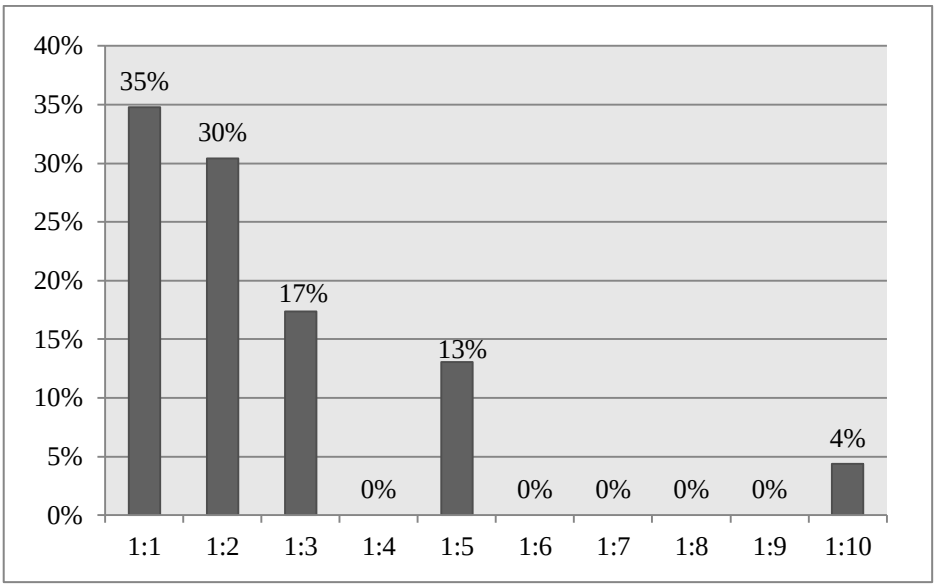
Na Tabela 1, apresenta-se a distribuição dos tipos de alinhamento encontrados. Um alinhamento 1:2, por exemplo, significa que, para uma sentença do sumário, foram localizadas duas sentenças com sobreposição de conteúdo nos textos-fonte. No Gráfico 1 apresenta-se a percentagem dos diferentes tipos de alinhamento.

Tabela 1 – Tipos de alinhamento encontrados

	Tipos de Alinhamento									
	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8	1:9	1:10
Número de ocorrências	8	7	4	0	3	0	0	0	0	1

Fonte: Elaborado pelo autor

Gráfico 1 – Tipos de alinhamento encontrados – distribuição percentual



Fonte: Elaborado pelo autor

5 INVESTIGAÇÃO DE ESTRATÉGIAS DE SELEÇÃO DE CONTEÚDO

Concluída a tarefa de alinhamento das sentenças dos sumários manuais às sentenças dos textos-fonte e de suas respectivas representações conceituais, a etapa seguinte consistiu em investigar como a representação em UNL poderia ser utilizada para a elaboração de estratégias de seleção de conteúdo relevante para composição do sumário. Buscou-se identificar as sentenças cujas representações conceituais codificassem informações selecionadas pelos sumarizadores humanos para fazer parte do sumário, visando elaborar estratégias consistentes de seleção de conteúdo que pudessem ser formalizadas e aplicadas às representações conceituais.

Convém ressaltar que a seleção de conteúdo seria apenas uma das tarefas necessárias para o desenvolvimento de um sistema de SAMM baseado em representações conceituais. No modelo de Sparck Jones (1998), apresentado na seção 2.1.1, essa atividade se localizaria na etapa de transformação, ou seja, após a etapa de análise e anteriormente à síntese.

Uma vez que o tipo de alinhamento realizado reflete sobreposição total ou parcial de conteúdo entre as sentenças do sumário e as sentenças dos textos-fonte, o fato de uma sentença do texto-fonte ter sido alinhada a uma sentença do sumário significa que pelo menos uma parte da informação da sentença do texto-fonte também consta no sumário, indicando tratar-se de conteúdo considerado relevante pelo sumarizador humano.

Nos casos de alinhamento, a sobreposição de conteúdo pode ser total ou parcial. Assim, a relevância da informação contida na sentença do texto-fonte pode ser maior ou menor, dependendo do grau de sobreposição de conteúdo. Entretanto, é razoável presumir que, de um modo geral, as sentenças dos textos-fonte alinhadas a sentenças do sumário contêm informações mais relevantes do que sentenças que não foram alinhadas. Assim, tentou-se identificar quais as características conceituais das sentenças que foram alinhadas, em comparação com as que não foram alinhadas.

Em Tosta (2014), por exemplo, a pressuposição é que as sentenças com os conceitos mais frequentes da coleção contêm as informações mais relevantes e, portanto, devem ser selecionadas para o sumário. Já em Sornlertlamvanich et al. (2001), na etapa de seleção de conteúdo, calcula-se uma pontuação para cada sentença com base no peso de cada UW. Esse peso reflete não somente a frequência do conceito, mas também a frequência inversa desse conceito na coleção. Em outras palavras, utiliza-se a frequência dos conceitos corrigida pela sua frequência inversa nos documentos (conforme apresentado na Seção 2.3.2).

Pode-se então questionar: dentre essas duas formas de selecionar conteúdo, qual melhor corresponde ao que um sumarizador humano de fato realiza? Há alguma outra forma de utilizar as informações conceituais de modo a se obter uma seleção de conteúdo que mais se aproxime da sumarização humana? Qual é o papel das relações semânticas na seleção de conteúdo realizada por humanos?

Em se tratando de uma investigação inicial, de caráter exploratório e bastante restrita no que diz respeito à quantidade de dados analisados, não se pretendeu fornecer respostas definitivas a essas questões, mas sim levantar hipóteses e apontar alguns caminhos que se mostrassem promissores para investigações futuras.

5.1 Seleção de conteúdo com base em informações conceituais

Apoiando-se na revisão literária, inicialmente duas estratégias de seleção de conteúdo com base em informações conceituais foram consideradas: selecionar as sentenças com base na frequência dos conceitos que as constituem, o que se assemelha a um dos métodos apresentados por Tosta (2014), e selecionar as sentenças com base na frequência dos conceitos multiplicada pela frequência inversa nos documentos (TF * IDF), conforme proposto por Sornlertlamvanich et al. (2001).

Na primeira estratégia, cada conceito é pontuado conforme sua frequência na coleção, ou seja, contabilizando-se o número de vezes que ele ocorre. Um conceito que ocorre 15 vezes no *cluster* recebe pontuação "15", enquanto que um conceito que aparece uma única vez recebe pontuação "1". Somando-se as pontuações dos conceitos que constituem uma sentença, obtém-se a pontuação dessa sentença, com base na qual o ranque de sentenças é elaborado. Na sentença "*Seven people have been rescued from the rubble*", por exemplo, os quatro conceitos identificados, lexicalizados em inglês como *seven*, *person*, *rescue* e *rubble*, ocorreram 7, 1, 1 e 3 vezes, respectivamente. Assim, a pontuação dessa sentença foi 7 + 1 + 1 + 3 = 12.

Observou-se, no entanto, que esse método de ranqueamento de sentenças tende a atribuir pontuação mais elevada a sentenças maiores, uma vez que elas contêm um número maior de conceitos, e pontuações mais baixas a sentenças menores. Em outras palavras, essa forma de ranqueamento privilegia sentenças mais longas e penaliza sentenças menores. Assim, propôs-se também uma estratégia de ranqueamento em que o fato de uma sentença ser muito longa ou muito curta fosse menos relevante para sua classificação. Com base nessa

métrica proposta, a pontuação de cada sentença foi calculada inicialmente da mesma forma, ou seja, somando-se a frequência de ocorrência dos conceitos nela presentes, porém essa pontuação foi dividida pelo número de conceitos da sentença. Em outras palavras, a pontuação de cada sentença foi normalizada em função de seu tamanho, expresso em número de conceitos.

Foram investigadas, portanto, três formas de pontuar e ranquear as sentenças com base em suas informações conceituais:

- Método F(UWs) = Frequência simples das UWs:

$$S(s) = \sum_{\forall UWi \in s} F(UWi)$$

onde

S é a função de pontuação das sentenças;

s é a sentença sendo pontuada;

F é a frequência de ocorrência do conceito; e

UWi é o conceito.

- Método F(UWs) * IDF (UWs) = Frequência das UWs corrigida pela frequência inversa nos documentos:

$$S(s) = \sum_{\forall UWi \in s} W(UWi) \quad (7)$$

$$W(UWi) = F(UWi) * IDF(UWi) \quad (8)$$

$$IDF(UWi) = \log\left(\frac{D(UWi)}{d(UWi)}\right) \quad (9)$$

onde

S é a função de pontuação das sentenças;

s é a sentença sendo pontuada;

W é a função que calcula o peso de cada conceito;

UWi é o conceito;

F é a frequência de ocorrência do conceito;

IDF é a frequência inversa do conceito nos documentos;

D(UWi) é o número de documentos do *corpus;* e

d(UWi) é o número de documentos em que a UW ocorre.

- Método F(UWs) / n. de UWs = Frequência das UWs normalizada pelo tamanho da sentença:

$$S(s) = \frac{\sum_{\forall UWi \in s} F(UWi)}{n(s)}$$

onde

- S é a função de pontuação das sentenças;
- s é a sentença sendo pontuada;
- F é a frequência de ocorrência do conceito;
- UWi é o conceito; e
- n(s) é o número de UWs na sentença *s*.

Essas estratégias de seleção de sentenças baseadas em informações conceituais foram comparadas entre si e também a um método de seleção tradicional baseado em uma característica superficial – a posição da sentença no texto-fonte. Segundo essa estratégia, as sentenças localizadas no início do texto ocupam as primeiras posições no ranque, enquanto que as sentenças localizadas no final ocupam as últimas posições.

Visando facilitar a comparação entre as estratégias de seleção de conteúdo investigadas, a pontuação obtida para cada sentença passou ainda por uma normalização que resultou, ao final, em sentenças pontuadas em uma escala de 0 a 10. O objetivo foi converter as escalas de pontuações originalmente obtidas, bastante variáveis de método para método, em uma escala única. Assim, em cada método, a sentença mais bem ranqueada teve pontuação final igual a 10, enquanto que a sentença menos bem ranqueada ficou com pontuação igual a 0. Os resultados antes e após essa normalização foram identificados como "Pontuação – Valor bruto" e "Pontuação – Valor normalizado", respectivamente (Apêndice A).

Nos casos em que duas sentenças tenham recebido a mesma pontuação, foi necessário adotar um critério de desempate para proceder com o ranqueamento. No caso, foi utilizado como critério o número de UWs: como o objetivo é realizar seleção de conteúdo para um eventual método de sumarização baseado em informações conceituais, caso duas ou mais sentenças tenham pontuação igual, selecionar a menor delas pode ser uma estratégia interessante para melhorar a taxa de compressão. Assim, nos casos de empate na pontuação, o segundo critério de ordenação foi o tamanho da sentença, medido em número de UWs, dando-se preferência para sentenças menores.

Por fim, a esses dados também se agregaram informações obtidas durante o processo de alinhamento. Conforme explanado anteriormente, caso uma sentença de um texto-fonte

tenha sido alinhada a uma sentença do sumário, isso quer dizer que ela apresenta conteúdo considerado relevante pelo sumarizador humano, uma vez que o alinhamento indica sobreposição de conteúdo entre o texto-fonte e o sumário. Assim, analisando se esses métodos são capazes de gerar ranques nos quais as sentenças mais bem pontuadas tenham sido alinhadas a sentenças dos sumários, é possível (i) comparar esses métodos entre si e (ii) e avaliar se a seleção de conteúdo por eles gerada correlaciona-se com a seleção de conteúdo realizada pelo sumarizador humano.

Idealmente, nesses métodos de seleção de conteúdo, as sentenças mais bem ranqueadas deveriam estar alinhadas a sentenças do sumário, pois isso significaria que elas trazem informações associadas às do sumário (sobreposição total ou parcial de conteúdo). Já as sentenças menos bem ranqueadas não deveriam estar alinhadas, ou seja, conteriam informações dissociadas das presentes no sumário.

Assim sendo, os métodos de seleção de conteúdo foram avaliados quanto à capacidade de gerar um ranqueamento cujas sentenças iniciais contenham informações relevantes, ou seja, que tenham sido alinhadas ao sumário. Mais uma vez, ressalta-se que essa seria apenas uma das etapas de um possível método de SAMM, visto que se trata apenas da seleção de conteúdo. Após essa etapa, seria necessário lidar com fenômenos multidocumento como redundância, contradição e complementaridade, de forma a se obter um sumário mais coeso, coerente e com melhor informatividade.

Cruzando as informações dos ranques de sentenças construídos e informações provindas da tarefa de alinhamento, foram construídas as Tabelas 4 a 27, que permitiram a comparação entre as estratégias de seleção de conteúdo. Essas tabelas são apresentadas no Apêndice A (Resultados do Ranqueamento de Sentenças), e estão divididas conforme o *cluster* (C1, C2 e C9), o texto-fonte (português / inglês) e o método de seleção de conteúdo utilizado para gerar o ranque. Em cada tabela, as sentenças estão ordenadas conforme a pontuação obtida na estratégia de seleção de conteúdo analisada.

5.2 Comparação das estratégias de seleção de conteúdo

Concluída a etapa de ranqueamento das sentenças com base nas quatro estratégias de seleção de conteúdo investigadas inicialmente (três conceituais e uma superficial), a etapa seguinte foi analisar os ranques para comparar essas diferentes formas de seleção.

Como observado na seção anterior, o critério adotado para que se considere uma sentença de um texto-fonte relevante ou não foi o alinhamento a uma ou mais sentenças do sumário: caso a sentença tenha sido alinhada, houve sobreposição de conteúdo e, portanto, ela foi considerada relevante (em comparação com as sentenças não alinhadas). Nesses termos, um método de seleção de conteúdo eficaz seria aquele em que o ranqueamento obtido trouxesse, em primeiro lugar, as sentenças dos textos-fonte que tenham sido alinhadas a sentenças do sumário, deixando por último as sentenças não alinhadas.

Para comparar entre si os métodos, fez-se então a seguinte pergunta: das n sentenças mais bem pontuadas em cada texto-fonte, quantas foram alinhadas a sentenças do sumário? Em outras palavras, das n sentenças mais bem ranqueadas, quantas podem ser consideradas relevantes, com base no alinhamento entre os textos-fonte e os sumários humanos?

O passo seguinte foi decidir o valor de n, ou seja, quantas sentenças seriam selecionadas em cada texto-fonte para fazer a comparação entre os métodos. Estabelecer um valor fixo pareceu problemático, visto que há uma variação razoavelmente grande no tamanho dos textos-fonte. Selecionar 6 sentenças, por exemplo, equivaleria a selecionar apenas 1/6 das sentenças do texto-fonte em inglês da coleção C1 (que tem ao todo 36 sentenças), mas significaria selecionar quase a metade das sentenças do texto-fonte em inglês da coleção C2 (que possui 13 sentenças).

Assim sendo, decidiu-se utilizar um valor proporcional ao tamanho do texto-fonte. Estabeleceu-se, arbitrariamente, que n seria equivalente a 1/5 do número de sentenças do texto-fonte, com arredondamento para baixo, se necessário. No texto-fonte em português do *cluster* C2, por exemplo, em que há 11 sentenças, n = (1/5) * 11 = 2,2. Após o arredondamento, obteve-se n = 2, o que quer dizer que, nessa coleção, a seleção de conteúdo analisada abrangeu as 2 primeiras sentenças.

Com base nos resultados dessa análise, construiu-se o Quadro 16, cujo objetivo foi permitir uma comparação mais objetiva entre as estratégias de seleção de conteúdo investigadas.

Quadro 16 – Comparação entre as estratégias de seleção de conteúdo investigadas

Documento (coleção e texto-fonte) / N. de sentenças	**Questões propostas para comparação entre os métodos**	**Métodos baseados na representação conceitual**			**Método superficial**
		F(UWs)[40]	**F(UWs) * IDF (UWs)[41]**	**F(UWs) / n. de UWs[42]**	**Posição no texto-fonte**
Cluster: C1 Texto-Fonte: EN 36 sentenças	Das 7 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	6	6	6	2
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	86%	86%	86%	29%
Cluster: C1 Texto-Fonte: PT 17 sentenças	Das 3 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	1	1	2	2
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	33%	33%	67%	67%
Cluster: C2 Texto-Fonte: EN 13 sentenças	Das 2 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	0	0	0	2
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	0%	0%	0%	100%
Cluster: C2 Texto-Fonte: PT 11 sentenças	Das 2 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	2	1	2	2
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	100%	50%	100%	100%
Cluster: C9 Texto-Fonte: EN 33 sentenças	Das 6 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	1	0	0	4
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	17%	0%	0%	67%
Cluster: C9 Texto-Fonte: PT 25 sentenças	Das 5 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	1	2	0	4
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	20%	40%	0%	80%
Dos 6 textos-fonte, em quantos deles o método de seleção gerou um percentual de sentenças alinhadas superior a 50%?		2	1	3	5

Fonte: Elaborado pelo autor

[40] Frequência simples das UWs
[41] Frequência das UWs corrigida pela frequência inversa nos documentos
[42] Frequência das UWs normalizada pelo tamanho da sentença

5.2.1 Comparação entre as estratégias baseadas em informações conceituais

A primeira análise realizada com base nas informações do Quadro 16 foi a comparação entre as três estratégias de seleção de conteúdo que utilizam informações conceituais.

No texto C1-EN, os três métodos foram capazes de selecionar conteúdo considerado relevante pelo sumarizador humano: dentre as 7 sentenças mais bem ranqueadas por esses métodos, 6 foram alinhadas ao sumário. No texto C2-PT, o primeiro e o terceiro métodos podem ser considerados bem sucedidos (Método F(UWs) e Método F(UWs) / n. de UWs), visto que geraram seleções de sentenças 100% alinhadas a sentenças do sumário.

Já nos textos C2-EN, C9-EN e C9-PT, os 3 métodos geraram ranques cujas primeiras posições foram primordialmente ocupadas por sentenças não alinhadas, ou seja, cujo conteúdo foi considerado não relevante pelo sumarizador humano.

No texto C1-PT, o método F(UWs) / n. de UWs saiu-se ligeiramente melhor, selecionando 2 sentenças alinhadas (dentre as 3 mais bem pontuadas), enquanto que os outros dois métodos selecionaram apenas 1 sentença alinhada.

De uma forma geral, pode-se observar que as diferenças entre os 3 métodos foram relativamente pequenas. Considerando-se uma estratégia de seleção de conteúdo bem sucedida quando mais de 50% das sentenças selecionadas estão alinhadas ao sumário (ou seja, contenham informação relevante), o método baseado em informações conceituais com melhor resultado foi o F(UWs) / n. de UWs (frequência das UWs normalizada pelo tamanho da sentença), conforme pode-se visualizar na última linha do Quadro 16.

Dentre as três estratégias de seleção baseadas unicamente em conceitos, o Método F(UWs) * IDF (UWs) (frequência de UWs corrigida pela frequência inversa nos documentos) foi o que gerou ranques com menor número de sentenças alinhadas dentre as mais bem pontuadas, ou seja, levaria à seleção de menos sentenças com conteúdo considerado relevante pelo sumarizador humano. Em apenas 1 dos 6 textos-fonte o percentual de sentenças alinhadas dentre as mais bem ranqueadas foi superior a 50%.

É difícil apontar, com exatidão, as razões para tal resultado. Entretanto, observando-se os dados coletados, nota-se que o tamanho do *corpus* e a própria lógica que embasa a fórmula parecem ser fatores relevantes. Nessa métrica, conceitos que ocorrem em todos os textos acabam recebendo peso igual a zero, o que seria uma forma de diminuir a influência de palavras de ocorrência comum na língua. Entretanto, em *corpora* pequenos, com apenas 2 ou 3 textos, por exemplo, a chance de uma UW ocorrer em todos os textos ainda é relativamente

grande, e a forma de cálculo acaba atribuindo peso zero a essas UWs, muitas vezes desprezando conceitos importantes.

Para explorar melhor essa hipótese, fez-se o levantamento dos conceitos mais frequentes das coleções. No *cluster* C9, por exemplo, os 10 conceitos mais frequentes foram os correspondentes aos itens lexicais *tornado*, *say*, *city*, *person*, *Missouri*, *storm*, *Joplin*, *Jeff Lehr*, *hit* e *strike*. Desses, todos, exceto *Jeff Lehr*, ocorreram em ambos os textos da coleção, recebendo pontuação igual a zero em virtude da forma de calcular o peso das UWs: embora a frequência (TF) desses conceitos seja elevada, a frequência inversa nos documentos (IDF) é igual a zero. O que aconteceu nessa coleção, portanto, foi que, dos 10 conceitos mais frequentes, 9 ficaram com peso igual a zero. A situação foi semelhante nos outros dois *clusters*, o que coloca em cheque a capacidade de essa métrica excluir conceitos comuns e preservar os mais relevantes, pelo menos em coleções pequenas. É necessário investigar a aplicação dessa estratégia de seleção de conteúdo em *corpora* com mais textos, para que se possa conhecer melhor o seu potencial.

Nos outros dois métodos baseados em conceitos – a frequência simples de UWs e a frequência de UWs normalizada pelo tamanho da sentença – o número de sentenças alinhadas selecionadas foi ligeiramente maior, embora seja difícil estabelecer a real significância dessa diferença com a quantidade de dados analisada.

O Método Frequência de UWs normalizada pelo tamanho da sentença gerou um maior número de sentenças alinhadas, comparado aos outros dois, aproximando-se um pouco mais da seleção de conteúdo realizada pelos sumarizadores humanos nos textos analisados. Em 3 dos 6 textos-fonte, essa estratégia foi capaz de gerar ranques cujas sentenças mais bem pontuadas estavam alinhadas em mais de 50% dos casos. Ou seja, em metade dos textos-fonte, pode-se considerar que houve uma boa correlação entre o conteúdo considerado relevante pelo sumarizador humano e o conteúdo das sentenças elencadas por esse método (mais da metade das sentenças nesses 3 textos-fonte apresentavam sobreposição parcial ou total de conteúdo com relação às sentenças dos sumários humanos). No Método Frequência simples das UWs, isso ocorreu em 2 dos 6 textos-fonte.

Uma vez que Tosta (2014) demonstrou que um método baseado na frequência simples de conceitos teve desempenho melhor do que métodos tradicionais que fazem a TA integral de textos-fonte, sugere-se que o método de seleção proposto neste estudo – Frequência de conceitos normalizada pelo tamanho da sentença – seja objeto de uma investigação mais detalhada, considerando-se os resultados obtidos. Apesar de serem poucos os dados analisados, os resultados indicam tratar-se de um método que poderia levar a uma seleção de

conteúdo um pouco mais próxima à realizada pelo sumarizador humano, comparado à Frequência simples de UWs.

5.2.2 Estratégias conceituais x estratégia superficial

A segunda análise feita a partir do Quadro 16 foi a comparação entre os métodos de seleção de conteúdo baseados em informações conceituais e um método superficial, mais especificamente a seleção de sentenças com base em sua posição no texto-fonte.

O único caso em que a estratégia de selecionar com base na posição da sentença gerou seleção de conteúdo com menos alinhamentos do que os métodos conceituais foi no texto C1-EN, em que apenas 2 das 7 sentenças mais bem ranqueadas estavam alinhadas ao sumário (nos métodos conceituais, 6 das 7 sentenças estavam alinhadas). Em dois textos, C1-PT e C2-PT, houve empate quanto ao número de sentenças alinhadas entre o Método Posição no texto-fonte e o método conceitual mais bem sucedido (Frequência das UWs normalizada pelo tamanho da sentença). Nos outros três textos, selecionar com base na posição no texto-fonte levou a mais alinhamentos dentre as sentenças mais bem ranqueadas do que qualquer um dos três métodos baseados em informações conceituais.

Dos 6 textos-fonte analisados, em 5 deles a seleção de sentenças com base em sua posição gerou ranques cujas sentenças mais bem pontuadas estavam alinhadas em mais de 50% dos casos. Em outras palavras, em 5 dos 6 textos-fonte, mais de metade das sentenças selecionadas com base nesse método traziam conteúdo relevante. No método conceitual mais bem sucedido – Frequência das UWs normalizada pelo tamanho da sentença – isso ocorreu em apenas 3 dos 6 textos analisados.

Se a frequência de conceitos não pareceu ser a principal motivação para a seleção de conteúdo nesses textos-fonte, cabe indagar quais outras estratégias foram utilizadas pelos humanos na sumarização desses textos. Também seria importante analisar as consequências de se utilizar informações conceituais como critério de seleção de conteúdo em textos como esses, ou seja, qual seria a seleção de conteúdo realizada com base em tais estratégias e como ela diverge da seleção de conteúdo realizada pelos sumarizadores humanos.

Embora essa pareça ser uma frente de investigação com bastante potencial para proporcionar esclarecimentos sobre a SHMM e, consequentemente, relevante para a SAMM, as limitações de tempo naturais em um projeto como esse não permitiram adentrar nessas questões o tanto quanto desejado. Ainda assim, uma breve análise foi realizada nesse sentido, na qual comparou-se o conteúdo selecionado pelo método conceitual mais bem sucedido neste

estudo – frequência de UWs normalizada pelo tamanho da sentença – com o conteúdo selecionado com base na posição da sentença no texto-fonte, verificando-se a relação desses com o sumário humano de referência. Fez-se isso em dois textos nos quais o método conceitual gerou seleção de conteúdo bastante distinta da do sumarizador humano: o texto-fonte em inglês da coleção C2 (C2-EN) e o texto-fonte em português da coleção C9 (C9-PT). Um resumo dessa análise é trazido nos Quadros 17 e 18.

Quadro 17 – Comparação entre o conteúdo selecionado por um método conceitual e um método superficial no texto-fonte C2-EN

Método F(UWs) / n. de UWs aplicado ao texto-fonte C2-EN				**Método Posição no texto-fonte aplicado ao texto-fonte C2-EN**		
Seleção de conteúdo gerada pelo método (sentenças mais bem ranqueadas)		**A sentença foi alinhada ao sumário? (Se sim, a qual sentença?)**		**Seleção de conteúdo gerada pelo método (sentenças mais bem ranqueadas)**		**A sentença foi alinhada ao sumário? (Se sim, a qual sentença?)**
S:05	"She didn't like what she saw," Gilberto Carvalho said.	não		S:00	President Dilma Rousseff has suspended the distribution and production of sex education films for schools in Brazil.	sim (S:00)
S:13	"If she doesn't do a U-turn and change her mind, I will urge all gay people not to vote for her again."	não		S:01	President Rousseff believes the footage is not suitable for youngsters.	sim (S:01)
Sumário humano de referência – *Cluster* C2						
S:00	A presidente Dilma Rousseff suspendeu a produção e circulação do kit anti-homofobia dos Ministérios da Educação e Saúde.					
S:01	O kit, contendo vídeos com cenas lésbicas e gays para supostamente combater a homofobia, foi considerado impróprio.					
S:02	A decisão foi tomada depois que a bancada evangélica ameaçou não votar projetos da Câmara até a retirada do material, pois considera o kit um estímulo ao homossexualismo.					
S:03	Ativistas dos direitos dos homossexuais, como o deputado Jean Wyllys, viram a suspensão como falta de comprometimento com os direitos humanos.					

Fonte: Elaborado pelo autor

Quadro 18 – Comparação entre o conteúdo selecionado por um método conceitual e um método superficial no texto-fonte C9-PT

Método F(UWs) / n. de UWs aplicado ao texto-fonte C9-PT			**Método Posição no texto-fonte aplicado ao texto-fonte C9-PT**		
Seleção de conteúdo gerada pelo método (sentenças mais bem ranqueadas)		**A sentença foi alinhada ao sumário? (Se sim, a qual sentença?)**	**Seleção de conteúdo gerada pelo método (sentenças mais bem ranqueadas)**		**A sentença foi alinhada ao sumário? (Se sim, a qual sentença?)**
S:14	Em Minneapolis, tornados provocaram o fechamento de estradas e rodovias devido à queda de árvores e fios elétricos, além de terem provocado vazamentos de gás e destruído casas.	não	S:00	O número de mortos após a passagem de um tornado pela cidade de Joplin, no Estado americano do Missouri, subiu para ao menos 89 nesta segunda-feira, informaram autoridades locais durante coletiva de imprensa exibida na TV.	sim (S:00)
S:21	A moradora Carla Tabares conta que ela, o marido e várias crianças se abrigaram na cozinha de um restaurante da cidade no momento da passagem do tornado.	não	S:01	"Temos 89 mortes confirmadas devido ao tornado", anunciou o prefeito de Joplin, Mark Rohr.	sim (S:00)
S:22	"Foi horrível, muito assustador. Estou grata por estar viva, e sinto muito pelos que não estão", disse ela.	não	S:02	A cidade fica perto das divisas com Kansas e Oklahoma.	não
S:25	Outro tornado atingiu o norte de Minneapolis neste domingo, causando destruição, matando uma pessoa e ferindo outras 30.	não	S:03	De acordo com John Miller, fotógrafo que trabalha para o jornal "The Springfield News-Leader", armazéns da Home Depot e do Wallmart foram destruídos, assim como postos de gasolina e prédios.	sim (S:01)
S:27	No sábado (21), tornados mataram um e danificaram cerca de 200 imóveis no nordeste do Kansas.	não	S:04	O tornado foi registrado quase um mês depois de uma série de fenômenos similares que matou 354 pessoas em sete Estados do país.	sim (S:07)

Sumário humano de referência – *Cluster* C9	
S:00	Um tornado atravessou a cidade de Joplin, no estado americano de Missouri, matando pelo menos 116 pessoas e ferindo outras 1.000.
S:01	O prefeito de Joplin, Mark Rohr, disse que a tempestade percorreu um caminho de 10 quilômetros, arrasando edifícios e danificando o hospital regional St. John, que teve que ser evacuado.
S:02	Cerca de 100 pacientes tiveram que ser retirados do centro médico.
S:03	Ventos fortes e granizo atingiram a cidade, que ainda está grande parte sem energia.
S:04	O tornado derrubou linhas de energia e muitos dos serviços de telefonia foram interrompidos.
S:05	O governador de Missouri, Jay Nixon, declarou estado de emergência e ativou as tropas da Guarda Nacional.
S:06	Ele advertiu que mais tempestades estão a caminho.
S:07	No mês passado, tornados e tempestades causaram a morte de 354 pessoas no estado do Alabama e em seis outros estados no sudoeste dos Estados Unidos.
S:08	A Casa Branca disse na segunda-feira que o presidente Barack Obama se manteve informado sobre as tempestades letais no meio-oeste durante sua viagem de seis dias à Europa.
S:09	Obama divulgou um comunicado expressando condolências às famílias das vítimas e ordenando à Agência Federal de Gestão de Emergências que ajude na resposta imediata e nos esforços de reconstrução.

Fonte: Elaborado pelo autor

No Quadro 17, observa-se que, quando a estratégia de selecionar conforme a frequência de UWs normalizada pelo tamanho da sentença foi aplicada ao texto-fonte C2-EN, as duas sentenças mais bem ranqueadas foram a S:05 e a S:13. Convém recordar que, na representação em UNL, os pronomes são substituídos por seus antecedentes nos texto, sempre que possível. Assim, todas as referências a "*she*" ou "*her*", nessas sentenças, foram representadas pelos conceitos correspondentes a "*president*" e "*Dilma Roussef*". Estando entre as UWs mais frequentes da coleção, esses foram os conceitos que mais pesaram na pontuação dessas sentenças.

O que se nota, entretanto, é que dificilmente um sumarizador humano selecionaria o conteúdo dessas duas sentenças isoladamente, ou seja, sem apresentar as devidas informações contextuais. Embora a avaliação do que é relevante ou não seja, em parte, uma tarefa subjetiva, selecionar unicamente as sentenças S:05 e S:13 no texto C2-EN não parece ser uma boa estratégia de sumarização: ou o sumarizador deveria apresentar também as sentenças que introduzem as informações referenciadas nessas sentenças, colocando o leitor a par do contexto ao qual elas pertencem, ou as deixaria de fora.

Com relação ao conteúdo que seria selecionado pelo Método Posição no texto-fonte, vê-se que as sentenças S:00 e S:01 não apresentariam esse problema relacionado à contextualização. Se ambas fossem selecionadas, seria possível compreender o conteúdo delas sem a necessidade de agregar informações contextuais.

Nesse caso específico, selecionar com base nas informações conceituais poderia levar a problemas de coesão/coerência, enquanto que a seleção com base na posição da sentença no texto-fonte não ocasionaria esse inconveniente.

A segunda comparação entre o método F(UWs) / n. de UWs e o Método Posição no texto-fonte é apresentada no Quadro 18. Com base na leitura do texto-fonte C9-PT, observa-se que o evento principal relatado é um tornado na cidade de Joplin, Estados Unidos. Entretanto, a primeira sentença selecionada pelo método conceitual foi a S:14, sendo que nenhuma sentença anterior que introduzisse o assunto principal do texto foi elencada dentre as mais bem pontuadas por esse método. Mais ainda: nenhuma sentença contendo o conceito Joplin foi selecionada, no método F(UWs) / n. de UWs.

Visando compreender o ocorrido, fez-se uma análise do texto-fonte C9-PT, na qual notou-se que, em muitos casos, as referências ao tornado em Joplin eram implícitas. Destacam-se a seguir alguns trechos do texto-fonte em que isso ocorreu:

"Temos 89 mortes confirmadas devido ao tornado, anunciou o prefeito (...)".

(= ao tornado que ocorreu em Joplin)

"O tornado foi registrado quase um mês depois (...)".

(idem)

"A cidade fica perto das divisas com Kansas e Oklahoma."

(= a cidade de Joplin)

"Muitos dos prédios públicos da cidade ficaram amplamente danificados (...)".

(= da cidade de Joplin)

Mesmo uma representação conceitual não foi capaz de captar esses casos de correferencialidade, deixando de atribuir ao tornado de Joplin o devido peso. O conceito que mais influenciou na seleção de sentenças gerada pelo método conceitual analisado foi o correspondente a "*tornado*", o que chega a ser coerente com o assunto dos textos-fonte; porém, na maioria dessas sentenças, trata-se de tornados em outras localidades, e não do tornado de Joplin (ex.: S:14, S:25 e S:27, no Quadro 18). As informações dessas sentenças podem ser consideradas complementares, e não essenciais, visto que sumarizador humano não as mencionou no sumário. As outras duas sentenças selecionadas, S:21 e S:22, contêm um relato pessoal sobre a passagem do tornado. Também se trata de uma informação que não está presente no sumário e, portanto, foi considerada de baixa relevância pelo sumarizador humano.

Nesses dois casos analisados, notou-se que a seleção com base na frequência de conceitos não proporcionaria uma introdução adequada aos assuntos apresentados, o que provavelmente prejudicaria a coesão e coerência textuais. Além disso, no segundo caso estudado, a seleção de conteúdo gerada pelo método não estaria associada às informações consideradas mais relevantes pelo sumarizador humano (ex.: o tornado de Joplin), talvez em virtude da não identificação de correferencialidade em algumas situações.

5.2.3 Relações UNL e seleção de conteúdo

As propostas para seleção de conteúdo utilizando a representação UNL exploradas até aqui se basearam na frequência dos conceitos. Entretanto, cabe questionar: as relações UNL podem ser utilizadas para levar a uma melhor seleção de conteúdo? Visando tentar identificar estratégias de seleção de sentenças que utilizassem não apenas os conceitos, mas também as relações UNL, fez-se uma análise da distribuição das relações binárias nos textos-fonte e nos sumários.

Foram contabilizadas todas as relações das coleções C1, C2 e C9, verificando sua frequência de ocorrência nos textos-fonte e nos sumários. Para cada relação, calculou-se um índice que recebeu o nome de representatividade, obtido da seguinte forma:

$$Representatividade\ da\ RL(x) = \frac{N.\ de\ ocorrências\ da\ RL(x)}{N.\ de\ ocorrências\ de\ todas\ as\ RLs}$$

Esse cálculo foi uma forma de verificar quais foram as relações mais frequentes nas coleções e quais mais variaram entre os textos-fonte e os sumários (Tabela 2). A relação `mod` (*modifier*), por exemplo, representou 11,4% das relações dos textos-fonte. Nos sumários, ela passou a representar 8,4% das relações. Essa queda de representatividade é um indício que a relação tende a ser excluída quando vai para o sumário, comparada às outras RLs.

A partir da análise dessas variações de representatividade, cada RL recebeu um determinado peso. Considerou-se que as relações que ganharam representatividade no sumário tendem a ser mais preservadas e, portanto, têm peso maior; as que perderam representatividade tendem a ser excluídas e recebem peso menor. A Tabela 3 mostra as relações ordenadas conforme a pontuação recebida, segundo uma escala de 0 a 10.

Tabela 2 – Distribuição das relações nos textos-fonte e sumários

Relação (RL)	Número de ocorrências nos textos-fonte	Represen-tatividade nos textos-fonte (%)	Número de ocorrências nos sumários	Represen-tatividade nos sumários (%)	Diferença de represen-tatividade Sumários - Textos-fonte
agt	168	9,9%	23	8,4%	-1,4%
and	147	8,6%	19	7,0%	-1,7%
aoj	154	9,0%	30	11,0%	+2,0%
bas	16	0,9%	5	1,8%	+0,9%
ben	5	0,3%	0	0,0%	-0,3%
cnt	21	1,2%	5	1,8%	+0,6%
con	3	0,2%	1	0,4%	+0,2%
dur	13	0,8%	3	1,1%	+0,3%
equ	9	0,5%	1	0,4%	-0,2%
exp	9	0,5%	0	0,0%	-0,5%
gol	43	2,5%	9	3,3%	+0,8%
ins	3	0,2%	0	0,0%	-0,2%
man	82	4,8%	10	3,7%	-1,1%
mod	195	11,4%	23	8,4%	-3,0%
nam	54	3,2%	12	4,4%	+1,2%
obj	362	21,2%	65	23,8%	+2,6%
opl	2	0,1%	0	0,0%	-0,1%
or	4	0,2%	1	0,4%	+0,1%
plc	99	5,8%	15	5,5%	-0,3%
plf	3	0,2%	0	0,0%	-0,2%
plt	2	0,1%	0	0,0%	-0,1%
pof	18	1,1%	3	1,1%	+0,0%
pos	36	2,1%	6	2,2%	+0,1%
ptn	2	0,1%	0	0,0%	-0,1%
pur	29	1,7%	3	1,1%	-0,6%
qua	108	6,3%	23	8,4%	+2,1%
rsn	26	1,5%	2	0,7%	-0,8%
scn	12	0,7%	1	0,4%	-0,3%
seq	1	0,1%	1	0,4%	+0,3%
src	14	0,8%	3	1,1%	+0,3%
tim	52	3,1%	5	1,8%	-1,2%
tmf	3	0,2%	1	0,4%	+0,2%
tmt	4	0,2%	2	0,7%	+0,5%
via	5	0,3%	1	0,4%	+0,1%
Total	1704	100,0%	273	100,0%	-

Fonte: Elaborado pelo autor

Tabela 3 – Pontuação atribuída a cada relação

Posição	Relação	Diferença de representatividade	Pontuação atribuída
1	obj	+2,6%	10,0
2	qua	+2,1%	9,1
3	aoj	+2,0%	8,9
4	nam	+1,2%	7,6
5	bas	+0,9%	7,0
6	gol	+0,8%	6,8
7	cnt	+0,6%	6,5
8	tmt	+0,5%	6,3
9	dur	+0,3%	6,0
10	seq	+0,3%	6,0
11	src	+0,3%	5,9
12	con	+0,2%	5,7
13	tmf	+0,2%	5,7
14	or	+0,1%	5,6
15	pos	+0,1%	5,6
16	via	+0,1%	5,5
17	pof	0,0%	5,5
18	opl	-0,1%	5,2
19	plt	-0,1%	5,2
20	ptn	-0,1%	5,2
21	equ	-0,2%	5,1
22	ins	-0,2%	5,1
23	plf	-0,2%	5,1
24	ben	-0,3%	4,9
25	plc	-0,3%	4,8
26	scn	-0,3%	4,8
27	exp	-0,5%	4,5
28	pur	-0,6%	4,3
29	rsn	-0,8%	4,0
30	man	-1,1%	3,3
31	tim	-1,2%	3,2
32	agt	-1,4%	2,8
33	and	-1,7%	2,4
34	mod	-3,0%	0,0

Fonte: Elaborado pelo autor

A pontuação das RLs foi utilizada para ranquear sentenças em uma das coleções de textos-fonte (C1). Cada sentença de ambos os textos-fonte dessa coleção foi pontuada somando-se o peso das RLs que a constitui. O valor obtido passou então por uma normalização similar à realizada para os métodos baseados em conceitos: a pontuação de cada sentença foi convertida para uma escala de 0 a 10, de modo a facilitar a comparação com as outras estratégias de seleção de conteúdo.

Por fim, os valores obtidos foram utilizados para gerar três ranques diferentes:

- sentenças pontuadas com base unicamente no peso de suas RLs (Método RLs);
- sentenças pontuadas com base no peso de suas RLs e frequência de seus conceitos, em uma proporção de 1:1 (Método RLs + UWs 1:1);

- sentenças pontuadas com base no peso de suas RLs e frequência de seus conceitos, em uma proporção de 1:3 (Método RLs + UWs 1:3).

As hipóteses que guiaram esses ranqueamentos foram, respectivamente:

- os sumarizadores humanos selecionam as sentenças com base unicamente nas relações nelas presentes;
- os sumarizadores humanos selecionam as sentenças com base tanto nas suas relações quanto em seus conceitos, sendo que esses têm igual importância;
- os sumarizadores humanos selecionam as sentenças principalmente com base em seus conceitos, mas também levam em conta as relações semânticas presentes.

Os resultados desses ranques são apresentados nas Tabelas 28 a 33 (Apêndice A – Resultados do Ranqueamento de Sentenças). As informações dessas tabelas foram sintetizadas no Quadro 19, no qual as três estratégias de seleção de conteúdo que utilizam informações sobre as relações UNL foram comparadas a um método que utiliza unicamente informações sobre conceitos – F(UWs) / n. de UWs (frequência de UWs normalizada pelo tamanho da sentença).

Quadro 19 – Estratégias de seleção de conteúdo utilizando RLs

Documento (coleção e texto-fonte) / N. de sentenças	**Questões propostas para comparação entre os métodos**	**Método RLs[43]**	**Método RLs + UWs 1:1[44]**	**Método RLs + UWs 1:3[45]**	**Método F(UWs) / n. de UWs[46]**
Cluster: C1 Texto-Fonte: EN 36 sentenças	Das 7 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	4	4	6	6
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	57%	57%	86%	86%
Cluster: C1 Texto-Fonte: PT 17 sentenças	Das 3 sentenças mais bem pontuadas, quantas foram alinhadas ao sumário?	1	1	1	2
	Qual foi o % de sentenças selecionadas pelo método alinhadas ao sumário?	33%	33%	33%	67%

Fonte: Elaborado pelo autor

[43] Sentenças pontuadas com base unicamente no peso de suas RLs
[44] Sentenças pontuadas com base no peso de suas RLs e frequência de seus conceitos, em uma proporção de 1:1
[45] Sentenças pontuadas com base no peso de suas RLs e frequência de seus conceitos, em uma proporção de 1:3
[46] Frequência de UWs normalizada pelo tamanho da sentença

O intuito foi verificar se a utilização de informações sobre as RLs levaria a uma seleção de conteúdo que melhor se correlacionasse com a seleção feita pelo sumarizador humano, o que foi avaliado pelo número de sentenças alinhadas. Entretanto, nos três métodos propostos, o número de sentenças alinhadas foi menor ou igual ao método conceitual usado como base para comparação, o que indica que agregar informações sobre as RLs, da forma como foi feito, não foi útil para aprimorar a seleção de conteúdo.

Selecionar as sentenças com as RLs que tendem a ser preservadas no sumário não levou a uma melhor seleção de conteúdo nos casos analisados, mesmo que o ranqueamento das sentenças fosse feito em conjunto com informações conceituais (ou seja, selecionando as sentenças que contivessem os conceitos mais frequentes e também as relações que mais tendessem a ser preservadas).

Uma possível explicação para isso é que, embora conceitos de ocorrência frequente muitas vezes estejam ligados a informações relevantes (ex.: *storm*, na coleção C9), o mesmo não tende a ocorrer com as RLs. Exemplificando, a RL obj (= *patient*) foi utilizada na coleção C1 envolvendo tanto alguns dos conceitos mais frequentes da coleção (como os conceitos referentes a *tornado* e *city*) quanto envolvendo conceitos que ocorreram apenas uma ou duas vezes (como o caso de *help* e *coordinate*). Observando-se a representação conceitual em UNL do *corpus* analisado, parece difícil, se não impossível, prever quando uma determinada relação estará conectando conceitos que veiculam informações relevantes ou não.

O simples fato de uma relação ser mais frequente no sumário pode não significar muito, uma vez que sua relevância parece ser mais fortemente influenciada pelos conceitos que ela interliga do que por sua frequência em si. Um conceito que aparece muitas vezes pode significar informação importante; uma relação que aparece muitas vezes pode não significar nada, dependendo dos conceitos que ela estiver conectando.

Embora baseados em uma quantidade bastante limitada de dados, os resultados obtidos indicaram que selecionar sentenças com base na frequência de suas relações UNL não se mostrou mais vantajoso do que selecioná-las unicamente com base em seus conceitos.

No cenário da SA monodocumento, o uso de RLs já havia sido estudado por Martins e Rino (2002a, 2002b). Mesmo utilizando uma metodologia diferente, as autoras reportaram dificuldade em preservar a ideia central das sentenças ao realizar a sumarização com base unicamente nas relações, sendo que, em muitas situações, a informação principal acabava sendo descartada na seleção de conteúdo. Martins e Rino também observaram que uma mesma relação poderia ser utilizada para representar informações irrelevantes em algumas sentenças, mas informações essenciais em outras.

Assim sendo, a utilização de informações sobre as RLs na sumarização requer um estudo bem mais minucioso. Usar somente a frequência de ocorrência, como feito aqui, não foi suficiente para gerar uma seleção de conteúdo que se aproximasse mais da realizada por um sumarizador humano. Martins e Rino (2002a, 2002b) fizeram observações extremamente interessantes no cenário monodocumento, porém essas investigações devem ser expandidas para o cenário multidocumento, para que se consiga utilizar as RLs de uma maneira mais produtiva na SAM.

6 VERIFICAÇÃO DAS HIPÓTESES

De acordo com a Hipótese 1, a análise das representações conceituais das sentenças dos textos-fonte e das representações das sentenças de seus sumários manuais permitiria o desenvolvimento de estratégias de seleção de conteúdo que pudessem ser formalizadas e aplicadas às representações dos textos-fonte de uma coleção.

No entanto, na análise das características das representações conceituais, tanto a utilização de informações sobre a frequência de conceitos como de RLs não levou à obtenção de estratégias de seleção de conteúdo aplicáveis a todos os textos-fonte. Isso não significa que essas estratégias não possam ser utilizadas na SAMM: significa que o conteúdo por elas selecionado não se correlacionou com a seleção de conteúdo feita pelos sumarizadores humanos em todos os textos do *corpus* investigado.

Há de se considerar que uma das estratégias propostas gerou ranques cujo conteúdo apresentou correlação elevada com a seleção realizada pelos sumarizadores humanos em metade dos textos-fonte. Na outra metade, isso não ocorreu. De acordo com os resultados obtidos, a gama de estratégias de seleção de conteúdo utilizadas pelos sumarizadores humanos vai além de selecionar as sentenças com os conceitos ou relações mais frequentes, envolvendo questões como coesão, coerência e correferencialidade.

Assim, por ora, a Hipótese 1 não pôde ser confirmada. Cabe ainda ressaltar que existem outras formas de explorar as representações conceituais além das utilizadas neste trabalho. Aqui foram propostas, ao todo, 6 estratégias de seleção de conteúdo (3 baseadas em conceitos e 3 baseados nas relações conceituais binárias). Outras poderiam ser desenvolvidas, visando explorar melhor o potencial do sistema de representação de conhecimento analisado.

Com relação à Hipótese 2, propôs-se que o alinhamento das sentenças dos textos-fonte às sentenças de um sumário humano permitiria comparações entre as estratégias de seleção de conteúdo desenvolvidas com base nas representações conceituais, bem como avaliar quão bem elas se correlacionam com as estratégias utilizadas por humanos.

De fato, foi possível comparar as estratégias de seleção com base no alinhamento, visto que esse reflete sobreposição de conteúdo entre sentenças dos textos-fonte e sentenças do sumário humano. A metodologia proposta mostrou-se viável e permitiu comparar quantitativamente as estratégias de seleção de conteúdo investigadas. O número de sentenças alinhadas dentre as mais bem pontuadas pelos métodos reflete o quanto cada método se aproximou da seleção de conteúdo considerado relevante pelos sumarizadores humanos: mais

sentenças alinhadas significa mais sobreposição de conteúdo. Dessa forma, a Hipótese 2 foi confirmada.

Por fim, de acordo com a Hipótese 3, estratégias de seleção de conteúdo que utilizam representações conceituais correlacionar-se-iam melhor com a sumarização humana do que uma estratégia de seleção baseada em conhecimento superficial.

Na realidade, a situação foi inversa: em 5 dos 6 textos analisados, a seleção de sentenças com base em uma característica superficial – sua posição no texto-fonte – gerou ranques cujas sentenças mais bem pontuadas estavam alinhadas em mais de 50% dos casos. Parafraseando, em 5 dos 6 textos-fonte, mais de metade das sentenças selecionadas com base nesse método continham informação considerada relevante pelos sumarizadores humanos.

No método conceitual mais bem sucedido – Frequência das UWs normalizada pelo tamanho da sentença – isso ocorreu em 3 dos 6 textos analisados. Portanto, a Hipótese 3 está rejeitada, considerando-se as estratégias comparadas. Provavelmente isso se deve à natureza do gênero textual investigado: em textos jornalísticos, a informação presente no início do texto tipicamente veicula a ideia principal da notícia. Certamente seria interessante confrontar esses métodos em um cenário diferente, envolvendo textos de outros gêneros.

7 CONSIDERAÇÕES FINAIS

A metodologia utilizada, embora tenha sido aplicada a um *corpus* pequeno, permitiu algumas observações:

- em termos gerais, houve pouca diferença entre as três estratégias de seleção de conteúdo baseadas na frequência de conceitos. Entretanto, com base nos critérios adotados, utilizar a frequência de UWs normalizada pelo tamanho da sentença mostrou-se eficaz em um número maior de textos-fonte (3, dos 6 analisados), comparado aos métodos que utilizam a frequência simples de UWs e a frequência das UWs corrigida pela frequência inversa nos documentos;
- entre selecionar as sentenças com base em informações conceituais e selecionar as sentenças localizadas no início dos textos-fonte, a seleção de sentenças com base na posição no texto-fonte teve melhor desempenho nas coleções analisadas. O ranqueamento de sentenças com base na posição no texto-fonte gerou a seleção de conteúdo que melhor se correlacionou com a que os sumarizadores humanos realizaram, uma vez que houve um número maior de sentenças alinhadas selecionadas por esse método;
- as sentenças iniciais dos textos-fonte não necessariamente contêm os conceitos mais frequentes da coleção. Em diversos casos, as sentenças com conceitos mais frequentes estavam em posição intermediária ou final no texto, e não concentradas em seu início;
- a utilização de informações sobre as relações binárias, da forma como foi feita, não gerou seleções de conteúdo mais próximas à realizada pelos sumarizadores humanos.

À primeira vista, chama a atenção o fato de um método de seleção de conteúdo baseado em conhecimento superficial ter sido mais bem sucedido do que métodos que utilizam conhecimento profundo, em nível conceitual. Ao mesmo tempo, não deixa de ser um resultado interessante, pois pode refletir uma dissociação entre sentenças localizadas no início dos textos e sentenças com os conceitos mais frequentes. Conforme observado ao longo do *corpus*, muitas vezes as sentenças no meio ou no final do texto eram as que continham os conceitos mais frequentes da coleção.

Se o fato de um texto pertencer ao gênero jornalístico significa que suas sentenças iniciais veiculam as informações mais relevantes, e se as sentenças iniciais não

necessariamente contêm os conceitos mais frequentes de uma coleção (como sugerido neste estudo), conclui-se que uma sentença relevante não é necessariamente a que contém os conceitos mais frequentes. Outros critérios, além da frequência de conceitos, parecem determinar a relevância de uma sentença, em diversas situações. Logo, um pressuposto que aparentemente faz sentido – o de que conceitos relevantes tendem a se repetir ao longo das coleções – talvez precise ser reavaliado ou, no mínimo, aplicado com alguma cautela.

7.1 Contribuições

Conforme mencionado anteriormente, deve-se considerar que esta investigação foi feita em pequena escala e, em razão disso, é preferível evitar generalizações ou conclusões definitivas. Ainda assim, com base nos resultados obtidos, observou-se que a seleção de sentenças com base na frequência de seus conceitos não refletiu adequadamente a seleção de conteúdo feita por humanos em todos os textos do *corpus*, mesmo que a frequência fosse corrigida para eliminar termos comuns (usando a IDF) ou normalizada para diminuir o efeito do tamanho da sentença.

Assim sendo, uma das contribuições deste trabalho é apresentar uma hipótese sobre a SHMM que afeta diretamente a SAMM e, portanto, merece ser investigada mais a fundo: para o sumarizador humano, a seleção de conteúdo multidocumento não se resume a identificar as sentenças contendo os conceitos mais frequentes da coleção. O processo vai além disso, aparentando ser muito mais complexo. Ademais, mesmo trabalhando no nível conceitual, formalizar as estratégias de sumarização que os humanos utilizam não é uma tarefa trivial.

Sob o ponto de vista metodológico, a proposta de estudar estratégias de sumarização a partir do alinhamento sentencial mostrou-se viável e profícua, permitindo uma série de análises e comparações entre os métodos investigados. Além disso, foi possível traçar um roteiro de como realizar uma representação em UNL com os recursos disponíveis atualmente. À primeira vista, a tarefa parecia simples, uma vez, que teoricamente, existem ferramentas capazes de executá-la automaticamente. Entretanto, esse processo teve de ser feito manualmente e mostrou-se bem mais complexo. O relato dessas atividades poderá ser útil para quem venha a estudar a UNL futuramente, facilitando pesquisas que envolvam a necessidade de gerar representações nessa interlíngua.

Ressalta-se, ainda, a futura disponibilização *online*, na página do projeto SUSTENTO[47], de documentos contendo as seguintes informações: (i) anotação conceitual via UNL dos 2 textos-fonte e do sumário de referência das coleções C1, C2 e C9 do CM2News; (ii) alinhamento das sentenças dos 2 textos-fonte às sentenças do sumário manual nas coleções C1, C2 e C9; e (iii) alinhamento das representações conceituais dos textos-fonte à representação do sumário manual nas referidas coleções, o que poderá ser proveitoso para futuras investigações na área.

7.2 Limitações

Dentre as principais limitações deste projeto, destacam-se: o fato de ter envolvido um único gênero textual – no caso, o jornalístico; a utilização de apenas duas línguas para tratar de SAMM; o baixo número de textos-fonte e sumários; e o fato de as representações em UNL e os alinhamentos terem sido realizados por uma única pessoa.

O gênero textual jornalístico é muito comum na área de SAMM e de grande relevância, visto que a sumarização de notícias é uma atividade de bastante interesse. Entretanto, não se sabe se as estratégias propostas comportar-se-iam de maneira análoga em outros gêneros. Portanto, essa é uma das limitações deste trabalho.

A utilização de apenas duas línguas nos textos-fonte poderia ser considerada uma limitação menos impactante no nosso caso, visto que, teoricamente, a partir do momento em que os textos são representados em UNL, a influência das línguas-fonte seria eliminada. Entretanto, na prática, a UNL ainda não se mostra tão independente de língua quanto se deseja (GALLARDO, 2005). Em virtude disso, uma investigação com maior número de línguas faz-se necessária, visando entender melhor a influência dos idiomas-fonte.

As limitações mais sérias correm por conta do baixo número de textos analisados e pelo fato de muitos dos dados utilizados terem sido produzidos por um único indivíduo.

Torna-se difícil realizar afirmações categóricas com relação ao desempenho dos métodos de seleção de conteúdo com a quantidade de material analisada. Havendo somente um sumário manual de referência para cada coleção, a gama de estratégias utilizadas pelos sumarizadores humanos certamente foi limitada. Além disso, os sumários produzidos por mais de uma pessoa muito provavelmente seriam diferentes, o que afetaria o alinhamento entre as sentenças dos sumários e as dos textos-fonte. Em função disso, é extremamente

[47] http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23#resource

desejável que *corpora* maiores sejam investigados. Isso poderia ser útil não somente para averiguar se o comportamento dos métodos seria o mesmo, mas também buscando identificar outras estratégias de seleção de conteúdo empregadas por humanos que pudessem ser formalizadas.

Quanto à representação dos textos em UNL e a tarefa de alinhamento, ambas foram realizadas por apenas uma pessoa, o que é uma fonte de viés a ser considerada.

Para gerar as representações em UNL, faz-se a indexação dos conceitos identificados nas sentenças às UWs dos dicionários UNL. Havendo somente um indivíduo realizando a UNLização, é possível que, quando um texto em português e um texto em inglês apresentem conceitos semanticamente relacionados, haja uma tendência a indexá-los a uma mesma UW, algo que talvez não ocorresse caso pessoas diferentes estivessem fazendo a representação dos textos. Em outras palavras, um anotador humano que já tenha escolhido uma determinada UW para representar um conceito pode apresentar uma tendência a escolher a mesma UW para representar conceitos similares com os quais venha a se deparar futuramente. Nessas situações, é possível que haja uma espécie de "hiper-regularização" da representação conceitual.

Outra questão relevante relacionada à UNL é que, às vezes, sentenças que veiculam a mesma informação podem ser representadas de maneiras diferentes, embora, em um cenário ideal, isso não devesse ocorrer. Além disso, parece haver uma curva de aprendizado nessa tarefa: uma mesma pessoa, ao representar a mesma sentença, pode gerar representações diferentes, conforme seu grau de familiarização com a linguagem UNL. Essa variação nas representações UNL pode causar flutuações nas frequências dos conceitos e relações e, consequentemente, afetar de alguma forma a avaliação dos métodos conceituais investigados.

Com relação ao alinhamento, a tarefa de identificar sobreposição de conteúdo entre textos é, de certa forma, subjetiva. Se outras pessoas estivessem envolvidas nessa atividade, poderia haver divergência quanto aos casos em que se julgou haver sobreposição de conteúdo ou não. Assim, a principal fonte de informação utilizada para comparação entre os métodos – quais sentenças dos textos-fonte foram alinhadas a sentenças do sumário – provém de uma tarefa na qual existe algum grau de subjetividade. Por outro lado, em Camargo (2013), a tarefa de alinhamento mostrou-se bem definida, com baixa discordância entre os anotadores. Ainda assim, baixa não significa nula, sendo, portanto uma fonte de variação que não se deve desprezar.

7.3 Trabalhos futuros

Neste estudo, sugeriu-se que selecionar sentenças para o sumário com base na frequência de seus conceitos, realizando-se uma normalização de acordo com o tamanho da sentença, pode ser mais interessante do que selecioná-las unicamente com base na frequência de conceitos. Ainda assim, em textos do gênero jornalístico, a seleção de conteúdo com base na posição da sentença no texto-fonte seria a que mais se aproxima da seleção realizada pelo sumarizador humano, dentre as estratégias investigadas. Considerando as limitações apresentadas, é necessário confirmar esses resultados utilizando-se uma quantidade maior de dados. Portanto, um estudo em maior escala seria de grande importância para confirmar qual dessas é, de fato, a melhor estratégia de seleção de conteúdo na SAMM de textos jornalísticos.

Outra frente de investigação importante seria expandir essa análise para outros gêneros textuais, especialmente aqueles em que a posição da sentença no texto-fonte deixa de fornecer informações confiáveis sobre sua relevância.

Também pode ser interessante explorar outras formas de se utilizar as informações das representações conceituais em UNL no âmbito da SAMM. As relações binárias, por exemplo, foram usadas neste estudo unicamente de maneira quantitativa, contabilizando-se as que mais tendiam a ser preservadas nos sumários e atribuindo-se a elas um peso em função dessa característica. Entretanto, elas podem ser exploradas de outros modos. Uma sugestão seria um trabalho similar ao de Martins e Rino (2002a, 2002b), em que foram desenvolvidas heurísticas de podas de sentenças com base nas RLs, porém expandindo-o para o cenário multidocumento.

Mais uma investigação relevante a se fazer diz respeito à seleção de conteúdo intrassentencial. As estratégias de seleção de conteúdo aqui investigadas aplicam-se a sentenças inteiras, ou seja, consistem em selecionar quais sentenças nos textos-fonte são relevantes e quais não. No entanto, para que se chegue à sumarização abstrativa, é necessário desenvolver estratégias de seleção de conteúdo também no nível intrassentencial, eliminando fragmentos de sentenças considerados pouco relevantes, fundindo sentenças com conteúdo relacionado e assim por diante. Quais estratégias de sumarização humana são usadas com essa finalidade? O alinhamento sentencial e/ou conceitual permite identificar tais estratégias? É possível formalizar essas estratégias e aplicá-las às representações conceituais dos textos-fonte?

Há ainda a possibilidade de se investigar os chamados fenômenos multidocumento[48], como informações redundantes, contraditórias e complementares. Sugere-se investigar, primeiramente, se esses fenômenos se projetam na representação conceitual em UNL. Caso isso aconteça, estratégias para tratar fenômenos multidocumento poderiam ser investigadas a partir das representações conceituais em UNL.

Por fim, propõe-se uma questão mais geral para reflexão. Se não é a presença dos conceitos mais frequentes que faz uma sentença ser considerada relevante, o que é? Se não são suas informações conceituais, quais são as características das sentenças do início de um texto jornalístico que as tornam relevantes? Como selecionar automaticamente essas sentenças, não simplesmente por elas estarem localizadas no início do texto (visto que isso é a *consequência* de elas serem relevantes, e não a *causa*), mas com base em suas características intrínsecas? Aparentemente, esse ainda continua sendo um dos grandes desafios da área de SA. Aprimorando nossos conhecimentos sobre isso, as estratégias de seleção de conteúdo então desenvolvidas poderão ser utilizadas em outros gêneros nos quais as sentenças mais importantes nem sempre estão posicionadas no início do texto.

[48] Tais fenômenos ocorrem porque os documentos de uma coleção a serem sumarizados tratam do mesmo assunto, sendo que tais textos possuem origem diversificada e foram escritos em diferentes momentos.

REFERÊNCIAS BIBLIOGRÁFICAS

AKABANE, A. T.; PARDO, T. A. S.; RINO, L. H. M. Explorando medidas de redes complexas para sumarização multidocumento. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL 2011), 8. – WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TILic), 2., 2011, Cuiabá. **Proceedings...** Cuiabá: UFMT, 2011.

ALANSARY, S.; NAGI, M. From language implicit structure to UNL explicit knowledge infrastructure. In: SYMPOSIUM ON NATURAL LANGUAGE PROCESSING (SNLP2013), 10., 2013, Phuket, Thailand. **Proceedings...** Phuket, Thailand, 2013.

ALANSARY, S.; NAGI, M.; ADLY, N. UNL Editor: An annotation tool for semantic analysis. In: INTERNATIONAL CONFERENCE ON LANGUAGE ENGINEERING, 11., 2011, Cairo, Egypt. **Proceedings...** Cairo, Egypt, 2011.

BARZILAY, R.; McKEOWN, K. R.; ELHADAD, M. Information fusion in the context of multi-document summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 37., 1999, Maryland, USA. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 1999. p. 550-557.

BAXENDALE, P. B. Machine-made index for technical literature: An experiment. **IBM Journal of Research and Development**, v. 2, n. 4, p. 354-361, 1958.

BOSSARD, A.; RODRIGUES, C. Combining a multi-document update summarization system – CBSEAS – with a genetic algorithm. In: HATZILYGEROUDIS, I.; PRENTZAS, J. (Eds.). **Combinations of Intelligent Methods and Applications**. Berlin Heidelberg: Springer-Verlag, 2011. p. 71-87.

BOUDIN, F.; HUET, S.; TORRES-MORENO, J.-M. A Graph-based Approach to Cross-language Multi-document Summarization. **Polibits**, México, n. 43, p. 113-118, 2011.

CAMARGO, R. T. **Investigação de estratégias de sumarização humana multidocumento**. 2013. 133 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2013. Disponível em: <http://www.nilc.icmc.usp.br/arianidf/Camargo_Dissertacao2013.pdf>. Acesso em: 21 mai. 2014.

CARDEÑOSA, J.; GELBUKH, A.; TOVAR, E. (Eds.). **Universal Networking Language**: Advances in theory and applications. México D.F.: Instituto Politécnico Nacional, 2005. 443 p. (Special issue of *Research on Computing Science*, Volume 12).

CARDEÑOSA, J. et al. A new knowledge representation model to support multilingual ontologies. A case study. In: INTERNATIONAL CONFERENCE ON SEMANTIC WEB AND WEB SERVICES (SWWS), 2008, Monterrey, Mexico. **Proceedings...** Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 313-319.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. 2014. 180 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – USP, São Carlos, SP, 2014.

CARDOSO, P. C. F.; PARDO, T. A. S.; NUNES, M. G. V. Métodos para Sumarização Automática Multidocumento usando modelos semântico-discursivos. In: RST BRAZILIAN MEETING, 3., 2011, Cuiabá, Brazil. **Proceedings...** Cuiabá, 2011. p. 59-74.

CASELI, H. M.; NUNES, M. G. V. Corpus paralelo e corpus paralelo alinhado: Propriedades e aplicações. **Estudos Lingüísticos**, Taubaté, v. 33, p. 581-586, 2004.

CASELI, H. M.; NUNES, M. G. V. Alinhamento sentencial e lexical de córpus paralelos: Recursos para a Tradução Automática. **Estudos Lingüísticos**, São Paulo, v. 34, p. 356-361, 2005. Disponível em: <http://www.nilc.icmc.usp.br/nilc/download/CaNuEstLin05.pdf>. Acesso em: 06 mai. 2014.

CASELI, H. M.; SILVA, A. M. P.; NUNES, M. G. V. Evaluation of methods for sentence and lexical alignment of Brazilian Portuguese and English parallel texts. In: BAZZAN, A. L. C.; LABIDI, S. (Eds.). BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE (SBIA 2004, LNAI 3171), 17., 2004, São Luís, Maranhão. **Proceedings...** São Luís, Maranhão, 2004. v. 17, p.184-193.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. Experiments with CST-based multidocument summarization. In: 2010 WORKSHOP ON GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING (TEXTGRAPHS-5) (ACL2010), 2010, Uppsala, Sweden. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2010. p. 74-82.

CASTRO JORGE, M. L.; PARDO, T. A. S. A Generative approach for multi-document summarization using the noisy channel model. In: WORKSHOP "A RST E OS ESTUDOS DO TEXTO", 3., 2011, Cuiabá/MT, Brasil. **Anais**... Cuiabá: Sociedade Brasileira de Computação, 2011. p. 75-87.

CLARKE, J.; LAPATA, M. Discourse constraints for document compression. **Computational Linguistics**, v. 36, n. 3, p. 411-441, 2010.

CONROY, J. M.; O'LEARY, D. P. Text summarization via hidden Markov models. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT ON INFORMATION RETRIEVAL (SIGIR '01), 24., 2001, New Orleans, LA, USA. **Proceedings...** New York: ACM, 2001. p. 406-407.

DAVIS, R.; SHROBE, H.; SZOLOVITS, P. What is knowledge representation? **AI Magazine,** v. 14, n. 1, p. 17-33, 1993.

DI FELIPPO, A.; DIAS-DA-SILVA, B. C. O processamento automático de línguas naturais enquanto engenharia do conhecimento linguístico. **Calidoscópio**, São Leopoldo, v. 7, n. 3, p. 183-191, 2009. Disponível em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/4871/2127>. Acesso em: 14 abr. 2014.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem**: O processamento automático das línguas naturais. 1996. 272 f. Tese (Doutorado em Letras) - Universidade Estadual Paulista, Araraquara, SP, 1996. Disponível em: <http://wiki.icmc.usp.br/images/a/ad/DiasDaSilva1996.pdf>. Acesso em: 21 mai. 2014.

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

DIAS-DA-SILVA, B. C.; DI FELIPPO, A.; HASEGAWA, R. Methods and tools for encoding the WordNet.Br sentences, concept glosses, and conceptual-semantic relations. In: VIEIRA, R. et al. (Eds.). COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE: 7th INTERNATIONAL WORKSHOP, PROPOR'06, 2006, Itatiaia. **Proceedings...** Berlin: Springer Verlag, 2006, p.120-130.

DIAS-DA-SILVA, B. C.; MONTILHA, G.; RINO, L. H. M. et al. **Introdução ao Processamento das Línguas Naturais e algumas aplicações**. São Carlos, SP: ICMC-USP, 2007. 121p. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Relatório NILC-TR-07-10).

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM (JACM)**, v. 16, n. 2, p. 264-285, 1969.

EVANS, D. K.; KLAVANS, J. L.; McKEOWN, K. R. Columbia Newsblaster: Multilingual news summarization on the web. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (HLT-NAACL-2004), 2004, Boston, USA. **Proceedings...** Boston, USA, 2004. p. 1-4. Disponível em: <http://acl.ldc.upenn.edu/hlt-naacl2004/demos/ pdf/evans.pdf>. Acesso em: 10 out. 2012.

EVANS, D. K.; McKEOWN, K.; KLAVANS, J. L. **Similarity-based multilingual multi-document summarization**. New York: Columbia University, 2005. (Technical Report CUCS-014-05).

FELLBAUM, C. **WordNet**: An electronic lexical database. Cambridge, MA: MIT Press, 1998. 423 p.

GALLARDO, C. Prologue. In: CARDEÑOSA, J.; GELBUKH, A.; TOVAR, E. (Eds.). **Universal Networking Language**: Advances in theory and applications. México D.F.: Instituto Politécnico Nacional, 2005. 443 p. (Special issue of *Research on Computing Science*, Volume 12).

GRISHMAN, R. **Computational Linguistics**: An introduction. Cambridge, UK: Cambridge University Press, 1986. 193 p.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258-268, 2010.

HATZIVASSILOGLOU, V. et al. SIMFINDER: A flexible clustering tool for summarization. In: WORKSHOP ON AUTOMATIC SUMMARIZATION AT NAACL, 2001, Pittsburgh, USA. **Proceedings...** Pittsburgh, USA, 2001. p.41-49.

HENNIG, L.; UMBRATH, W.; WETZKER, R. An ontology-based approach to text summarization. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECHNOLOGY (WI-IAT '08), 2008,

Sydney, Australia. **Proceedings...** Washington, DC: Institute of Electrical and Electronics Engineers (IEEE), 2008, v. 3, p. 291-294.

JACKENDOFF, R. **Semantics and cognition**. Cambridge, Massachusetts: MIT Press, 1983.

JACKENDOFF, R. **Consciousness and the computational mind**. Cambridge, Massachusetts: MIT Press, 1985.

JURAFSKY, D; MARTIN, J. **Speech and language processing**: An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2000.

KATRAGADDA, R.; PINGALI, P.; VARMA, V. Sentence position revisited: A robust light-weight update summarization "baseline" algorithm. In: INTERNATIONAL WORKSHOP ON CROSS LINGUAL INFORMATION ACCESS: ADDRESSING THE INFORMATION NEED OF MULTILINGUAL SOCIETIES (CLIAWS3), 3., 2009, Boulder, Colorado. **Proceedings...** Boulder, Colorado: Association for Computational Linguistics, 2009. p. 46-52.

KAY, M. Preface. In: VÉRONIS, J. (Ed.). **Parallel text processing**: Alignment and use of translation corpora. Dordrecht: Kluwer Academic Publishers, 2000. 402 p.

KAY, P.; FILLMORE, C. J. Grammatical constructions and linguistic generalizations: The 'What's X doing Y?' Construction. **Language**, v. 75, p. 1-33, 1999.

KUMAR, Y. J.; SALIM, N.; RAZA, B. Cross-document structural relationship identification using supervised machine learning. **Applied Soft Computing**, v. 12, n. 10, p. 3124-3131, 2012.

LENCI, A. et al. A. Multilingual summarization by integrating linguistic resources in the MLIS-MUSI Project. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 3., 2002, Las Palmas de Gran Canaria, España. **Proceedings...** Las Palmas de Gran Canaria, España: European Language Resources Association, 2002. p. 1464-1471.

LI, L.; WANG, D.; SHEN, C.; LI, T. Ontology-enriched multi-document summarization in disaster management. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 33., 2010, Geneva. **Proceedings...** New York, NY: ACM, 2010. p. 819-820.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT (WAS-2004), POST-CONFERENCE WORKSHOP OF ACL 2004, Barcelona, Spain. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2004. p. 74-81.

LIN, C.-Y.; HOVY, E. From single to multi-document summarization: A prototype system and its evaluation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL '02), 40., 2002, Philadelphia, PA. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2002. p. 457-464.

LIN, C.-Y.; HOVY, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: HUMAN TECHNOLOGY CONFERENCE (HLT-NAACL-2003), 2003, Edmonton, Canada. **Proceedings...** Edmonton: North American Chapter of the Association for Computational Linguistics, 2003. p. 71-78.

LOUIS, A.; JOSHI, A.; NENKOVA, A. Discourse indicators for content selection in summarization. In: ANNUAL MEETING OF THE SPECIAL INTEREST GROUP ON DISCOURSE AND DIALOGUE (SIGDIAL '10), 11., 2010, Tokyo, Japan. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2010. p. 147-156.

LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, v. 39, n. 2, p. 267-300, 2013.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of research and development**, v. 2, n. 2, p. 159-165, 1958.

MANI, I. **Automatic Summarization**. Amsterdam: John Benjamins Publishing, 2001. xi + 286 p.

MANI, I.; BLOEDORN, E. Multi-document summarization by graph search and matching. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-97), 14., 1997, Providence, Rhode Island. **Proceedings…** American Association for Artificial Intelligence, 1997, p. 622-628.

MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. Cambridge, MA: MIT Press, 1999. 434 p.

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory**: A theory of text organization. Marina del Rey, CA: Information Sciences Institute, University of Southern California, 1987. (Technical Report ISI/RS-87-190).

MARCU, D. **The theory and practice of discourse parsing and summarization**. Cambridge, Massachusetts: The MIT Press, 2000.

MARTIN, P. Knowledge representation in RDF/XML, KIF, Frame-CG and Formalized-English. In: INTERNATIONAL CONFERENCE ON CONCEPTUAL STRUCTURES: INTEGRATION AND INTERFACES (ICCS 2002), 10., 2002, Borovets, Bulgaria. **Proceedings...** Berlin: Springer, 2002. p. 77-91.

MARTINS, C. B. **UNLSumm**: Um sumarizador automático de textos UNL. 2002. 100 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de São Carlos, São Carlos, SP, 2002.

MARTINS, C. B.; RINO, L. H. M. Pruning UNL texts for summarizing purposes. In: NATURAL LANGUAGE PROCESSING PACIFIC RIM SYMPOSIUM, 6., 2001, Tokyo, Japan. **Proceedings...** Tokyo, Japan, 2001. p. 539-544.

MARTINS, C. B.; RINO, L. H. M. Revisiting UNLSumm: Improvement through a case study. In: WORKSHOP ON MULTILINGUAL INFORMATION ACCESS AND

NATURAL LANGUAGE PROCESSING, 2002, Sevilha, Espanha. **Proceedings...** Sevilha: Universidad de Sevilla, 2002a. p.71-79.

MARTINS, C. B.; RINO, L. H. M. **Heurísticas de poda de sentenças para a Sumarização Automática de textos UNL**: Estudo de casos. São Carlos, SP: ICMC-USP, 2002b. 51 p. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Relatório NILC-TR-02-11).

MARTINS, R. T. **De volta ao Crátilo**. São Paulo, SP: Universidade Presbiteriana Mackenzie, 2007. (Relatório Científico). Disponível em: <http://www.ronaldomartins.pro.br/cratylus/RelatorioCientificoCratilo.pdf>. Acesso em: 17 out. 2013.

MARTINS, R. T. UNL: **Corpus**. 2008. Disponível em: <http://www.ronaldomartins.pro.br/unlx/corpus.htm>. Acesso em: 17 out. 2013.

MARTINS, R. T. Le Petit Prince in UNL. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC2012), 2012, Istanbul, Turkey. **Proceedings...** Istanbul: European Language Resources Association (ELRA), 2012. p. 3201-3204

MARTINS, R. T.; AVETISYAN, V. Generative and enumerative lexicons in the UNL framework. In: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND INFORMATION TECHNOLOGIES (CSIT2009), 7., WORKSHOP "UNL - Universal Networking Language", 2009, Yerevan, Armenia. **Proceedings...** Yerevan, Armenia, 2009.

MARTINS, R. T.; HASEGAWA, R.; NUNES, M. G. V. HERMETO: A NL analysis environment. In: WORKSHOP DA TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA - TIL'04, 2004, Salvador. **Anais...**, Salvador, 2004. p.64-71.

MARTINS, R. T.; RINO, L. H. M.; NUNES, M. G. V.; OLIVEIRA JR., O. N. The UNL distinctive features: inferences from a NL-UNL enconverting task. In: INTERNATIONAL WORKSHOP ON UNL, OTHER INTERLINGUAS AND THEIR APPLICATIONS (LREC'2002), 1., 2002, Las Palmas, Canary Islands, Spain. **Proceedings...** Las Palmas, Canary Islands, Spain, 2002. v. 1, p.8-13.

MARTINS, R. T.; RINO, L. H. M.; NUNES, M. G. V.; MONTILHA, G.; OLIVEIRA JR., O. N. An interlingua aiming at communication on the Web: How language-independent can it be? In: WORKSHOP ON APPLIED INTERLINGUAS: PRACTICAL APPLICATIONS OF INTERLINGUAL APPROACHES TO NLP (Pre-Conference Workshop in conjunction with ANLP-NAACL2000), 2000, Seattle, Washington, USA. **Proceedings...** Seattle, Washington, USA, 2000.

MAZIERO, E. G. **Identificação automática de relações multidocumento**. 2012. 106 f. Dissertação (Mestrado em Ciências - Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC-USP), São Carlos, SP, 2012.

McENERY, T. Corpus Linguistics. In: MITKOV, R. (Ed.). **The Oxford Handbook of Computational Linguistics**. New York: Oxford University Press, 2003. p. 448-463.

McKEOWN, K. R.; KLAVANS, J. L.; HATZIVASSILOGLOU, V.; BARZILAY, R.; ESKIN, E. Towards multidocument summarization by reformulation: Progress and prospects. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-99), 16., 1999, Orlando, Florida. **Proceedings...** Menlo Park, CA: American Association for Artificial Intelligence, 1999. p. 453-460.

MIHALCEA, R.; TARAU, P. A language independent algorithm for single and multiple document summarization. In: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (IJCNLP), 2005, Jeju Island, Korea. **Proceedings...** Berlin/Heidelberg: Springer-Verlag, 2005. p. 19-24.

NENKOVA, A.; PASSONNEAU, R.; MCKEOWN, K. The pyramid method: Incorporating human content selection variation in summarization evaluation. **ACM Transactions on Speech and Language Processing (TSLP)**, v. 4, n. 2, p. 1-23, 2007.

NUNES, M. G. V.; OLIVEIRA JR., O. N. O processo de desenvolvimento do revisor gramatical ReGra. In: SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE (SEMISH), 27. (CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 20.), 2000, Curitiba. **Anais...** Curitiba, 2000. v. 1, p. 6 (resumo) (artigo completo na versão em CD-ROM).

NUNES, M. G. V.; MARTINS, R. T.; RINO, L. H. M.; OLIVEIRA JR., O. N. The use of the Universal Networking Language for devising an automatic sentence generator for Brazilian Portuguese. **Cadernos de Computação**, São Carlos, v. 2(2), p. 73-97, 2001.

O'DONNELL, M. Variable-length on-line document generation. In: EUROPEAN WORKSHOP ON NATURAL LANGUAGE GENERATION, 6., 1997, Duisburg, Germany. **Proceedings...** Duisburg: Gerhard-Mercator University, 1997.

ORĂSAN, C. Automatic summarization in the informational age. In: RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING – INTERNATIONAL CONFERENCE (RANLP - 2009), 7., 2009, Borovets, Bulgaria. **Proceedings...** Stroudsburg, PA: Association on Computational Linguistics, 2009.

ORĂSAN, C.; CHIOREAN, O. A. Evaluation of a cross-lingual Romanian-English multi-document summariser. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC2008), 6., 2008, Marrakesh. **Proceedings...** Marrakesh, 2008. Disponível em: <http://clg.wlv.ac.uk/papers/539_paper.pdf>. Acesso em: 3 set. 2012.

PANDIAN, S. L.; KALPANA, S. UNL based document summarization based on level of users. **International Journal of Computer Applications**, New York, v. 66, n. 24, p. 28-36, 2013.

PARDO, T. A. S. **GistSumm – GIST SUMMarizer**: Extensões e novas funcionalidades. São Carlos: ICMC-USP, 2005. 6p. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Relatório NILC-TR-05-05).

PARDO, T. A. S. **Sumarização Automática**: Principais conceitos e sistemas para o português brasileiro. São Carlos, SP, Brasil: ICMC-USP, 2008. Disponível em: <http://www.icmc.usp.br/pessoas/taspardo/NILCTR0804-Pardo.pdf>. Acesso em: 10 abr.

2014. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional, Relatório NILC-TR-08-04).

PARDO, T. A. S.; RINO, L. H. M. DMSumm: Review and assessment. In: RANCHOD, E.; MAMEDE, N. J. (Eds.). **Advances in Natural Language Processing**. Berlin/Heidelberg: Springer-Verlag, 2002. p. 263-273. (Lecture Notes in Computer Science, v. 2389).

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. GistSumm: A summarization tool based on a new extractive method. In: MAMEDE, N. J. et al. (Eds.). WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE - WRITTEN AND SPOKEN – PROPOR, 6., 2003, Faro/Portugal. **Proceedings...** Berlin/Heidelberg: Springer-Verlag, 2003. 210-218. (Lecture Notes in Artificial Intelligence, v. 2721).

PARDO, T. A. S.; GASPERIN, C. V.; CASELI, H. M.; NUNES, M. G. V. Computational Linguistics in Brazil: An overview. In: NAACL HLT 2010 YOUNG INVESTIGATORS WORKSHOP ON COMPUTATIONAL APPROACHES TO LANGUAGES OF THE AMERICAS, 2010, Los Angeles, California, USA. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2010. p. 1-7. Disponível em: <http://dl.acm.org/citation.cfm?id=1868701.1868702>. Acesso em: 10 abr. 2014.

RADEV, D. R. A common theory of information fusion from multiple text sources: Step one: Cross-document structure. In: SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1., 2000, Hong Kong (SIGDIAL '00). **Proceedings...** (Volume 10). Stroudsburg, PA: Association for Computational Linguistics, 2000. p. 74-83.

RADEV, D. R.; McKEOWN, K. R. Generating natural language summaries from multiple on-line sources. **Computational Linguistics** (Special issue on natural language generation), v. 24, n. 3, p. 469-500, 1998.

RADEV, D. et al. MEAD - a platform for multidocument multilingual text summarization. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2004), 4., 2004, Lisbon, Portugal. **Proceedings...** Paris: ELRA, 2004.

RIBALDO, R. **Investigação de mapas de relacionamento para sumarização multidocumento**. Monografia de conclusão de curso (Bacharelado em Ciências de Computação) – USP, São Carlos, 2013.

RIBALDO, R.; PARDO, T. A. S.; RINO, L. H. M. Sumarização Automática Multidocumento com mapas de relacionamento. In: STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL 2011), 2., 2011, Cuiabá, Brazil. **Proceedings...** Cuiabá: UFMT, 2011. p. 1-3.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based methods for multi-document summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE - INTERNATIONAL CONFERENCE, 10., 2012, Coimbra, Portugal. **Proceedings...** Berlin/Heidelberg: Springer-Verlag, 2012. p. 260-271. (Lecture Notes in Computer Science, v. 7243).

RINO, L. H. M.; SENO, E. R. M. A importância do tratamento co-referencial para a sumarização automática de textos. **Estudos Lingüísticos**, São Paulo, v. 35, p. 1179-1188, 2006.

RINO, L. H. M.; PARDO, T. A. S.; SILLA JR., C. N.; KAESTNER, C. A. A.; POMBO, M. A comparison of automatic summarization systems for Brazilian Portuguese texts. In: BAZZAN, A. L. C.; LABIDI, S. (Eds.). BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE (SBIA 2004) (Lecture Notes in Artificial Intelligence 3171), 17., 2004, São Luís. **Proceedings…** São Luís, 2004. p. 235-244.

ROARK, B.; FISHER, S. OGI / OHSU baseline multilingual multi-document summarization system. In: MULTILINGUAL SUMMARIZATION EVALUATION (MSE) (ACL WORKSHOP), 2005, Michigan, USA. **Proceedings...** Michigan, USA, 2005. Disponível em: <http://www.cslu.ogi.edu/~fishers/publications/mse05.pdf>. Acesso em: 24 ago. 2012.

SAGGION, H.; LAPALME, G. Concept identification and presentation in the context of technical text summarization. In: NAACL-ANLP WORKSHOP ON AUTOMATIC SUMMARIZATION, 2000, Seattle, Washington. **Proceeedings...** Stroudsburg, Pennsylvania: Association for Computational Linguistics, 2000. p. 1-10.

SAGGION, H. et al. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2002), 3., 2002, Las Palmas, Gran Canaria, Spain. **Proceedings...** Paris: ELRA, 2002. p. 747-754.

SALTON, G. **Automatic text processing**: The transformation, analysis, and retrieval of. Reading, MA: Addison-Wesley, 1989.

SALTON, G. et al. Automatic text structuring and summarization. **Information Processing & Management**, v. 33, n. 2, p. 193-207, 1997.

SARDINHA, T. B. **Lingüística de Corpus**. Barueri, SP: Manole, 2004. 410 p.

SCHIFFMAN, B.; NENKOVA, A.; McKEOWN, A. Experiments in multi-document summarization. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH (HLT ’02), 2., 2002, San Diego, CA, USA. **Proceedings...** San Francisco, CA: Morgan Kaufmann Publishers, 2002. p.52-58.

SCHILDER, F.; KONDADADI, R. FastSum: Fast and accurate query-based multi-document summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL-08): HUMAN LANGUAGE TECHNOLOGIES, SHORT PAPERS (HLT-Short '08), 46., 2008, Columbus, Ohio. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2008. p. 205-208.

SENO, E. R. M.; NUNES, M. G. V. Automatic alignment of common information in comparable sentences of Portuguese. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL), 6., 2008, Vila Velha, Espírito Santo. **Anais...** Vila Velha, Espírito Santo, 2008. p.331-335.

SORNLERTLAMVANICH, V.; POTIPITI, T.; CHAROENPORN, T. UNL document summarization. In: INTERNATIONAL WORKSHOP ON MULTIMEDIA ANNOTATION (MMA'2001), 1., 2001, Tokyo, Japan. **Proceedings...** Tokyo, Japan, 2001.

SOUZA, C. F. R.; PEREIRA, M. B.; NUNES, M. G. V. Algoritmos de sumarização extrativa de textos em português. In: IV WORKSHOP DE COMPUTAÇÃO (WORKCOMP´2001), 4., 2001, São José dos Campos. **Anais...** São José dos Campos: Instituto Tecnológico de Aeronáutica - ITA, 2001.

SOWA, J. F. **Conceptual structures**: Information processing in mind and machine. Boston: Addison-Wesley Longman Publishing Co., Inc., 1984. 481 p.

SPARCK JONES, K. Automatic summarizing: factors and directions. In: MANI, I.; MAYBURY, M. T. (Eds.). **Advances in automatic text summarization**. Cambridge, Massachusetts: MIT Press, 1998. p.1-12.

SPARCK JONES, K.; GALLIERS, J. R. **Evaluating Natural Language Processing systems:** An analysis and review. Berlin: Springer-Verlag, 1996. 228 p.

SPECIA, L; RINO, L. H. M. **Representação semântica**: Alguns modelos ilustrativos. São Carlos, SP: ICMC-USP, 2002. 33p. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Relatório NILC-TR-02-12).

SPECIA, L.; RINO, L. H. M. Um gerador de estruturas conceituais UNL para o português. **Scientia - Revista do Programa Interdisciplinar de Pós-Graduação em Computação Aplicada**, São Leopoldo, RS, v. 14(2), p. 1-20, 2004. Disponível em: <http://www.nilc.icmc.usp.br/nilc/pessoas/specia/publications/Scientia-SpeciaRino.pdf>. Acesso em: 15 mai. 2014.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M. S. Fuzzy genetic semantic based text summarization. In: DEPENDABLE, AUTONOMIC AND SECURE COMPUTING (DASC), IEEE NINTH INTERNATIONAL CONFERENCE ON, 9., 2011, Sydney. **Proceedings...** Washington, DC: IEEE Computer Society, 2011. p. 1184-1191.

SVORE, K. M.; VANDERWENDE, L.; BURGES, C. J. C. Enhancing single-document summarization by combining RankNet and third-party sources. In: JOINT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL NATURAL LANGUAGE LEARNING (EMNLP-CoNLL), 2007, Prague, Czech Republic. **Proceedings…** Stroudsburg, PA: The Association for Computational Linguistics (ACL), 2007. p. 448-457.

TIECHER, A. L. A UNL a serviço da globalização do conhecimento. In: HOESCHL, H. C. (Org.). **UNL no Brasil**: trabalhando pela inclusão digital. Florianópolis, SC: Ijuris, 2003. p. 61-65. Disponível em: <http://www.i3g.org.br/editora/livros/unlnobrasil.pdf>. Acesso em: 12 set. 2012.

TOSTA, F. E. S. **Aplicação de conhecimento léxico-conceitual na sumarização automática multidocumento multilíngue**. 2014. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, 2014.

TOSTA, F. E. S.; DI FELIPPO, A.; PARDO, T. A. S. **Aplicação de métodos clássicos de Sumarização Automática no contexto multidocumento multilíngue**: Primeiras aproximações. São Carlos, SP: ICMC-USP, 2012. 18 p. Disponível em: <http://www.nilc.icmc.usp.br/arianidf/NILC-TR-12-02-TostaEtAl.pdf>. Acesso em: 06 mai. 2013. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Relatório NILC-TR-12-02).

TOSTA, F. E. S.; DI FELIPPO, A; PARDO, T. A. S. Estudo de métodos clássicos de sumarização no cenário multidocumento multilíngue. In: WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TILIC), 3., 2013, Fortaleza. **Anais...** Fortaleza, 2013. p. 34-36. Disponível em: <http://www.icmc.usp.br/pessoas/taspardo/TILic2013-TostaEtAl.pdf>. Acesso em: 17 abr. 2014.

UCHIDA, H.; ZHU, M.; DELLA SENTA, T. **The UNL, a gift for a millennium**. Tokyo: The United Nations University - Institute of Advanced Studies, 1999. Disponível em: <http://www.undl.org/publications/gm/index.htm>. Acesso em: 4 nov. 2013.

UNDL FOUNDATION. Universal Networking Language (UNL). Especificações. Versão 3.0. Tradução: Geraldo Macedo. In: HOESCHL, H. C. (Org.). **UNL no Brasil**: Trabalhando pela inclusão digital. Florianópolis, SC: Ijuris, 2003. Disponível em: <http://www.i3g.org.br/editora/livros/unlnobrasil.pdf>. Acesso em: 12 set. 2012. p. 3-60.

UNDL FOUNDATION. **Universal Networking Language (UNL) Specifications - Version 2005**. 2005. Disponível em: <http://www.undl.org/unlsys/unl/unl2005/>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **UNL Documents**. 2006. Disponível em: <http://www.undl.org/ unldoc/>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **UNL-EOLSS**. 2009. Disponível em: <http://www.undlfoundation.org/eolss/>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **UNL Documents of UNL-EOLSS Project**. 2010. Disponível em: <http://www.undl.org/unl-eolss/unldoc.html>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **List of UNL corpora**. 2012a. Disponível em: <http://www.unlweb.net/ wiki/List_of_UNL_Corpora>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **LPP** [Le Petit Prince - UNL Wiki]. 2012b. Disponível em: <http://www.unlweb.net/wiki/LPP>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **IGLU**. 2012c. Disponível em: <http://www.unlweb.net/wiki/IGLU>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **UNL Reference Corpus**. 2012d. Disponível em: <http://www.unlweb.net/wiki/UC>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **Introduction to UNL**. 2013a. Disponível em: <http://www.unlweb.net/wiki/index.php?title=Introduction_to_UNL>. Acesso em: 16 out. 2013.

UNDL FOUNDATION. **UNLization**. 2013b. Disponível em: <http://www.unlweb.net/wiki/UNLization>. Acesso em: 16 out. 2013.

UNDL FOUNDATION. **UNLdev**. 2013c. Disponível em: <http://dev.undlfoundation.org/index.jsp>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **IAN**. 2013d. Disponível em: <http://www.unlweb.net/wiki/IAN>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **UNL2010**. 2013e. Disponível em: <http://www.unlweb.net/wiki/UNL2010>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **Universal Words**. 2013f. Disponível em: <http://www.unlweb.net/wiki/Universal_Words>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. **Scope**. 2013g. Disponível em: <http://www.unlweb.net/wiki/Scope>. Acesso em: 17 out. 2013.

UNDL FOUNDATION. [**UNLarium - Portuguese Dictionary**]. 2014a. Disponível em: <http://www.unlweb.net/unlarium/index.php?lang=pt>. Acesso em: 27 mar. 2014.

UNDL FOUNDATION. **Grammar**. 2014b. Disponível em: <http://www.unlweb.net/wiki/Grammar>. Acesso em: 27 mar. 2014.

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 (TSLP)**, v. 6, n. 4, p. 1-20, 2010.

WAN, X. An exploration of document impact on graph-based multi-document summarization. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP '08), 2008, Waikiki, Honolulu, Hawaii. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2008. p. 755-762.

WAN, X.; LI, H.; XIAO, J. Cross-language document summarization based on machine translation quality prediction. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Uppsala, Sweden. **Proceedings...** Stroudsburg, PA: Association for Computational Linguistics, 2010. p. 917-926.

WELTY, C. A. **An integrated representation for software development and discovery**. 1995. Tese (PhD em Ciências da Computação) - Rensselaer Polytechnic Institute, Troy, NY, 1995. Disponível em: < http://www.cs.vassar.edu/~weltyc/papers/phd/HTML/dissertation-1.html>. Acesso em: 20 jan. 2015.

WHITE, J.; DOYON, J.; TALBOTT, S. Task tolerance of MT output in integrated text processes. In: ANLP-NAACL 2000 WORKSHOP: EMBEDDED MT SYSTEMS

WORKSHOP, 2000, Seattle, WA. **Proceedings…** Stroudsburg, PA: Association for Computational Linguistics, 2000. p. 9-16

WU, C.-W.; LIU, C.-L. Ontology-based text summarization for business news articles. **Computers and Their Applications**, v. 2003, p. 389-392, 2003.

APÊNDICE A – Resultados do ranqueamento de sentenças

Tabela 4 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método F(UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:23]	[S:07]	20	97	10,0
[S:34]	[S:06]	27	87	8,9
[S:29]	[S:08]	18	86	8,8
[S:01]	[S:06]	17	84	8,6
[S:44]	[S:06]	26	80	8,2
[S:18]	-	31	76	7,7
[S:32]	[S:06]	19	75	7,6
[S:38]	[S:06]	18	60	6,0
[S:36]	[S:06]	13	55	5,5
[S:26]	-	19	52	5,2
[S:02]	[S:06]	10	51	5,1
[S:00]	-	12	51	5,1
[S:40]	[S:06]	13	48	4,7
[S:17]	-	15	46	4,5
[S:24]	-	16	46	4,5
[S:11]	-	23	46	4,5
[S:05]	-	14	44	4,3
[S:41]	[S:06]	8	42	4,1
[S:42]	-	17	42	4,1
[S:43]	-	13	41	4,0
[S:15]	-	12	29	2,7
[S:39]	[S:06]	8	26	2,4
[S:33]	-	14	24	2,2
[S:35]	[S:06]	5	23	2,0
[S:13]	-	17	23	2,0
[S:03]	-	8	22	1,9
[S:04]	-	6	21	1,8
[S:06]	-	5	18	1,5
[S:37]	-	4	17	1,4
[S:09]	-	10	15	1,2
[S:07]	-	6	12	0,9
[S:31]	-	9	12	0,9
[S:10]		7	10	0,6
[S:25]	-	4	8	0,4
[S:30]	[S:08]	7	8	0,4
[S:16]	-	2	4	0,0

Fonte: Elaborado pelo autor

Tabela 5 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método F(UWs) * IDF (UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:44]	[S:06]	26	17,46	10,0
[S:01]	[S:06]	17	12,04	6,7
[S:38]	[S:06]	18	11,74	6,5
[S:32]	[S:06]	19	11,44	6,3
[S:18]	-	31	11,44	6,3
[S:02]	[S:06]	10	10,24	5,5
[S:23]	[S:07]	20	10,24	5,5
[S:11]	-	23	9,63	5,2
[S:26]	-	19	9,33	5,0
[S:42]	-	17	9,03	4,8
[S:34]	[S:06]	27	9,03	4,8
[S:36]	[S:06]	13	8,73	4,6
[S:29]	[S:08]	18	8,13	4,3
[S:40]	[S:06]	13	7,83	4,1
[S:33]	-	14	7,22	3,7
[S:39]	[S:06]	8	6,62	3,3
[S:15]	-	12	6,32	3,1
[S:17]	-	15	6,02	3,0
[S:24]	-	16	6,02	3,0
[S:31]	-	9	5,72	2,8
[S:13]	-	17	5,72	2,8
[S:37]	-	4	4,52	2,0
[S:09]	-	10	4,21	1,8
[S:35]	[S:06]	5	3,91	1,7
[S:03]	-	8	3,61	1,5
[S:43]	-	13	3,61	1,5
[S:10]	-	7	3,01	1,1
[S:41]	[S:06]	8	3,01	1,1
[S:00]	-	12	3,01	1,1
[S:07]	-	6	2,71	0,9
[S:25]	-	4	2,41	0,7
[S:30]	[S:08]	7	2,11	0,6
[S:05]	-	14	2,11	0,6
[S:16]	-	2	1,20	0,0
[S:06]	-	5	1,20	0,0
[S:04]	-	6	1,20	0,0

Fonte: Elaborado pelo autor

Tabela 6 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método F(UWs) / n. de UWs

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:41]	[S:06]	8	5,25	10,0
[S:02]	[S:06]	10	5,10	9,6
[S:01]	[S:06]	17	4,94	9,2
[S:23]	[S:07]	20	4,85	9,0
[S:29]	[S:08]	18	4,78	8,9
[S:35]	[S:06]	5	4,60	8,4
[S:37]	-	4	4,25	7,6
[S:00]	-	12	4,25	7,6
[S:36]	[S:06]	13	4,23	7,5
[S:32]	[S:06]	19	3,95	6,8
[S:40]	[S:06]	13	3,69	6,2
[S:06]	-	5	3,60	6,0
[S:04]	-	6	3,50	5,7
[S:38]	[S:06]	18	3,33	5,3
[S:39]	[S:06]	8	3,25	5,1
[S:34]	[S:06]	27	3,22	5,1
[S:43]	-	13	3,15	4,9
[S:05]	-	14	3,14	4,9
[S:44]	[S:06]	26	3,08	4,7
[S:17]	-	15	3,07	4,7
[S:24]	-	16	2,88	4,2
[S:03]	-	8	2,75	3,9
[S:26]	-	19	2,74	3,9
[S:42]	-	17	2,47	3,2
[S:18]	-	31	2,45	3,2
[S:15]	-	12	2,42	3,1
[S:16]	-	2	2,00	2,1
[S:25]	-	4	2,00	2,1
[S:07]	-	6	2,00	2,1
[S:11]	-	23	2,00	2,1
[S:33]	-	14	1,71	1,4
[S:09]	-	10	1,50	0,9
[S:10]	-	7	1,43	0,7
[S:13]	-	17	1,35	0,5
[S:31]	-	9	1,33	0,5
[S:30]	[S:08]	7	1,14	0,0

Fonte: Elaborado pelo autor

Tabela 7 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método Posição no texto-fonte

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Posição no texto-fonte	Pontuação - Valor normalizado
[S:00]	-	12	1	10,0
[S:01]	[S:06]	17	2	9,7
[S:02]	[S:06]	10	3	9,4
[S:03]	-	8	4	9,1
[S:04]	-	6	5	8,9
[S:05]	-	14	6	8,6
[S:06]	-	5	7	8,3
[S:07]	-	6	8	8,0
[S:09]	-	10	9	7,7
[S:10]	-	7	10	7,4
[S:11]	-	23	11	7,1
[S:13]	-	17	12	6,9
[S:15]	-	12	13	6,6
[S:16]	-	2	14	6,3
[S:17]	-	15	15	6,0
[S:18]	-	31	16	5,7
[S:23]	[S:07]	20	17	5,4
[S:24]	-	16	18	5,1
[S:25]	-	4	19	4,9
[S:26]	-	19	20	4,6
[S:29]	[S:08]	18	21	4,3
[S:30]	[S:08]	7	22	4,0
[S:31]	-	9	23	3,7
[S:32]	[S:06]	19	24	3,4
[S:33]	-	14	25	3,1
[S:34]	[S:06]	27	26	2,9
[S:35]	[S:06]	5	27	2,6
[S:36]	[S:06]	13	28	2,3
[S:37]	-	4	29	2,0
[S:38]	[S:06]	18	30	1,7
[S:39]	[S:06]	8	31	1,4
[S:40]	[S:06]	13	32	1,1
[S:41]	[S:06]	8	33	0,9
[S:42]	-	17	34	0,6
[S:43]	-	13	35	0,3
[S:44]	[S:06]	26	36	0,0

Fonte: Elaborado pelo autor

Tabela 8 – Ranque de sentenças na coleção C1 – Texto-fonte em português – Método F(UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:01]	[S:02]	22	61	10,0
[S:17]	-	22	58	9,2
[S:16]	-	21	51	7,2
[S:07]	-	23	50	6,9
[S:04]	[S:00] / [S:04]	17	45	5,6
[S:10]	[S:07]	18	45	5,6
[S:08]	-	20	45	5,6
[S:02]	-	23	45	5,6
[S:06]	[S:01]	24	44	5,3
[S:12]	-	18	41	4,4
[S:09]	[S:05]	14	39	3,9
[S:15]	-	15	37	3,3
[S:11]	-	10	35	2,8
[S:14]	-	14	34	2,5
[S:03]	[S:03]	15	31	1,7
[S:18]	-	12	28	0,8
[S:00]	[S:02]	13	25	0,0

Fonte: Elaborado pelo autor

Tabela 9 – Ranque de sentenças na coleção C1 – Texto-fonte em português – Método F(UWs) * IDF (UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:06]	[S:01]	24	9,33	10,0
[S:12]	-	18	7,53	7,6
[S:07]	-	23	7,53	7,6
[S:08]	-	20	6,62	6,4
[S:01]	[S:02]	22	6,62	6,4
[S:02]	-	23	6,62	6,4
[S:03]	[S:03]	15	6,32	6,0
[S:00]	[S:02]	13	6,02	5,6
[S:17]	-	22	5,72	5,2
[S:09]	[S:05]	14	5,42	4,8
[S:04]	[S:04] / [S:00]	17	5,12	4,4
[S:14]	-	14	4,21	3,2
[S:15]	-	15	3,61	2,4
[S:16]	-	21	3,31	2,0
[S:11]	-	10	2,11	0,4
[S:18]	-	12	2,11	0,4
[S:10]	[S:07]	18	1,81	0,0

Fonte: Elaborado pelo autor

Tabela 10 – Ranque de sentenças na coleção C1– Texto-fonte em português – Método F(UWs) / n. de UWs

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:11]	-	10	3,50	10,0
[S:09]	[S:05]	14	2,79	5,7
[S:01]	[S:02]	22	2,77	5,6
[S:04]	[S:04] / [S:00]	17	2,65	4,9
[S:17]	-	22	2,64	4,8
[S:10]	[S:07]	18	2,50	4,0
[S:15]	-	15	2,47	3,8
[S:14]	-	14	2,43	3,6
[S:16]	-	21	2,43	3,6
[S:18]	-	12	2,33	3,0
[S:12]	-	18	2,28	2,7
[S:08]	-	20	2,25	2,5
[S:07]	-	23	2,17	2,0
[S:03]	[S:03]	15	2,07	1,4
[S:02]	-	23	1,96	0,7
[S:00]	[S:02]	13	1,92	0,5
[S:06]	[S:01]	24	1,83	0,0

Fonte: Elaborado pelo autor

Tabela 11 – Ranque de sentenças na coleção C1 – Texto-fonte em português – Método Posição no texto-fonte

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Posição no texto-fonte	Pontuação - Valor normalizado
[S:00]	[S:02]	13	1	10,0
[S:01]	[S:02]	22	2	9,4
[S:02]	-	23	3	8,8
[S:03]	[S:03]	15	4	8,1
[S:04]	[S:04] / [S:00]	17	5	7,5
[S:06]	[S:01]	24	6	6,9
[S:07]	-	23	7	6,3
[S:08]	-	20	8	5,6
[S:09]	[S:05]	14	9	5,0
[S:10]	[S:07]	18	10	4,4
[S:11]	-	10	11	3,8
[S:12]	-	18	12	3,1
[S:14]	-	14	13	2,5
[S:15]	-	15	14	1,9
[S:16]	-	21	15	1,3
[S:17]	-	22	16	0,6
[S:18]	-	12	17	0,0

Fonte: Elaborado pelo autor

Tabela 12 – Ranque de sentenças na coleção C2 – Texto-fonte em inglês – Método F(UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:13]	-	17	109	10,0
[S:11]	-	18	96	8,6
[S:06]	-	14	75	6,4
[S:05]	-	8	63	5,1
[S:10]	[S:03]	14	57	4,5
[S:03]	[S:02]	17	53	4,0
[S:04]	-	12	46	3,3
[S:00]	[S:00]	10	34	2,0
[S:08]	[S:02]	16	32	1,8
[S:07]	[S:00]	12	26	1,2
[S:01]	[S:01]	6	24	1,0
[S:02]	[S:01]	10	19	0,4
[S:09]	[S:03]	7	15	0,0

Fonte: Elaborado pelo autor

Tabela 13 – Ranque de sentenças na coleção C2 – Texto-fonte em inglês – Método F(UWs) * IDF (UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:13]	-	17	17,76	10,0
[S:11]	-	18	17,46	9,8
[S:06]	-	14	11,14	5,7
[S:10]	[S:03]	14	10,24	5,1
[S:03]	[S:02]	17	8,43	3,9
[S:05]	-	8	7,83	3,5
[S:04]	-	12	5,12	1,8
[S:09]	[S:03]	7	4,52	1,4
[S:08]	[S:02]	16	4,52	1,4
[S:02]	[S:01]	10	3,91	1,0
[S:07]	[S:00]	12	3,31	0,6
[S:00]	[S:00]	10	2,71	0,2
[S:01]	[S:01]	6	2,41	0,0

Fonte: Elaborado pelo autor

Tabela 14 – Ranque de sentenças na coleção C2 – Texto-fonte em inglês – Método F(UWs) / n. de UWs

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:05]	-	8	7,88	10,0
[S:13]	-	17	6,41	7,6
[S:06]	-	14	5,36	5,8
[S:11]	-	18	5,33	5,7
[S:10]	[S:03]	14	4,07	3,6
[S:01]	[S:01]	6	4,00	3,5
[S:04]	-	12	3,83	3,2
[S:00]	[S:00]	10	3,40	2,5
[S:03]	[S:02]	17	3,12	2,0
[S:07]	[S:00]	12	2,17	0,4
[S:09]	[S:03]	7	2,14	0,4
[S:08]	[S:02]	16	2,00	0,2
[S:02]	[S:01]	10	1,90	0,0

Fonte: Elaborado pelo autor

Tabela 15 – Ranque de sentenças na coleção C2 – Texto-fonte em inglês – Método Posição no texto-fonte

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Posição no texto-fonte	Pontuação - Valor normalizado
[S:00]	[S:00]	10	1	10,0
[S:01]	[S:01]	6	2	9,2
[S:02]	[S:01]	10	3	8,3
[S:03]	[S:02]	17	4	7,5
[S:04]	-	12	5	6,7
[S:05]	-	8	6	5,8
[S:06]	-	14	7	5,0
[S:07]	[S:00]	12	8	4,2
[S:08]	[S:02]	16	9	3,3
[S:09]	[S:03]	7	10	2,5
[S:10]	[S:03]	14	11	1,7
[S:11]	-	18	12	0,8
[S:13]	-	17	13	0,0

Fonte: Elaborado pelo autor

Tabela 16 – Ranque de sentenças na coleção C2 – Texto-fonte em português – Método F(UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:00]	[S:00]	27	53	10,0
[S:05]	[S:02]	24	44	8,0
[S:03]	-	26	42	7,5
[S:01]	[S:01]	19	33	5,5
[S:06]	-	16	25	3,6
[S:07]	-	10	17	1,8
[S:04]	-	11	17	1,8
[S:02]	[S:02]	8	16	1,6
[S:10]	-	12	12	0,7
[S:09]	[S:02]	10	11	0,5
[S:08]	-	8	9	0,0

Fonte: Elaborado pelo autor

Tabela 17 – Ranque de sentenças na coleção C2 – Texto-fonte em português – Método F(UWs) * IDF (UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:03]	-	26	9,03	10,0
[S:05]	[S:02]	24	5,42	5,4
[S:00]	[S:00]	27	5,12	5,0
[S:04]	-	11	4,82	4,6
[S:01]	[S:01]	19	4,82	4,6
[S:06]	-	16	4,21	3,8
[S:10]	-	12	3,91	3,5
[S:07]	-	10	3,91	3,5
[S:08]	-	8	2,71	1,9
[S:02]	[S:02]	8	2,11	1,2
[S:09]	[S:02]	10	1,20	0,0

Fonte: Elaborado pelo autor

Tabela 18 – Ranque de sentenças na coleção C2 – Texto-fonte em português – Método F(UWs) / n. de UWs

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:02]	[S:02]	8	2,00	10,0
[S:00]	[S:00]	27	1,96	9,6
[S:05]	[S:02]	24	1,83	8,3
[S:01]	[S:01]	19	1,74	7,4
[S:07]	-	10	1,70	7,0
[S:03]	-	26	1,62	6,2
[S:06]	-	16	1,56	5,6
[S:04]	-	11	1,55	5,5
[S:08]	-	8	1,13	1,3
[S:09]	[S:02]	10	1,10	1,0
[S:10]	-	12	1,00	0,0

Fonte: Elaborado pelo autor

Tabela 19 – Ranque de sentenças na coleção C2 – Texto-fonte em português – Método Posição no texto-fonte

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Posição no texto-fonte	Pontuação - Valor normalizado
[S:00]	[S:00]	27	1	10,0
[S:01]	[S:01]	19	2	9,0
[S:02]	[S:02]	8	3	8,0
[S:03]	-	26	4	7,0
[S:04]	-	11	5	6,0
[S:05]	[S:02]	24	6	5,0
[S:06]	-	16	7	4,0
[S:07]	-	10	8	3,0
[S:08]	-	8	9	2,0
[S:09]	[S:02]	10	10	1,0
[S:10]	-	12	11	0,0

Fonte: Elaborado pelo autor

Tabela 20 – Ranque de sentenças na coleção C9 – Texto-fonte em inglês – Método F(UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:19]	-	20	127	10,0
[S:25]	-	21	120	9,4
[S:16]	-	26	109	8,6
[S:37]	[S:09]	20	75	5,8
[S:06]	-	14	68	5,3
[S:15]	-	23	68	5,3
[S:39]	-	22	58	4,5
[S:03]	[S:01]	17	54	4,2
[S:01]	[S:00]	12	50	3,8
[S:12]	[S:08]	15	47	3,6
[S:35]	-	16	43	3,3
[S:14]	[S:07]	12	41	3,1
[S:31]	-	17	41	3,1
[S:38]	-	8	38	2,9
[S:32]	-	20	36	2,7
[S:00]	-	11	31	2,3
[S:23]	-	6	30	2,2
[S:24]	-	6	30	2,2
[S:33]	-	11	28	2,1
[S:13]	[S:09]	15	28	2,1
[S:08]	-	10	27	2,0
[S:04]	[S:03]	9	26	1,9
[S:09]	[S:05] /[S:06]	10	26	1,9
[S:41]	-	12	26	1,9
[S:30]	[S:01] / [S:02]	11	23	1,7
[S:29]	-	12	22	1,6
[S:07]	-	11	21	1,5
[S:05]	[S:04]	8	17	1,2
[S:36]	[S:09]	9	17	1,2
[S:10]	-	6	16	1,1
[S:11]	-	6	16	1,1
[S:02]	-	4	12	0,8
[S:34]	-	2	2	0,0

Fonte: Elaborado pelo autor

Tabela 21 – Ranque de sentenças na coleção C9 – Texto-fonte em inglês – Método F(UWs) * IDF (UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:16]	-	26	24,38	10,0
[S:19]	-	20	12,34	5,1
[S:15]	-	23	8,73	3,6
[S:35]	-	16	8,43	3,5
[S:39]	-	22	7,53	3,1
[S:25]	-	21	6,32	2,6
[S:03]	[S:01]	17	5,42	2,2
[S:31]	-	17	5,42	2,2
[S:32]	-	20	5,42	2,2
[S:12]	[S:08]	15	5,12	2,1
[S:37]	[S:09]	20	5,12	2,1
[S:41]	-	12	4,82	2,0
[S:24]	-	6	4,21	1,7
[S:36]	[S:09]	9	3,61	1,5
[S:29]	-	12	3,61	1,5
[S:23]	-	6	3,01	1,2
[S:08]	-	10	3,01	1,2
[S:13]	[S:09]	15	3,01	1,2
[S:38]	-	8	2,41	1,0
[S:14]	[S:07]	12	2,41	1,0
[S:05]	[S:04]	8	2,11	0,9
[S:00]	-	11	2,11	0,9
[S:06]	-	14	2,11	0,9
[S:30]	[S:01] / [S:02]	11	1,81	0,7
[S:33]	-	11	1,81	0,7
[S:04]	[S:03]	9	1,51	0,6
[S:02]	-	4	1,20	0,5
[S:10]	-	6	1,20	0,5
[S:09]	[S:05] /[S:06]	10	1,20	0,5
[S:07]	-	11	1,20	0,5
[S:34]	-	2	0,60	0,2
[S:01]	[S:00]	12	0,60	0,2
[S:11]	-	6	0,00	0,0

Fonte: Elaborado pelo autor

Tabela 22 – Ranque de sentenças na coleção C9 – Texto-fonte em inglês – Método F(UWs) / n. de UWs

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:19]	-	20	6,35	10,0
[S:25]	-	21	5,71	8,8
[S:23]	-	6	5,00	7,5
[S:24]	-	6	5,00	7,5
[S:06]	-	14	4,86	7,2
[S:38]	-	8	4,75	7,0
[S:16]	-	26	4,19	6,0
[S:01]	[S:00]	12	4,17	5,9
[S:37]	[S:09]	20	3,75	5,1
[S:14]	[S:07]	12	3,42	4,5
[S:03]	[S:01]	17	3,18	4,1
[S:12]	[S:08]	15	3,13	4,0
[S:02]	-	4	3,00	3,7
[S:15]	-	23	2,96	3,7
[S:04]	[S:03]	9	2,89	3,5
[S:00]	-	11	2,82	3,4
[S:08]	-	10	2,70	3,2
[S:35]	-	16	2,69	3,2
[S:10]	-	6	2,67	3,1
[S:11]	-	6	2,67	3,1
[S:39]	-	22	2,64	3,1
[S:09]	[S:05] /[S:06]	10	2,60	3,0
[S:33]	-	11	2,55	2,9
[S:31]	-	17	2,41	2,6
[S:41]	-	12	2,17	2,2
[S:05]	[S:04]	8	2,13	2,1
[S:30]	[S:01] / [S:02]	11	2,09	2,0
[S:07]	-	11	1,91	1,7
[S:36]	[S:09]	9	1,89	1,7
[S:13]	[S:09]	15	1,87	1,6
[S:29]	-	12	1,83	1,6
[S:32]	-	20	1,80	1,5
[S:34]	-	2	1,00	0,0

Fonte: Elaborado pelo autor

Tabela 23 – Ranque de sentenças na coleção C9 – Texto-fonte em inglês – Método Posição no texto-fonte

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Posição no texto-fonte	Pontuação - Valor normalizado
[S:00]	-	11	1	10,0
[S:01]	[S:00]	12	2	9,7
[S:02]	-	4	3	9,4
[S:03]	[S:01]	17	4	9,1
[S:04]	[S:03]	9	5	8,8
[S:05]	[S:04]	8	6	8,4
[S:06]	-	14	7	8,1
[S:07]	-	11	8	7,8
[S:08]	-	10	9	7,5
[S:09]	[S:05] /[S:06]	10	10	7,2
[S:10]	-	6	11	6,9
[S:11]	-	6	12	6,6
[S:12]	[S:08]	15	13	6,3
[S:13]	[S:09]	15	14	5,9
[S:14]	[S:07]	12	15	5,6
[S:15]	-	23	16	5,3
[S:16]	-	26	17	5,0
[S:19]	-	20	18	4,7
[S:23]	-	6	19	4,4
[S:24]	-	6	20	4,1
[S:25]	-	21	21	3,8
[S:29]	-	12	22	3,4
[S:30]	[S:01] / [S:02]	11	23	3,1
[S:31]	-	17	24	2,8
[S:32]	-	20	25	2,5
[S:33]	-	11	26	2,2
[S:34]	-	2	27	1,9
[S:35]	-	16	28	1,6
[S:36]	[S:09]	9	29	1,3
[S:37]	[S:09]	20	30	0,9
[S:38]	-	8	31	0,6
[S:39]	-	22	32	0,3
[S:41]	-	12	33	0,0

Fonte: Elaborado pelo autor

Tabela 24 – Ranque de sentenças na coleção C9 – Texto-fonte em português – Método F(UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:14]	-	16	61	10,0
[S:21]	-	15	51	8,2
[S:13]	-	19	51	8,2
[S:22]	-	14	50	8,1
[S:00]	[S:00]	19	46	7,4
[S:25]	-	12	43	6,8
[S:04]	[S:07]	15	43	6,8
[S:05]	[S:07]	15	39	6,1
[S:27]	-	11	35	5,4
[S:18]	-	11	34	5,3
[S:07]	[S:01]	18	34	5,3
[S:17]	-	15	31	4,7
[S:06]	[S:05]	12	30	4,6
[S:01]	[S:00]	10	29	4,4
[S:24]	-	10	29	4,4
[S:20]	[S:09]	14	28	4,2
[S:16]	-	8	25	3,7
[S:26]	-	14	24	3,5
[S:10]	-	10	22	3,2
[S:03]	[S:01]	12	22	3,2
[S:19]	[S:09]	12	22	3,2
[S:15]	[S:06]	11	20	2,8
[S:02]	-	5	12	1,4
[S:28]	-	5	9	0,9
[S:09]	[S:01] / [S:02]	4	4	0,0

Fonte: Elaborado pelo autor

Tabela 25 – Ranque de sentenças na coleção C9 – Texto-fonte em português – Método F(UWs) * IDF (UWs)

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:22]	-	14	9,93	10,0
[S:21]	-	15	6,62	6,6
[S:00]	[S:00]	19	4,52	4,4
[S:17]	-	15	3,61	3,4
[S:07]	[S:01]	18	3,61	3,4
[S:04]	[S:07]	15	3,31	3,1
[S:15]	[S:06]	11	3,01	2,8
[S:13]	-	19	2,71	2,5
[S:16]	-	8	2,41	2,2
[S:27]	-	11	2,41	2,2
[S:03]	[S:01]	12	2,41	2,2
[S:26]	-	14	2,41	2,2
[S:05]	[S:07]	15	2,41	2,2
[S:24]	-	10	2,11	1,9
[S:20]	[S:09]	14	2,11	1,9
[S:01]	[S:00]	10	1,81	1,6
[S:18]	-	11	1,81	1,6
[S:14]	-	16	1,81	1,6
[S:02]	-	5	1,51	1,3
[S:06]	[S:05]	12	1,51	1,3
[S:10]	-	10	1,20	0,9
[S:09]	[S:01] / [S:02]	4	0,60	0,3
[S:28]	-	5	0,60	0,3
[S:25]	-	12	0,60	0,3
[S:19]	[S:09]	12	0,30	0,0

Fonte: Elaborado pelo autor

Tabela 26 – Ranque de sentenças na coleção C9– Texto-fonte em português – Método F(UWs) / n. de UWs

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Pontuação - Valor bruto	Pontuação - Valor normalizado
[S:14]	-	16	3,81	10,0
[S:25]	-	12	3,58	9,2
[S:22]	-	14	3,57	9,1
[S:21]	-	15	3,40	8,5
[S:27]	-	11	3,18	7,8
[S:16]	-	8	3,13	7,6
[S:18]	-	11	3,09	7,4
[S:01]	[S:00]	10	2,90	6,8
[S:24]	-	10	2,90	6,8
[S:04]	[S:07]	15	2,87	6,6
[S:13]	-	19	2,68	6,0
[S:05]	[S:07]	15	2,60	5,7
[S:06]	[S:05]	12	2,50	5,3
[S:00]	[S:00]	19	2,42	5,1
[S:02]	-	5	2,40	5,0
[S:10]	-	10	2,20	4,3
[S:17]	-	15	2,07	3,8
[S:20]	[S:09]	14	2,00	3,6
[S:07]	[S:01]	18	1,89	3,2
[S:03]	[S:01]	12	1,83	3,0
[S:19]	[S:09]	12	1,83	3,0
[S:15]	[S:06]	11	1,82	2,9
[S:28]	-	5	1,80	2,8
[S:26]	-	14	1,71	2,5
[S:09]	[S:01] / [S:02]	4	1,00	0,0

Fonte: Elaborado pelo autor

Tabela 27 – Ranque de sentenças na coleção C9 – Texto-fonte em português – Método Posição no texto-fonte

Sentença do texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Tamanho da sentença do texto-fonte (quantidade de UWs)	Posição no texto-fonte	Pontuação - Valor normalizado
[S:00]	[S:00]	19	1	10,0
[S:01]	[S:00]	10	2	9,6
[S:02]	-	5	3	9,2
[S:03]	[S:01]	12	4	8,8
[S:04]	[S:07]	15	5	8,3
[S:05]	[S:07]	15	6	7,9
[S:06]	[S:05]	12	7	7,5
[S:07]	[S:01]	18	8	7,1
[S:09]	[S:01] / [S:02]	4	9	6,7
[S:10]	-	10	10	6,3
[S:13]	-	19	11	5,8
[S:14]	-	16	12	5,4
[S:15]	[S:06]	11	13	5,0
[S:16]	-	8	14	4,6
[S:17]	-	15	15	4,2
[S:18]	-	11	16	3,8
[S:19]	[S:09]	12	17	3,3
[S:20]	[S:09]	14	18	2,9
[S:21]	-	15	19	2,5
[S:22]	-	14	20	2,1
[S:24]	-	10	21	1,7
[S:25]	-	12	22	1,3
[S:26]	-	14	23	0,8
[S:27]	-	11	24	0,4
[S:28]	-	5	25	0,0

Fonte: Elaborado pelo autor

Tabela 28 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método RLs

Sentença texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Pontuação normalizada
[S:24]	-	10,00
[S:26]	-	9,31
[S:43]	-	7,59
[S:40]	[S:06]	7,08
[S:44]	[S:06]	6,84
[S:34]	[S:06]	6,80
[S:39]	[S:06]	6,52
[S:15]	-	5,94
[S:25]	-	5,85
[S:00]	-	5,80
[S:41]	[S:06]	5,72
[S:06]	-	5,71
[S:37]	-	5,70
[S:10]	-	5,64
[S:13]	-	5,53
[S:30]	[S:08]	5,43
[S:33]	-	5,36
[S:04]	-	5,27
[S:31]	-	4,91
[S:35]	[S:06]	4,72
[S:11]	-	4,61
[S:23]	[S:07]	4,48
[S:09]	-	4,46
[S:07]	-	4,43
[S:03]	-	4,43
[S:18]	-	4,23
[S:02]	[S:06]	4,00
[S:38]	[S:06]	3,92
[S:16]	-	3,58
[S:36]	[S:06]	3,28
[S:05]	-	3,28
[S:29]	[S:08]	3,23
[S:32]	[S:06]	3,21
[S:01]	[S:06]	2,89
[S:17]	-	1,33
[S:42]	-	0,00

Fonte: Elaborado pelo autor

Tabela 29 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método RLs + UWs 1:1

Sentença texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Pontuação normalizada
[S:34]	[S:06]	10,00
[S:44]	[S:06]	9,41
[S:24]	-	9,00
[S:23]	[S:07]	8,98
[S:26]	-	8,96
[S:29]	[S:08]	6,97
[S:18]	-	6,91
[S:40]	[S:06]	6,78
[S:43]	-	6,57
[S:01]	[S:06]	6,52
[S:00]	-	5,99
[S:32]	[S:06]	5,98
[S:38]	[S:06]	5,24
[S:41]	[S:06]	5,13
[S:11]	-	4,57
[S:02]	[S:06]	4,50
[S:39]	[S:06]	4,37
[S:36]	[S:06]	4,27
[S:15]	-	4,16
[S:05]	-	3,30
[S:13]	-	3,29
[S:33]	-	3,24
[S:06]	-	2,99
[S:37]	-	2,90
[S:04]	-	2,90
[S:35]	[S:06]	2,62
[S:03]	-	2,29
[S:10]	-	2,23
[S:25]	-	2,22
[S:30]	[S:08]	1,88
[S:17]	-	1,87
[S:31]	-	1,81
[S:09]	-	1,70
[S:07]	-	1,41
[S:42]	-	0,42
[S:16]	-	0,00

Fonte: Elaborado pelo autor

Tabela 30 – Ranque de sentenças na coleção C1 – Texto-fonte em inglês – Método RLs + UWs 1:3

Sentença texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Pontuação normalizada
[S:23]	[S:07]	10,00
[S:34]	[S:06]	9,27
[S:29]	[S:08]	8,66
[S:44]	[S:06]	8,53
[S:01]	[S:06]	8,40
[S:18]	-	7,74
[S:32]	[S:06]	7,49
[S:38]	[S:06]	5,99
[S:26]	-	5,90
[S:36]	[S:06]	5,37
[S:24]	-	5,36
[S:00]	-	5,30
[S:40]	[S:06]	5,16
[S:02]	[S:06]	5,05
[S:11]	-	4,60
[S:43]	-	4,49
[S:41]	[S:06]	4,34
[S:05]	-	4,20
[S:17]	-	4,14
[S:42]	-	3,53
[S:15]	-	2,99
[S:39]	[S:06]	2,75
[S:33]	-	2,37
[S:13]	-	2,29
[S:35]	[S:06]	2,18
[S:04]	-	2,04
[S:03]	-	2,03
[S:06]	-	1,79
[S:37]	-	1,68
[S:09]	-	1,29
[S:31]	-	1,04
[S:07]	-	0,97
[S:10]	-	0,93
[S:25]	-	0,74
[S:30]	[S:08]	0,69
[S:16]	-	0,00

Fonte: Elaborado pelo autor

Tabela 31 – Ranque de sentenças na coleção C1 – Texto-fonte em português – Método RLs

Sentença texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Pontuação normalizada
[S:10]	[S:07]	10,00
[S:08]	-	7,72
[S:15]	-	7,23
[S:02]	-	7,16
[S:03]	[S:03]	6,78
[S:14]	-	6,39
[S:04]	[S:04] / [S:00]	5,54
[S:09]	[S:05]	5,13
[S:16]	-	5,06
[S:11]	-	3,95
[S:17]	-	3,52
[S:06]	[S:01]	2,94
[S:07]	-	1,59
[S:12]	-	0,93
[S:01]	[S:02]	0,63
[S:18]	-	0,13
[S:00]	[S:02]	0,00

Fonte: Elaborado pelo autor

Tabela 32 – Ranque de sentenças na coleção C1 – Texto-fonte em português – Método RLs + UWs 1:1

Sentença texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Pontuação normalizada
[S:10]	[S:07]	10,00
[S:08]	-	8,54
[S:02]	-	8,18
[S:17]	-	8,16
[S:16]	-	7,90
[S:04]	[S:04] / [S:00]	7,14
[S:01]	[S:02]	6,84
[S:15]	-	6,80
[S:09]	[S:05]	5,80
[S:14]	-	5,72
[S:07]	-	5,49
[S:03]	[S:03]	5,43
[S:06]	[S:01]	5,29
[S:11]	-	4,33
[S:12]	-	3,46
[S:18]	-	0,62
[S:00]	[S:02]	0,00

Fonte: Elaborado pelo autor

Tabela 33 – Ranque de sentenças na coleção C1 – Texto-fonte em português – Método RLs + UWs 1:3

Sentença texto-fonte	Foi alinhada a qual(is) sentença(s) do sumário?	Pontuação normalizada
[S:01]	[S:02]	10,00
[S:17]	-	9,58
[S:16]	-	7,87
[S:07]	-	7,11
[S:10]	[S:07]	6,92
[S:08]	-	6,60
[S:02]	-	6,52
[S:04]	[S:04] / [S:00]	6,29
[S:06]	[S:01]	5,65
[S:09]	[S:05]	4,58
[S:12]	-	4,54
[S:15]	-	4,33
[S:14]	-	3,38
[S:11]	-	3,31
[S:03]	[S:03]	2,61
[S:18]	-	0,84
[S:00]	[S:02]	0,00

Fonte: Elaborado pelo autor