Programa de Pós-Graduação em Lingüística

PPGL

# *Sumarização Automática Multidocumento Multilíngue: seleção de conteúdo e tratamento da redundância com base em conhecimento léxico-conceitual*

Yasmin Vizeu Camargo

SÃO CARLOS

2020

Universidade Federal de São Carlos

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

# SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO MULTILÍNGUE: SELEÇÃO DE CONTEÚDO E TRATAMENTO DA REDUNDÂNCIA COM BASE EM CONHECIMENTO LÉXICO-CONCEITUAL

YASMIN VIZEU CAMARGO
Bolsista: CAPES

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

Orientadora: Profª. Drª. Ariani Di Felippo

São Carlos/SP
2020

**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 da candidata Yasmin Vizeu Camargo, realizada em 19/03/2020:

Profa. Dra. Ariani Di Felippo
UFSCar

Prof. Dr. Jackson Wilke da Cruz Souza
UNIFAL

Profa. Dra. Cláudia Dias de Barros
IFSP

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Jackson Wilke da Cruz Souza, Cláudia Dias de Barros e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Ariani Di Felippo

## AGRADECIMENTOS

De todas as partes de um trabalho acadêmico, acredito que esta é a mais pessoal (e, na minha opinião, a mais especial). Não por acreditar que nas demais haja pouco, verdadeiramente, de quem escreve, mas sim por saber que aqui é onde os conceitos, as teorias, o rigor científico e as normas da ABNT pouco interferem. Sendo assim, eu não poderia começar os meus agradecimentos de outra maneira.

Agradeço a Deus, pelo sustento. Por não se esquecer de mim um dia sequer, se mostrando presente das mais diferentes formas, sobretudo nas vezes em que eu supliquei por amparo e por respostas. Por me fortalecer e me proteger em mais uma etapa, com o carinho e o cuidado que só um Pai é capaz de despender.

Agradeço à minha mãe, Maria José, quem eu me lembro, desde pequena, dizer que a educação é a única maneira de transformar realidades. Não tenho dúvidas de que nada do que eu construí até então existiria se não fossem o seu incentivo e a sua dedicação em criar e educar filhas conscientes da liberdade que a educação possibilita.

Ao meu pai, Antonio Carlos, que sempre se empenhou em não me deixar duvidar da minha capacidade e por vibrar com as minhas conquistas como se fossem suas — e são. Por ser, desde sempre, exemplo de quem ama o que faz, o que me inspira todos os dias e me mostra que acreditar no que se faz é o que diferencia emprego de trabalho.

À minha irmã, Keith, por ser espelho e exemplo de quem eu quero ser quando crescer. Pelas conversas e por partilhar a vida comigo, com todos os seus altos e baixos, desde que eu nasci. Obrigada por ser companhia, escuta e apoio. Agradeço por me dar o privilégio de dizer que sou sua irmã, posto esse do qual eu me orgulho muito.

À minha amiga querida, Magê, por ser sempre um acalento a cada encontro, a cada café, a cada companhia para as atividades corriqueiras. Em cada um desses momentos — e talvez fosse nesses que eu mais precisasse —, contar com as suas palavras e com a sua amizade foi como um respiro.

Ao meu amigo de longa data, Caio, por ser sempre inspiração nos momentos em que estivemos juntos, dos almoços às peças de teatro. Nossa amizade, que atravessa tantos anos, me lembra sempre de como eu sou grata por ser amiga de uma pessoa tão cheia de luz e coragem para lutar pelo que acredita, o que me faz querer ser uma pessoa melhor sempre.

À amiga que o mestrado me deu de presente, Renata, por cada uma de nossas conversas — que não precisavam de muito para se estenderem até perdermos a hora — e, principalmente, por ser companhia nessa experiência intensa que é a academia. Por cada um

dos tantos cafés que tomamos juntas, agradeço por partilhar comigo esses últimos dois anos, que teriam sido muito mais difíceis sem alguém com quem contar.

À minha orientadora, Ariani, pela paciência e confiança, e pelos ensinamentos compartilhados ao longo de todo o processo.

Aos colegas do NILC, em especial ao Roney, pela importante ajuda nos momentos em que tropecei nos limites entre a Linguística e a Computação. Pelas conversas sempre muito enriquecedoras e pela atenção comigo e com o meu trabalho.

À CAPES, pelo apoio financeiro.

## RESUMO

As aplicações de Sumarização Automática Multidocumento Multilíngue (SAMM) geram, a partir de uma coleção de textos em diferentes línguas, um sumário em uma das línguas-fonte. Assim, a SAMM lida com os problemas da Sumarização Automática Multidocumento (SAM), como a identificação de conteúdo relavante e o tratamento da redundância, além das múltiples línguas-fonte. Para a produção de sumários multilíngues em português, o método CFUL é o de melhor desempenho. Sendo extrativo, ele basicamente pontua as sentenças dos texto-fonte em suas línguas originais com base na frequência simples de seus conceitos nominais na coleção e seleciona as mais bem ranqueadas em português para o sumário, evitando a redundância entre tais sentençcas com base na verificação da sobreposição de palavras entre elas. Neste trabalho, propôs-se o método extrativo CFULHiper. Nele, a seleção de conteúdo também é feita com base na frequência dos conceitos nominais da coleção, mas considerando adicionalmente uma pontuação diferenciada para os conceitos superordenados que se encontram em relação hierárquica a outros na coleção, sob a hipótese de que eles veiculam informações mais genéricas e, portanto, relevantes para sumários informativos. Ademais, o CFULHiper objetiva evitar a redundância com base na sobreposição de conceitos, buscando capturar mais adequadamente a similaridade de conteúdo entre as sentenças selecionadas. Para desenvolver o CFULHiper, selecionou-se o *corpus* CM2News, que é composto por 20 coleções bilíngues (português e inglês) de notícias, cujos nomes dos textos-fonte foram anotados com conceitos extraídos da WordNet de Princeton. Tal *corpus* foi estendido pela inclusão de 10 novas coleções, o que resultou na versão 2.0 do CM2News. O CM2News 2.0 foi submetido a um pré-processamento automático no qual, para cada coleção, realizou-se: (i) identificação das relações conceituais hierárquicas em cada uma das 30 coleções e (ii) cálculo da frequência simples e da acumulada dos conceitos em cada uma das 30 coleções. Para calcular a frequência acumulada de um hiperônimo *x*, a frequência simples de *x* é somada à frequência simples de seus hipônimos. Na sequência, aplicou-se automaticamente o CFULHiper a cada coleção do *corpus*, produzindo sumários em português com 70% de compressão. Os 30 sumários gerados pelo método foram avaliados manualmente quanto à qualidade linguística (DUC'05), segundo sua gramaticalidade, não-redundância, clareza referencial, foco temático e estrutura/coerência, e à informatividade, automaticamente, via ROUGE. Os extratos gerados pelo CFULHiper apresentaram resultados ligeiramente melhores na avaliação de informatidade, quando comparado a outros métodos, indicando que informações mais genéricas são, de fato, relevantes para compor extratos multilíngues. A sobreposição conceitual, no entanto, não teve impacto no tratamento da redundância porque sentenças selecionadas exclusivamente de um único texto-fonte já não apresentam muita redundância entre si e também porque as coleções multidocumento tendem a apresentar baixa sinonímia e polissemia e, assim, aplicar uma medida de sobreposição lexical ou conceitual não gera diferença na identificação da similaridade entre as sentenças.



**Palavras-chave:** sumarização multidocumento multilíngue; conhecimento léxico-conceitual; seleção de conteúdo; redundância; relação hierárquica.

## ABSTRACT

Multilingual Multi-document Summarization consists in automatically producing, from a collection of texts on the same topic and in different languages, a summary in one of the source languages. Thus, this task deals with the problems of Multi-document Summarization, such as the identification of relevant content and the treatment of redundancy, and with the multiplicity of source languages. For the production of multilingual summaries in Portuguese, CFUL is the method with the best performance. CFUL is extractive and thus it punctuates the source sentences in their original languages based on the simple frequency of their nominal concepts in the collection and it selects the best-ranked ones in Portuguese for the summary, avoiding redundancy based on word overlapping between them. In this work, the CFULHiper extractive method is proposed. It also selects content based on the simple frequency of the nominal concepts, but it additionally takes into account a differentiated score for the superordinate concepts that are in hierarchical relations with others in the collection. The method assumes that superordinate concepts convey generic information, which is relevant to compose informative summaries. Moreover, CFULHiper avoids redundancy based on concept overlapping, capturing sentence similarity in a more intelligent manner. To develop CFULHiper, we have selected the CM2News *corpus*, which consists of 20 bilingual collections (Portuguese and English) of news, whose nouns of the source texts were annotated with concepts from WordNet of Princeton. The *corpus* was extended with the inclusion of 10 new collections, resulting at the second version of CM2News. The CM2News 2.0 *corpus* was submitted to an automatic pre-processing. For each collection, we have performed: (ii) identification of the conceptual hierarquical relations across the source-texts, and (iii) calculation of the simple and cumulative frequencies of the nominal concepts. To calculate the accumulated frequency of a hyperonym *x*, the simple frequency of *x* is added to the simple frequency of its hyponyms. Then, we automatically applied CFULHiper to each collection of the *corpus*, producing 30 summaries in Portuguese with 70% compression. We have evaluated the linguistic quality (gramaticality, non-redundancy, referential clarity, focus and estructure/coherence) and the informativeness (ROUGE) of all summaires generated by CFULHiper. The informativeness of the CFULHiper extracts is slightly better, which indicates that more generic information is relevant for composing multilingual extracts. The conceptual overlap, however, had no impact on the treatment of redundancy. Since sentences selected exclusively from a single source-text no longer have much redundancy between themselves, and multi-document *clusters* tend to have few cases of synonymy and polysemy, the application of a lexical or conceptual overlap measure basically generates the same results for similarity identification.



**Keywords**: multilingual multi-document summarization; lexical-conceptual knowledge; content selection; redundancy; hierarquical relation.

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE TABELAS

## LISTA DE SIGLAS

CM2NEWS – *Corpus* Multidocumento Bilíngue de Textos Jornalísticos

CM3NEWS – *Corpus* Multidocumento Trilíngue de Textos Jornalísticos

C – *Cluster* ou coleção

CST – *Cross-document Structure Theory*

DEMS – *Dissimilarity Engine for Multi-document Summarization*

DUC – *Document Understanding Conference*

L – Língua

MULSen – *Multilingual Sense Estimator from NILC*

NILC – Núcleo Interinstitucional de Linguística Computacional

PLN – Processamento Automático das Línguas Naturais

ROUGE – *Recall-Oriented Understudy of Gisting Evaluation*

RST – *Rethorical Structure Theory*

SA – Sumarização Automática

SAM – Sumarização Automática Multidocumento

SAMM – Sumarização Automática Multidocumento Multilíngue

SYNSET – *Synonym set* (conjunto de formas sinônimas)

TA – Tradução automática

WN.Pr – WordNet de Princeton

# ÍNDICE

# 1 INTRODUÇÃO

## 1.1 Contextualização

Com a ascensão da internet, é cada vez mais crescente o volume de dados disponível *online*. Segundo estudo realizado pela *International Data Corporation*, atingiu-se a marca de 33 *zettabytes* em informações na *web* em 2018, com previsão de 175 *zettabytes* para 2025 (REINSEL; GANTZ; RYDNING, 2018). Muita dessa informação está em formato textual, circulando em diferentes plataformas (portais de notícias *online*, redes sociais, *blogs*, entre outras) e em diferentes línguas.

Nesse cenário, identificar uma maneira de facilitar o acesso a toda essa informação pelos usuários torna-se importante. Assim, as aplicações de Sumarização Automática (SA), subárea do Processamento Automático das Línguas Naturais (PLN), podem ser alternativas relevantes para se lidar com esse problema, já que têm como objetivo, a partir de uma fonte, extrair as informações mais relevantes de modo que seja possível apresentá-las de forma condensada a um usuário e/ou aplicação (MANI; MAYBURY, 1999).

Considerando que uma das principais características do grande volume de informações que circula atualmente é a multiplicidade de línguas, há vários trabalhos que buscam explorar essa questão no cenário da SA. Tais investigações podem ser organizadas em 3 diferentes abordagens, isto é, independente de língua, *cross-language* e multilíngue (ORĂSAN, 2009). Em todas elas, busca-se comumente gerar sumários extrativos[1] (ou extratos), isto é, sumários compostos por sentenças extraídas integralmente dos textos-fonte, que sejam, ao mesmo tempo, informativos e genéricos. Por "informativo", entende-se a presença das informações mais relevantes do(s) texto(s)-fonte e, por "genérico", entende-se que seu público-alvo não é específico.

A abordagem independente de língua caracteriza-se por não se basear em conhecimento linguístico para realizar a seleção de conteúdo. No caso, os métodos dessa abordagem capturam as informações mais relevantes do(s) texto(s)-fonte com base em conhecimento empírico-estatístico e, com isso, podem ser aplicáveis a praticamente todas as línguas (p.ex.: COWIE *et al.*, 1998, RADEV *et al.*, 2004, LITVAK; LAST; FRIEDMAN, 2010).

[1] Na sumarização extrativa, as sentenças-fonte são pontuadas e ranqueadas em função de algum critério de relevância, que busca capturar a importância do conteúdo veiculado pela(s) sentença(s) no(s) texto(s)-fonte. Diante do ranque, selecionam-se as sentenças mais bem pontuadas e que não sejam redundantes entre si para compor o sumário extrativo ou extrato (NENKOVA; MCKEOWN, 2011).

A abordagem *cross-language* caracteriza-se por partir de um material-fonte em uma língua L*x* e gerar um sumário desse material em uma língua L*y*. Tal sumarização pode ser mono ou multidocumento, dependendo da quantidade de documentos a ser sumarizada (p.ex.: ORĂSAN; CHIOREAN, 2008, WAN; LI; XIAO, 2010, BOURDIN; HUET; TORRES-MORENO, 2011).

Os métodos/sistemas da abordagem *cross-language* englobam uma etapa central, que é a de tradução automática (TA). A TA pode ser realizada de duas formas: antes ou depois do processo de extração de conteúdo (WAN; LI; XIAO, 2010). Na primeira delas, denominada *early-translation*, os textos-fonte são traduzidos automaticamente para a língua-alvo e sumarizados com base em métodos exclusivamente multidocumento desenvolvidos para a língua-alvo. Na segunda, denominada *late-translation*, os textos-fonte são sumarizados por um método multidocumento existente para a língua-fonte e, na sequência, o sumário é automaticamente traduzido para a língua-alvo. Em ambas as abordagens *cross-language*, tem-se material-fonte monolíngue, permitindo que um método de SA mono ou multidocumento, baseado em conhecimento linguístico (simples ou profundo) ou em conhecimento empírico/estatístico, possa ser aplicado (cf. GUPTA; LEHAL, 2010, KUMAR; SALIM; RAZA, 2012).

A abordagem multilíngue é, necessariamente, multidocumento, pois parte de um conjunto de ao menos 2 textos que abordam o mesmo assunto, sendo 1 texto em uma *Lx* (a do usuário) e 1 texto em uma língua *Ly* (língua estrangeira) para gerar um sumário na *Lx*, que é a língua de interesse do usuário. A tal processo de sumarização, dá-se o nome específico de Sumarização Automática Multidocumento Multilíngue (SAMM) (p.ex.: EVANS; KLAVANS; MCKEOWN, 2004, TOSTA, 2014, DI-FELIPPO; TOSTA; PARDO, 2016). Os métodos de SAMM também podem englobar uma etapa de TA, que é realizada antes do processo de SA (*early-translattion*) e consiste em traduzir o(s) texto(s)-fonte em língua estrangeira (L*y*) para a língua de interesse do usuário (L*x*). Após a TA, um método de Sumarização Automática Multidocumento (SAM) pode ser aplicado.

Para a geração de extratos multidocumento multilíngues em português, tem-se os trabalhos de Tosta, Di-Felippo e Pardo (2013) e Tosta (2014). Esse último fora mais recentemente publicado em Di-Felippo, Tosta e Pardo (2016). Tosta, Di-Felippo e Pardo (2013) desenvolveram dois métodos superficiais segundo a abordagem *early-translation*, nos quais o tratamento da redundância é feito via *word overlap*[2]. Um dos métodos pontua e

[2] A medida *word overlap* calcula a redundância ou similaridade entre sentenças com base na sobreposição das palavras de classe aberta idênticas (JURAFSKY; MARTIN, 2001). Para calcular a *word overlap* entre as

ranqueia as sentenças-fonte pela posição que elas ocupam em seus respectivos textos-fonte e o outro realiza o mesmo processo com base na frequência dos itens lexicais das sentenças na coleção de textos-fonte. A medida *word overlap* também é empregada para substituir uma sentença selecionada que foi traduzida por uma similar que seja proveniente do texto original em português. Segundo os parâmetros de avaliação intrínseca da *Document Understanding Conference* (DUC)[3] (gramaticalidade, não-redundância, clareza referencial, foco temático e estrutura/coerência) (DANG, 2005), os autores verificaram que o método pautado na localização das sentenças obteve em média as mais altas pontuações quanto aos 5 parâmetros. Entre os parâmetros linguísticos, a gramaticalidade foi o que obteve a menor média, o que pode ser explicado por problemas gerados pela TA dos textos-fonte.

Buscando alternativas para esse cenário, Tosta (2014) e Di-Felippo, Tosta e Pardo (2016) investigaram dois métodos que, uma vez pautados em conhecimento profundo do tipo léxico-conceitual, evitam a TA integral dos textos-fonte e capturam mais adequadamente o conteúdo relevante da coleção, gerando sumários mais informativos e coesos/coerentes. Neles, a seleção do conteúdo tem início com a pontuação e o ranqueamento das sentenças originais em função da frequência de ocorrência, na coleção, dos conceitos expressos por seus nomes comuns e a redundância também é evitada pela aplicação da medida *word overlap.*

O CF (*concept frequency*) é um desses métodos, o qual realiza a seleção das sentenças apenas com base no ranque. Ele seleciona a primeira sentença do ranque para compor o sumário e a traduz automaticamente para o português caso esteja em língua estrangeira. Esse processo é repetido até que se atinja a taxa de compressão (isto é, tamanho desejado do sumário). O outro método, o CFUL (*concept frequency + user language*), seleciona somente as sentenças mais bem ranqueadas provenientes dos textos em português.

Os métodos CF e CFUL foram comparados ao melhor método superficial de Tosta, Di-Felippo e Pardo (2013), tido como *baseline.* Quanto à gramaticalidade, avaliada com base nos parâmetros da DUC'05, e à informatividade, avaliada pela aplicação das medidas ROUGE, verificou-se que o método CFUL supera o CF e o *baseline.* Isso indica que um extrato composto exclusivamente por sentenças originais na língua do usuário é (i)

---

sentenças de um par (S1 e S2), divide-se o número total de palavras idênticas entre as sentenças pela soma do número total de palavras de cada sentença, excluindo-se as *stopwords*, números e símbolos. O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo de 1, mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante.

[3] Em 2008, a DUC passou a ser a trilha de sumarização automática da *Text Analysis Conference* (TAC) (Disponível em: https://duc.nist.gov/).

informativo (isto é, veicula a informação principal da coleção), posto que os conceitos que ocorrem nos textos em língua estrangeira também foram levados em consideração para a criação do ranque sentencial, e (ii) mais gramatical, já que não inclui sentenças traduzidas.

Diante do exposto, vê-se que os resultados da aplicação de conhecimento léxico-conceitual avançaram o então estado atual da SAMM. No entanto, havia alguns aspectos que ainda podiam ser explorados. Um deles dizia respeito à pontuação das sentenças, que no CFUL é feita somente com base na frequência simples dos conceitos na coleção, desconsiderando-se, por exemplo, qualquer tipo de relação entre os conceitos. Outro ponto dizia respeito à identificação da redundância, feita no CFUL com base somente na sobreposição lexical. Tendo em vista que a medida estatística utilizada é a *word overlap*, ressalta-se que esta não captura a similaridade que se dá entre palavras sinônimas (intratextual).

## 1.2 Objetivos e hipóteses

Tendo em vista os resultados promissores das investigações sobre a aplicação de conhecimento léxico-conceitual na SAMM, objetivou-se investigar o processo de seleção de conteúdo, considerando: (i) pontuação ou peso diferenciado para os conceitos superordenados que estão em relação de hiperonímia com outro(s) na coleção, e (ii) tratamento da redundância com base na sobreposição de conceitos (*concept overlap*).

O aspecto descrito em (i) foi definido a partir da hipótese de que os conceitos superordenados (ou "hiperônimos") podem auxiliar na identificação de conteúdo genérico na coleção, relevante para compor extratos informativos/genéricos. Essa é a relação que se dá, por exemplo, entre os conceitos "veículo" (hiperônimo) e "carro" (hipônimo). Tal estratégia é, aliás, frequentemente utilizada na desambiguação e anotação de sentidos. Tal hipótese também se sustenta se considerada a taxa de compressão, que gera extratos mais genéricos quanto menor for o sumário. O aspecto descrito em (ii) foi definido com base na hipótese de que a similaridade entre sentenças é mais adequadamente calculada ao se incluir a ocorrência de expressões distintas de um mesmo conceito, o que não é possível por meio da medida *word overlap*.

## 1.3 Metodologia

- **Tarefa 1** - Revisão da literatura: essa tarefa consistiu na análise constante da bibliografia fundamental e de referências pertinentes à pesquisa que surgiram no decorrer da

investigação. A bibliografia englobou trabalhos sobre SAMM e Semântica Lexical Computacional.

- **Tarefa 2 -** Proposição do método CFULHiper: modelou-se um método que permitisse investigar os aspectos ainda não explorados segundo a revisão da literatura. Assim, propôs-se um método extrativo de SAMM que, assim como o CFUL, também se baseia na frequência dos conceitos nominais para selecionar conteúdo, sendo, portanto, de abordagem profunda. Tal método, o CFULHiper, caracteriza-se por privilegiar conceitos genéricos e evitar a redundância pela sobreposição conceitual. Nele, a pontuação diferenciada dos conceitos genéricos é definida pela frequência acumulada, que pressupõe a identificação dos conceitos em relação hierárquica nas coleções.

- **Tarefa 3 -** Seleção e extensão do *corpus*: selecionou-se o CM2News (DI-FELIPPO, 2016), posto que é um *corpus* com 20 coleções bilíngues (português e inglês) constituídas por: (i) 1 notícia em inglês e 1 em português sobre certo evento, (ii) anotação conceitual dos nomes dos textos-fonte, (iii) 1 sumário multilíngue de referência em português e (iv) 2 extratos automáticos multilíngues em português. Neste trabalho, construíram-se 10 novas coleções, resultando na versão 2.0 do *corpus*, o que englobou a anotação conceitual dos nomes dos textos-fonte e a produção de sumários multilíngues de referência em português para as coleções acrescidas.

- **Tarefa 4 -** Aplicação do método e construção dos extratos: aplicou-se automaticamente o CFULHiper ao CM2News (2.0). Para tanto, cada coleção foi submetida a 2 pré-processos automáticos: (i) identificação das relações conceituais hierárquicas nos textos-fonte, e (iii) cálculo da frequência simples e acumulada dos conceitos nominais. Diante disso, as sentenças-fonte de cada coleção foram pontuadas e ranqueadas em função da frequência de seus conceitos e as mais bem ranqueadas foram selecionadas para o extrato até que se atingisse a taxa de compressão. Para evitar redundância, verificou-se a sobreposição dos conceitos nominais entre elas com base em um *threshold* empiricamente estabelecido. Uma vez selecionadas, as sentenças foram justapostas e ordenadas (manualmente) segundo sua posição no respectivo texto-fonte, resultando nos extratos.

- **Tarefa 5 -** Avaliação do método CFULHiper: avaliou-se intrinsecamente o método CFULHiper. Assim, os extratos automáticos foram avaliados quanto à qualidade

linguística e à informatividade. A qualidade linguística foi manualmente avaliada segundo os critérios da DUC'05 (DANG, 2005) e a informatividade foi autromaticamete avaliada via ROUGE (LIN, 2004). Os resultados foram comparados aos do método CFUL (TOSTA, 2014) e ao melhor método de Tosta, Di-Felippo e Pardo (2013).

## 1.4 Estrutura da dissertação

Este texto está estruturado em 6 Seções. Na Seção 2, apresenta-se a revisão da literatura. Na Seção 3, destacam-se o *corpus* selecionado e a sua extensão. Na Seção 4, apresenta-se o método de SAMM investigado neste trabalho e sua aplicação ao *corpus* selecionado e estendido. Na Seção 5, apresenta-se a avaliação intrínseca dos sumários extrativos gerados pelo método proposto, considerando a qualidade linguística e a informatividade. Na Seção 6, tecem-se algumas considerações sobre o trabalho realizado e apontam-se possíveis desdobramentos desta pesquisa ora descrita, além das contribuições para a área.

# 2 REVISÃO DA LITERATURA

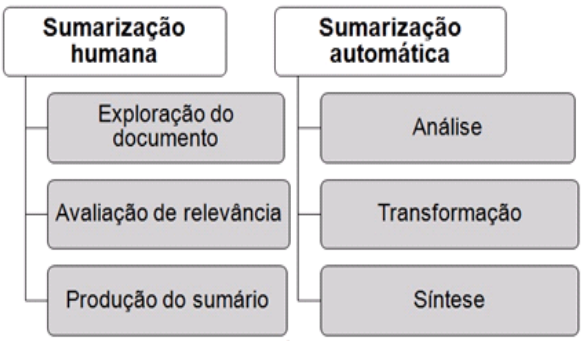
Nesta seção, apresenta-se uma revisão detalhada da literatura, com o objetivo de se criar um panorama dos mais diversos estudos desenvolvidos na área e pertinentes a este trabalho. Na Seção 2.1, apresentam-se as noções gerais sobre SA, que englobam as etapas ou processos que a constituem, os vários fatores que a influenciam e os tipos de avaliação. Na Seção 2.2, apresenta-se a Sumarização Automática Multidocumento (SAM), uma vez que o cenário multilíngue em que este trabalho se insere é necessariamente multidocumento. Na Seção 2.3, apresentam-se os principais trabalhos da literatura que lidam com a multiplicidade de línguas-fonte na SA, enfatizando a tarefa denominada Sumarização Automática Multidocumento Multilíngue (SAMM), a qual é o foco deste trabalho. Na Seção 2.4, destacam-se os trabalhos sobre SAMM envolvendo o português.

## 2.1 Noções básicas sobre Sumarização Automática

Subárea do Processamento de Línguas Naturais (PLN), a Sumarização Automática (doravante SA) ocupa-se da produção automática de sumários a partir de um único texto ou de uma coleção de textos, que sejam suficientemente informativos (MANI, 2001).

Para tanto, sugere-se uma arquitetura específica para o funcionamento de um sistema de Sumarização Automática, baseada no processamento humano para a tarefa (SPARCK JONES, 1998; MANI, MAYBURY, 1999; SPARCK JONES, 1999; MANI, 2001). Tal arquitetura está representada na Figura 1:

Figura 1 – Arquitetura da Sumarização Automática.




Fonte*:* adaptada de Endres-Niggemeyer (1998) e Mani e Mayburry (1999)

Tendo o texto ou coleção de textos como objeto de entrada, a etapa de análise consiste na interpretação desse(s) texto(s)-fonte com o objetivo de representá-lo(s) formalmente. A etapa de transformação é caracterizada por condensar a representação formal gerada na etapa

anterior de modo que uma representação interna do sumário seja obtida. Essa etapa é considerada a mais importante do processo pois é nela que a seleção do conteúdo acontece; essa, por sua vez, é fruto do ranqueamento das sentenças mais relevantes[4] do(s) texto(s)-fonte de modo que as sentenças melhor pontuadas sejam selecionadas para compor o sumário, que, na etapa de síntese, é produzido pela representação interna gerada pela etapa da transformação. Nesse momento, aplicam-se métodos de justaposição[5] sobre as sentenças do ranque para definir a apresentação do sumário final.

No entanto, alguns fatores influenciam diretamente etapas específicas desse processo, como ilustrado no Quadro 2 e detalhado a seguir.

Quadro 1 – Fatores que influenciam o processo de SA

| **Fatores** | **Especificações** |
|---|---|
| Taxa de compressão | ----- |
| Audiência | Genérica |
| | Focada no interesse do usuário |
| Função | Sumário indicativo |
| | Sumário informativo |
| | Sumário crítico |
| Formato | Extrato |
| | *Abstract* |
| Gênero | ----- |
| Número de textos-fonte | Monodocumento |
| | Multidocumento |
| Número de línguas-fonte | Monolíngue |
| | Multilíngue |
| Abordagem | Superficial |
| | Profunda |
| | Híbrida |

Fonte: elaborado pela autora.

[4] Para definir o que é considerado relevante para a seleção do conteúdo, diversos métodos de ranqueamento de sentenças, com diferentes critérios, podem ser utilizados, o que será detalhado nesta seção.

[5] Tais métodos de justaposição podem ser baseados na ordenação, fusão ou correferenciação das sentenças, por exemplo (c.f. SPARCK JONES, 1993).

Por “taxa de compressão”, compreende-se o tamanho do sumário que será obtido. Normalmente, trata-se de uma taxa percentual que recairá sobre número total de palavras do(s) texto(s)-fonte[6]. Por exemplo, com uma taxa de compressão de 70% para um texto-fonte de 1.000 palavras, o sumário em questão corresponderá a 30% do texto de entrada, ou seja, 300 palavras.

Por “audiência”, compreende-se o tipo de público para o qual o sumário está sendo produzido, o que influencia diretamente em seu teor informacional. No caso dos sumários genéricos, as informações que os compõem devem ser abrangentes o suficiente, de modo que o usuário não precise de conhecimento prévio para consumi-lo. No caso dos sumários focados em interesses dos usuários, suas informações são customizadas de modo a atender usuários com dado conhecimento prévio sobre o assunto.

Por “função”, compreende-se o objetivo para o qual o sumário está sendo produzido. Assim, os sumários podem ser (i) informativos, (ii) indicativos ou (iii) críticos. No caso dos sumários informativos, são selecionadas as informações mais relevantes do(s) texto(s)-fonte, de modo que a leitura do(s) original(is) possa ser substituída sem perdas. No caso dos indicativos, as informações são selecionadas de modo a oferecer uma introdução sobre determinado assunto, logo, esse tipo de sumário não substitui a leitura do(s) texto(s) original(is). Por fim, no caso dos críticos, são selecionadas as informações principais do(s) texto(s)-fonte, com o acréscimo de conteúdos avaliativos.

Por “forma”, compreende-se a maneira como o sumário será composto, podendo ser “extrativa” (o que produz extratos) ou “abstrativa” (o que produz abstratos). No caso dos extratos, sua composição é marcada por selecionar integralmente as sentenças mais relevantes do(s) texto(s)-fonte, diferindo-se dos abstratos, que são sumários produzidos com base em um processo de reescrita dessas sentenças.

Por “número de textos-fonte”, compreende-se a composição do material que será usado como entrada para a produção dos sumários, podendo a SA ser monodocumento (isto é, partindo de um único texto) ou multidocumento (isto é, partindo de uma coleção de textos que abordam um mesmo assunto).

Por “número de línguas”, compreende-se a multiplicidade de línguas que envolvem o(s) texto(s)-fonte e o sumário, podendo a SA envolver apenas uma língua (sendo, portanto, monolíngue) ou mais de uma língua. Nesta última, há, ainda 3 desdobramentos: a SA pode ser (i) *cross-language*, (ii) multilíngue ou (iii) independente de língua. Um método/sistema

[6] No caso de uma coleção com mais de um texto-fonte, a taxa normalmente recai sobre o maior texto do conjunto.

*cross-language* caracteriza-se por utilizar um ou mais textos em uma L*x* para produzir um sumário em uma L*y* (p.ex.: ORĂSAN e CHIOREAN, 2008, WAN, LI, XIAO, 2010, BOURDIN et al., 2011). Abaixo, nas Figuras 2 e 3, apresentam-se ilustrações da sumarização *cross-language* nos cenários monodocumento e multidocumento.

Na SA, as avaliações podem ser intrínsecas ou extrínsecas. No caso da primeira, o foco é avaliar o sistema/método por meio da análise de seus sumários, enquanto que a segunda tem como objetivo avaliar a utilidade desses sumários para determinada tarefa, como, por exemplo, a recuperação de informação (SPARCK JONES, GALLIERS, 1996). Nos trabalhos de SA, a avaliação intrínseca é a mais frequentemente utilizada.

Na avaliação intrínseca, são analisadas a qualidade linguística e a informatividade do sumário (MANI, 2001). Por "qualidade linguística" entendem-se os fatores relacionados à coerência, à coesão, à gramaticalidade, entre outros; por "informatividade" compreende-se a qualidade de informações relevantes que o sumário veicula. Normalmente, a avaliação de informatividade é realizada de maneira automática, enquanto que a qualidade linguística é avaliada manualmente.

Para avaliação da qualidade linguística, não há um consenso sobre qual é a melhor forma de realizá-la. Há diversos autores cujas propostas são diversas: Nenkova e Passonneau (2004) estabelecem um método de pirâmide no qual leva-se em consideração sumários de referência para determinar, por meio de pontuação, as informações relevantes que o sumário gerado automaticamente deve conter, permitindo sua comparação; Louis e Nenkova (2013) inovam na proposta de 3 métodos de avaliação automática, que consideram (i) a similaridade entre texto-fonte e sumário gerado automaticamente, (ii) adição de sumários pseudomodelos (isto é, sumários automáticos escolhidos por humanos) aos sumários de referência elaborados por humanos, para se tomar como base, e (iii) sumários automaticamente gerados, para servirem de sumários de referência; e, por fim, a TAC (DANG, 2008) propõe 5 aspectos a serem avaliados, em uma escala de 1 a 5: (i) gramaticalidade, que está relacionada aos padrões de ortografia, à pontuação e à sintaxe; (ii) coerência, que relaciona-se a uma organização textual suficientemente boa, de modo que o sentido do texto se preserve; (iii) não-redundância, que garante que não existam informações repetitivas ao longo do texto; (iv) foco, que visa a relação entre as partes do texto, formando um todo; e (v) clareza referencial, que refere-se à presença de componentes linguísticos que conectem, de maneira satisfatória, os elementos que compõem o sumário.

Para avaliação da informatividade, o pacote de medidas ROUGE (do inglês *Recall-Oriented Understudy for Gisting Evaluation*) (LIN, 2004) é amplamente utilizado, sendo adotado em conferências internacionais. Tais medidas têm como princípio comparar a quantidade de *n-gramas* (isto é, palavras) em comum entre um sumário produzido automaticamente e um sumários de referência. O pacote avalia os resultados sob 3 esferas: (i) precisão, (ii) cobertura e (iii) medida-f. A primeira relaciona-se ao comparativo de *n-gramas* em comum entre o sumário gerado automaticamente e os sumários de referência, relacionado ao total de *n-gramas* do sumário automático; a segunda relaciona-se ao comparativo de *n-gramas* em comum entre o sumário gerado automaticamente e os sumários de referência, relacionado ao total de *n-gramas* do sumário de referência; a terceira tem como objetivo ponderar as duas medidas anteriores, calculando a média harmônica entre a precisão e a cobertura. Abaixo, ilustram-se os cálculos das três medidas:

$$(1)\ \text{Precisão} = \frac{\text{n−gramas em comum entre sumário automático e humano}}{\text{n−gramas do sumário automático}}$$

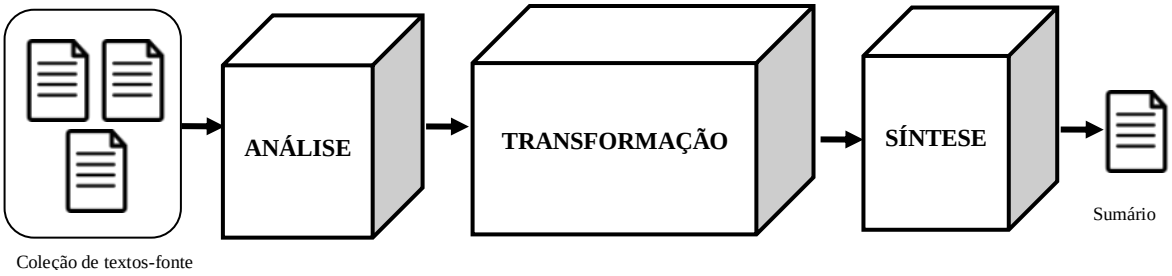
$$(2)\ \text{Cobertura} = \frac{\text{n−gramas em comum entre sumário automático e humano}}{\text{n−gramas do sumário humano}}$$

$$(3)\ \text{Medida} - \text{f} = \frac{\text{2x Precisão x Cobertura}}{\text{Precisão x Cobertura}}$$

## 2.2 A Sumarização Automática Multidocumento

A Sumarização Automática Multidocumento (SAM) objetiva, a partir de uma coleção de dois ou mais textos de fontes distintas sobre um mesmo assunto, gerar um sumário suficientemente informativo, coerente e coeso (MANI, 2001; ORĂSAN, 2009) (Figura 2).

Figura 2 – Arquitetura genérica da SAM




Fonte: Spark Jones (1993)

Esse tipo de sumarização esbarra em diferentes desafios: além das questões como coerência e coesão — que são tão caras à SA —, problemas de complementaridade, contradição e redundância de informações são frequentes no cenário multidocumento, sendo objeto de investigação de diversos trabalhos (c.f. SOUZA, 2015; SOUZA, DI-FELIPPO, PARDO, 2011). Esta última, especialmente, é o fenômeno mais comum, considerando que a multiplicidade de textos-fonte que abordam um mesmo assunto abre margem para a seleção de informações redundantes e repetidas na etapa de transformação.

As pesquisas em SAM tornaram-se mais comuns no início dos anos 2000 (em língua inglesa), graças à popularização de buscadores de notícias e sintetizadores de informação, que demandam esforços de recuperação e extração de informações, cenários nos quais a SA contribui. No entanto, investigações de SAM em português brasileiro são recentes, mas já reúnem trabalhos tanto sob o paradigma superficial (PARDO, 2005, AKABANE; PARDO; RINO, 2011), profundos (CASTRO JORGE; PARDO, 2010, CARDOSO, 2014, LI *et al*, 2010, ZACARIAS, 2016, DE LUCA, 2019) e híbridos (CASTRO JORGE; PARDO, 2011, CASTRO JORGE; AGOSTINI; PARDO, 2011, RIBALDO; RINO; PARDO, 2012, RIBALDO; CARDOSO; PARDO, 2013, CASTRO JORGE, 2015).

### *2.2.1 A SAM superficial*

Gupta e Lehal (2010) e Kumar, Salim e Raza (2012) classificam os métodos/sistemas de SAM superficiais em três grupos segundo suas estratégias de identificação de informação relevante: (i) métodos/sistemas que se baseiam em atributos linguísticos (como, por exemplo, frequência de palavras), (ii) baseados em *cluster* e centroide (que agrupam as sentenças de uma dada coleção com base na similaridade lexical, estatisticamente calculada), e (iii) os que lançam mão de grafos (modelando sentenças em nós e similaridade lexical em arestas).

Para o português, em especial, Pardo (2005) desenvolveu o método GistSumm, baseado em atributo linguístico. Nele, a etapa de análise consiste na segmentação das sentenças da coleção, ranqueando-as sob o critério de frequência de palavras, com o argumento de que a ideia principal de um texto pode ser expressa pelas palavras mais frequentes nele. Assim, o ranque elenca as sentenças que acumulam as palavras mais frequentes do texto e, com base nisso, a etapa de transformação seleciona o conteúdo que comporá o sumário da seguinte maneira: a sentença mais bem pontuada do ranque, considerada a *gist sentence*, inicia o sumário; em seguida, são selecionadas as sentenças que (i) compartilhem ao menos um radical em comum com a dita *gist sentence* e (ii) tenha uma

pontuação maior do que a média da pontuações de todas as sentenças do texto; as sentenças são selecionadas até que se atinja a taxa de compressão, que, no sistema baseado nesse método, é selecionada pelo usuário. Em seguida, na etapa de síntese, as sentenças selecionadas são justapostas conforme a ordem em que aparecem no ranque. As avaliações de informatividade para este sistema foram realizadas de maneira intrínseca e extrínseca. Na primeira, o sistema atingiu 51% quanto à informatividade e, na segunda, o sistema recebeu uma média de 3.12, em uma escala de 0 a 4, quanto à sua utilidade.

Akabane, Pardo e Rino (2011) desenvolveram o RCSumm, que é um método/sistema baseado em grafos e redes complexas. Tal sistema, na etapa de análise, segmenta as sentenças dos textos-fonte e representa-as, em uma rede complexa, por meio de nós cujas arestas indicam relações/pesos entre pares de sentenças. Para ranquear as sentenças, testaram-se 3 medidas, a saber: (i) grau; (ii) coeficiente de aglomeração e (iii) caminho mínimo. Por "grau", entende-se a quantidade de arestas diferentes ligadas a um determinado nó, o que indica o quão esse nó é conectado aos seus vizinhos; tal medida pauta-se no argumento de que a informatividade de uma sentença está estritamente relacionada ao número de relações que ela estabelece com as demais. Por "coeficiente de aglomeração", entende-se o quão conectados são os vértices indiretamente relacionados; isto é, o quão interligados são os vértices *x* e *z* quando *x* está conectado a *y*, que, por sua vez, está conectado a *z*, por exemplo. E, finalmente, por "caminho mínimo" entende-se a sequência mínima de arestas que conecta um determinado nó a outro; dessa forma, o argumento é o de que vértices conectados por uma distância menor detêm informações mais relevantes do que vértices que se conectam por uma distância maior. Assim, na etapa de transformação são selecionadas as sentenças mais bem pontuadas do ranque até que se atinja a taxa de compressão, que é determinada pelo usuário. Ao final, na síntese, calcula-se a redundância entre as sentenças selecionadas, por meio de uma medida que envolve cossenos, e gera-se o sumário. Na avaliação, realizada por meio da medida ROUGE (LIN, 2004), a medida de "grau" foi a que obteve, de um modo geral, os melhores resultados.

Na seção seguinte, apresentam-se os métodos/sistemas de SAM inseridos no paradigma profundo.

### *2.2.2 A SAM profunda*

Os métodos/sistemas de SAM de abordagem profunda caracterizam-se pelo uso massivo de conhecimento linguístico, por meio de recursos, com o objetivo de obter sumários coesos,

coerentes e suficientemente informativos. Os métodos inscritos nessa abordagem classificam-se em 3 grupos, com base no tipo de conhecimento que mobilizam, a saber: (i) sintático (BARZILAY et al, 1999), semântico-discursivo (RADEV e MCKEON, 1998; CASTRO JORGE E PARDO, 2010; CARDOSO, 2014), conhecimento semântico ou conceitual (LI et al, 2010; ZACARIAS, 2016; LUCA, 2019).

Em Barzilay (1999), o critério de relevância adotado leva em consideração conhecimento sintático. Nesse método, a etapa de análise consiste em segmentar as sentenças e, por meio de um *parser* sintático, agrupar as estruturas do tipo "predicado-argumento" que sejam similares sob o argumento de que tal configuração sintática teoricamente expressa tópicos. Na etapa de transformação, selecionam-se as estruturas mais frequentes para que, na síntese, elas sejam reordenadas para gerar um sumário abstrativo.

Radev e McKeon (1998) encabeçaram o primeiro trabalho de SAM baseado em conhecimento semântico-discursivo, utilizando-se da teoria CST (RADEV, 2000). Tal teoria, que é baseada na RST[7] (do inglês *Rhetorical Structure Theory*) (MANN, THOMPSON, 1987), visa conectar as sentenças possibilita conectar sentenças (ou outras unidades textuais) de documentos diferentes, estruturando seus conteúdos (RADEV, 2000); tais conexões estabelecem relações entre as sentenças, que podem ser de diversas ordens.

Jorge e Pardo (2010) e Cardoso (2014), por exemplo, desenvolveram métodos de SAM com base em conhecimento semântico-discursivo, codificado pelas relações do modelo *Cross-Document Structure Theory* (CST) (RADEV, 2000). A CST, que é baseada na *Rhetorical Structure Theory*) (RST[8]) (MANN; THOMPSON, 1987), é um modelo que permite conectar as sentenças (ou outras unidades textuais) de documentos diferentes, estruturando seu conteúdo. A CST é bastante utilizada na SAM por ser robusta, permitindo lidar com os fenômenos gerados pela multiplicidade de textos, como a redundância, a complementaridade e a contradição.

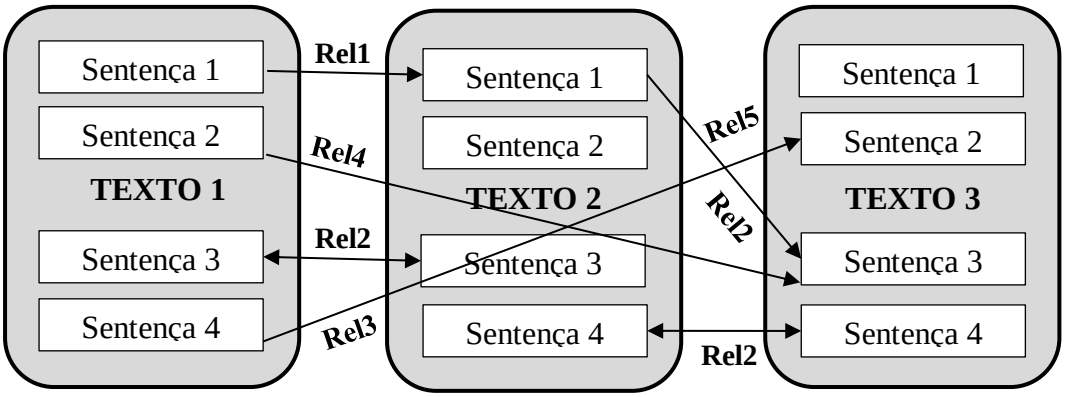
Castro Jorge e Pardo (2010) desenvolveram o sistema CSTSumm, que modela os textos-fonte de uma coleção em um grafo, no qual as sentenças são representadas pelos nós e as arestas codificam as relações CST entre as sentenças, como ilustrado na Figura 3.

---

[7] A RST é uma teoria/modelo linguístico descritivo pautado na classificação dos segmentos discursivos de um texto individual em núcleo (informação principal) ou satélite (informação adicional), relacionando-os por meio de relações retóricas, como Elaboration, List, Causa, Result, Justity, entre outras.

[8] A RST é uma teoria/modelo linguístico descritivo pautado na classificação dos segmentos discursivos de um texto individual em núcleo (informação principal) ou satélite (informação adicional), relacionando-os por meio de relações retóricas, como *Elaboration, List, Cause, Result, Justity*, entre outras.

Figura 3 – Esquema genérico de análise multidocumento




Fonte: Maziero (2012).

Para tanto, os autores utilizaram as 14 relações CST (Quadro 2) propostas por Pardo e Aleixo (2008) a partir da anotação manual do *corpus* CSTNews.

Quadro 2 – Conjunto de relações CST de Pardo e Aleixo (2008)

| | |
|---|---|
| *Identity* | *Elaboration* |
| *Equivalence* | *Contradiction* |
| *Summary* | *Citation* |
| *Subsumption* | *Attribution* |
| *Overlap* | *Modality* |
| *Historical Background* | *Indirect speech* |
| *Follow-up* | *Translation* |

Fonte: Pardo e Aleixo (2008)

Uma vez que os textos-fonte tenham sido modelados em grafos, as sentenças são ranqueadas com base no número de conexões no grafo, sendo que as que estabelecem um maior número de conexões figuram o topo do ranque. Assim, tais sentenças são selecionadas, até que se atinja a taxa de compressão. O diferencial do método de Castro Jorge e Pardo (2010), no entanto, encontra-se na aplicação de operadores de seleção de conteúdo ao ranque, na etapa de transformação; tais operadores atuam codificando as preferências do usuário (como, por exemplo, "informação contextual") e reorganizando o ranque de modo que o sumário apresente a informação solicitada pelo usuário. Assim, na síntese, as sentenças são selecionadas a partir do novo ranque gerado pelo operador. Na avaliação, o sistema apresentou bons resultados de informatividade e qualidade linguística, nos critérios tomados como base. Para tanto, Castro Jorge e Pardo (2010) utilizaram o *corpus* CSTNews, composto

por 50 coleções (que contam, cada uma, com 2 ou 3 textos do gênero jornalístico, em português), sendo que cada coleção aborda um tópico diferente. Tal *corpus* é, ainda, enriquecido com anotações de diversos conhecimentos linguísticos, em diferentes níveis, como: etiquetação morfossintática e sintática, interconexão entre os textos-fonte via CST, anotação dos sentidos de nomes e verbos, anotação de aspectos informativos, anotação discursiva via RST, entre outras. O CSTNews conta, ainda, com sumários humanos (abstratos) monodocumento e multidocumento, além de sumários automáticos multidocumento e extratos humanos multidocumento.

Cardoso (2014) desenvolveu o método RC-4, publicado mais recentemente em Cardoso e Pardo (2015, 2016). O RC-4 baseia-se na combinação de conhecimento semântico-discursivo das teorias RST e CST. Nele, as sentenças são pontuadas e ranqueadas levando em consideração 2 critérios de relevância: (i) a saliência da sentença em seu respectivo texto-fonte (que utiliza como base o modelo de saliência proposto por Marcu (1997)[9]); e (ii) correlação com os fenômenos multidocumento, indicada pela CST. Avaliado pelo pacote de medidas ROUGE (LIN, 2004), o método RC-4 configura-se como o método profundo de melhor desempenho para o português.

Quanto ao desempenho, o método RC-4 foi avaliado por meio da medida ROUGE e comparado a outros sumarizadores multidocumento para o português. No geral, o RC-4 produz sumários melhores que os do CSTSumm, considerado o estado-da-arte da abordagem profunda, e similares aos obtidos pelo RSumm (RIBALDO; CARDOSO; PARDO, 2013, RIBALDO; CARDOSO; PARDO, 2016), considerado o estado-da-arte da abordagem híbrida.

Quanto ao emprego de conhecimento conceitual, destacam-se, para o português, os trabalhos de Zacarias (2016) e De Luca (2019).

Zacarias (2016) investigou o emprego de determinadas propriedades hierárquicas como critério de relevância em métodos de SAM que levam em consideração a representação dos textos-fonte por conhecimento léxico-conceitual. Trata-se de um trabalho pautado em grafos e hierarquias, cujas métricas analisadas foram: (i) *Centrality*, (ii) *Simple Frequency*, (iii) *Cumulative Frequency*, (iv) *Closeness*, (v) *Level*.

[9] Tal modelo é baseado no uso de um conjunto promocional formado pelas unidades mais salientes de cada nó interno de uma árvore RST, sob a hipótese de que as unidades textuais que se encontram no conjunto promocional do topo de uma árvore são mais importantes do que as unidades que se encontram em níveis mais baixos.

A métrica "*Simple Frequency*" consiste na frequência de ocorrência de conceitos nos textos-fonte. "*Cumulative Frequency*", por sua vez, pauta-se na soma da frequência de ocorrência de um conceito *x* na coleção, somado à frequência da ocorrência de seus conceitos hipônimos. A métrica "*Centrality*" leva em consideração o número de ligações que um determinado conceito possui com outros da hierarquia. Por "*Closeness*", considera-se o relacionamento entre os conceitos mais importantes de uma representação conceitual, com o objetivo de determinar sua relevância; trata-se de uma medida que calcula a relevância de um conceito em relação a outros e não em sua forma isolada. E, por fim, a medida "*Level*" aplica-se com o objetivo de determinar a localização (em termos de níveis) de um conceito *x* em sua hierarquia, o que pode expressar generalidade ou especificidade, a depender de sua altura.

Assim, Zacarias (2016) propôs e avaliou 2 métodos: CFSumm (do inglês *Concept Frequency Summarization*) e LCHSumm (do inglês *Lexical Conceptual Hierarchy Summarization*). O primeiro tem como etapa de análise a segmentação das sentenças e a anotação de seus substantivos com seus respectivos conceitos na WordNet de Princeton (doravante WN.Pr) (FELLBAUM, 1998). No método CFSumm, de Zacarias (2016), a etapa de transformação consiste na pontuação das sentenças conforme a frequência de ocorrência de seus conceitos em toda a coleção, selecionando-se as mais bem pontuadas até que se atinja a taxa de compressão, definida em 70%. Em seguida, a etapa de síntese consiste em justapor as sentenças na ordem em que foram selecionadas e, depois, ordená-las pela ordem de ocorrência nos textos-fonte. O método LCHSumm, por sua vez, compartilha da mesma etapa de análise adotada pelo método CFSumm para, na etapa de transformação, aplicar um conjunto de regras, obtidas via aprendizado de máquina e estabelecidas a partir das métricas *Simple Frequency, Cumulative Frequency, Closeness* e *Level*, que foram utilizadas como atributos para tanto. Os métodos tiveram sua informatividade avaliada pelo pacote de medidas ROUGE (LIN, 2004) e foram comparados ao método *baseline* GistSumm (PARDO, 2005). Como resultado, verificou-se que o método CFSumm apresenta resultados mais satisfatórios do que o método LCHSumm, ainda que ambos tenham apresentado bom desempenho.

Para avaliar, Zacarias (2016) utilizou o pacote de medidas ROUGE (LIN, 2004), e comparou os resultados obtidos pelos métodos entre si e com os resultados obtidos pelo sistema GistSumm (PARDO, 2005), como pode ser observado na Tabela 1:

Tabela 1 – Média da ROUGE para os métodos CFSumm, LCHSumm e GistSumm

| Método | Média ROUGE-1 | | | Média ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| **CFSumm** | 0.40106 | 0.45736 | **0.42262** | 0.22012 | 0.26009 | **0.23543** |
| **LCHSumm** | 0.38173 | 0.44618 | 0.40760 | 0.18963 | 0.23473 | 0.20765 |
| **GistSumm** | 0.38373 | 0.45129 | 0.41113 | 0.20516 | 0.25200 | 0.22403 |

Fonte: Zacarias (2016)

De Luca (2019) também propôs métodos de SAM baseados em conhecimento léxico-conceitual. No caso, a autora propôs quatro métodos, sendo que todos eles também evitam a redundância com base na medida *word overlap*. Os métodos de De Luca são: (i) LCFSummN, baseado na frequência (simples) de ocorrência dos nomes na coleção; (ii) LCFSummN-V, baseado na combinação da frequência dos nomes e verbos; (iii) LCFSummN-pond, baseado na média ponderada da frequência dos nomes e (iv) LCFSummN-V-pond, baseado na média ponderada da frequência dos nomes e verbos. Para a avaliação, a autora aplicou os quatro métodos a cinco *clusters* do *corpus* CSTNews e os extratos resultantes foram analisados intrinsecamente quanto à informatividade, via ROUGE, e à qualidade linguística segundo os critérios da DUC'05.

Para fins de comparação, considerou-se outro método profundo da literatura, o CSTSumm (JORGE; PARDO, 2010), considerado estado-da-arte dessa abordagem. Considerando a medida-f como parâmetro principal, pois esta é uma média ponderada das medidas de precisão e cobertura, destaca-se, com base nos resultados da Tabela 2, que o método LCFSummN-pond apresenta os melhores resultados de informatividade entre os métodos léxico-conceituais, mas não supera o estado-da-arte. Embora não supere os valores obtidos pelo CSTSumm, as medidas-f de LCFSummN-pond são bastante próximas às medidas do método baseado em conhecimento discursivo, indicando que os extratos gerados por LCFSummN-pond têm bom nível de informatividade.

Tabela 2 – Comparação geral das avaliações ROUGE dos métodos de De Luca (2019).

| Método | Média ROUGE-1 | | | Média ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| **LCFSummN** | 0.415082 | 0.407554 | 0.41079 | 0.20653 | 0.20392 | 0.20497 |
| **LCFSummN-pond** | 0.433772 | 0.423922 | **0.42852** | 0.21298 | 0.20755 | **0.21009** |
| **LCFSummN-V** | 0.425422 | 0.413144 | 0.41868 | 0.20916 | 0.20378 | 0.20617 |
| **LCFSummN-V-pond** | 0.420848 | 0.401462 | 0.41051 | 0.19979 | 0.19025 | 0.19469 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **CSTSumm** | 0.48052 | 0.427008 | **0.451854** | 0.243502 | 0.216736 | **0.229146** |

Fonte: De Luca (2019)

Sobre a qualidade linguística, os dados da Tabela 3 evidenciam que os métodos baseados em conhecimento léxico-conceitual apresentam, no geral, extratos com boa qualidade.

Tabela 3 – Comparação entre os diferentes métodos LCFSumm.

| Critérios | LCFSummN | LCFSummN-V | LCFSummN-pond | LCFSummN-V-pond |
|---|---|---|---|---|
| Gramaticalidade | 4.5 | **4.6** | **4.6** | 4.5 |
| Não-redundância | **4.5** | 4.4 | 4.2 | 4.0 |
| Clareza referencial | **4.2** | 4.0 | **4.2** | 3.8 |
| Foco | **4.4** | 4.2 | 4.2 | 3.7 |
| Estrutura/Coerência | **4.0** | **4.0** | 3.7 | 3.3 |

Fonte: De Luca (2019)

Além disso, observa-se na Tabela 3 que o LCFSummN obteve as melhores médias na maioria dos critérios da DUC'05. Isso indica que os conceitos nominais, sendo os mais frequentes nas coleções, veiculam de fato as informações mais relevantes. Frente ao método LCFSummN, a inclusão dos conceitos verbais aos nominais no método LCFSummN-V melhorou apenas a gramaticalidade dos extratos.

### *2.2.3 A SAM híbrida*

Para o português, há vários métodos híbridos, a saber: Castro Jorge e Pardo (2011), Castro Jorge, Agostini e Pardo (2011), Ribaldo, Rino e Pardo (2012), Ribaldo, Cardoso e Pardo (2016), Camargo (2013) e Castro Jorge (2015).

O método/sistema descrito em Castro Jorge e Pardo (2011) e Castro Jorge, Agostini e Pardo (2011) associa atributos linguísticos superficiais (p.ex.: localização nos textos-fonte) e profundos (p.ex.: quantidade de relações CST).

Em Ribaldo, Rino e Pardo (2012), em particular, exploram-se medidas estatísticas aplicadas a grafos e redes, em combinação com relações CST. Nesse trabalho, os textos-fonte são modelados em um grafo em que as sentenças são representadas como nós e as relações CST entre as sentenças como arestas. As arestas codificam o número de relações CST de uma sentença. Os nós mais altamente conectados são selecionados para o sumário, uma vez que estes representam as informações mais relevantes da coleção. Conforme descrito pelos autores, esse método apresentou bons resultados, aproximando-se dos melhores métodos desenvolvidos para o português. Ribaldo, Cardoso e Pardo (2016), por

sua vez, exploraram medidas estatísticas aplicadas a grafos e redes em combinação com subtópicos (RSumm).

Camargo (2013) investigou estratégias de Sumarização Humana Multidocumento (SHM), em um trabalho de estudo de *corpus* que levou em consideração textos-fonte e seus respectivos sumários elaborados por humanos. Para identificar quais estratégias de seleção de conteúdo relevante um ser humano adota na hora de elaborar um sumário manual, a autora tomou como base o *corpus* CSTNews e a investigação pautou-se no alinhamento sentencial dos textos-fonte aos seus respectivos sumários elaborados manualmente, de modo que foi possível identificar características recorrentes na seleção de informações relevantes realizada por humanos, como, por exemplo, a localização das sentenças no texto, redundância do conteúdo, tamanho da sentença e ocorrência de palavras frequentes na coleção, o que evidencia a presença de atributos de abordagem híbridos em estratégias eficientes de sumarização. Tal caracterização foi submetida a um aprendizado de máquina, gerando um conjunto de regras que foram testadas no *corpus*, resultando em uma precisão de mais de 70%.

Por fim, Castro Jorge (2015) desenvolveu o método MTRST-MCAD, modelado segundo o esquema *Noisy Channel* (em português, *Canal Ruidoso*) (SHANNON, 1948). Para o desenvolvimento desse método, assumiu-se que a fonte do canal (isto é, processo de sumarização) produz um sumário multidocumento que passa por um canal ruidoso no qual algum tipo de ruído é introduzido (normalmente, elementos característicos dos fenômenos multidocumento), produzindo um conjunto maior de textos. No geral, a SAM modelada via *Noisy Channel* englobou três componentes probabilísticos: (i) P(S) é a probabilidade do sumário e descreve o modelo que captura padrões da boa construção de um sumário multidocumento em termos de coerência; (ii) P(C|S) é o canal ruidoso (ou a etapa de transformação na Figura 5) e modela a seleção de conteúdo via diversos atributos que representam os fatores que influenciam a sumarização, e (iii) P(S|C) é a etapa de decodificação, sendo responsável pela busca do melhor sumário de acordo com os modelos P(C|S) e P(S), os quais são inferidos a partir de um *corpus* de textos-fonte e seus correspondentes extratos humanos multidocumento. No MTRST-MCAD, o modelo de transformação (MT) engloba atributos baseados na representação dos textos-fonte via RST.

O modelo de coerência, por sua vez, foi desenvolvido com base no modelo de entidades e com informações discursivas via CST. Aliás, "MCAD" indica que o modelo de coerência foi aplicado após o processo de decodificação. Quanto ao desempenho, o MTRST-MCAD foi comparado aos principais métodos/sistemas multidocumento para o português

via ROUGE e os resultados evidenciam que seus extratos possuem informatividade que os tornam, em algumas das medidas, estado-da-arte.

## 2.3 A Sumarização Automática Multidocumento Multilíngue

Diante da imensa quantidade de informação veiculada em várias línguas na *web*, o multilinguismo também tem sido foco de pesquisas no âmbito da SA, as quais são motivadas pelo interesse em permitir o acesso à informação que inicialmente tenha sido veiculada em uma língua distinta daquela de interesse do usuário. Os métodos/sistemas de SA que envolvem mais de uma língua podem ser organizados em 3 grupos: (i) *cross-language*, (ii) multilíngue e (iii) independentes de língua (ORĂSAN, 2009).

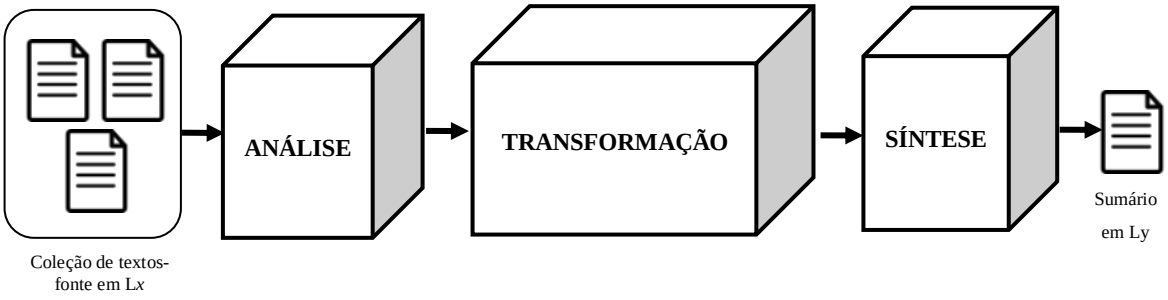
### *2.3.1 A Sumarização cross-language*

A sumarização *cross-language* pode ser monodocumento (Figura 4) ou multidocumento (Figura 5). Em ambas, a língua do sumário é diferente da língua do(s) texto(s)-fonte.

Figura 4 – Ilustração da arquitetura da SA cross-language monodocumento.


Fonte*:* Tosta (2014).

Figura 5 – Ilustração da arquitetura da SA cross-language multidocumento




Fonte*:* Tosta (2014).

De um modo geral, a SA *cross-language* pode seguir duas abordagens, a abordagem *early translation* ou *late translation*, assim, o processo de sumarização envolve necessariamente a etapa de tradução automática (TA) dos textos-fonte ou dos sumários. Tendo em vista que os tradutores automáticos não são aplicações de PLN totalmente precisas, a abordagem *late translation* apresenta certa vantagem frente à *early translation*, pois os textos-fonte não são traduzidos na íntegra, mas sim apenas as sentenças selecionadas para compor o sumário (WAN; LI; XIAO, 2010). Ademais, ressalta-se que os problemas gerados pela TA dos textos-fonte na abordagem *early translation* podem influenciar na aplicação dos métodos de SA mono e multidocumento.

Um exemplo de método/sistema *cross-language* monodocumento é o de Wan, Li e Xiao (2010), o qual gera um sumário em chinês a partir de um texto-fonte jornalístico em inglês. Para tanto, utiliza-se a abordagem *late translation* e, por isso, o texto-fonte em inglês é sumarizado e, em seguida, o sumário é traduzido para o chinês.

Quanto aos sistemas de SA *cross-language* multidocumento, em que se parte de uma coleção de textos em uma língua L*x* que abordam um mesmo assunto para produzir um sumário em uma língua L*y*, destacam-se os trabalhos de Orăsan e Chiorean (2008) e Bourdin, Huet e Torres-Moreno (2011). Orăsan e Chiorean (2008) apresentam um sistema em que uma coleção de textos-fonte em romeno é sumarizada por um método superficial de SAM e o sumário é traduzido para o inglês por um tradutor de livre acesso romeno-inglês. Dessa forma, a abordagem característica desse método é a *late translation*. O método de Orăsan e Chiorean (2008) é *query-based* e, por isso, o sumário deve apresentar a informação da coleção que satisfaz a uma consulta do usuário.

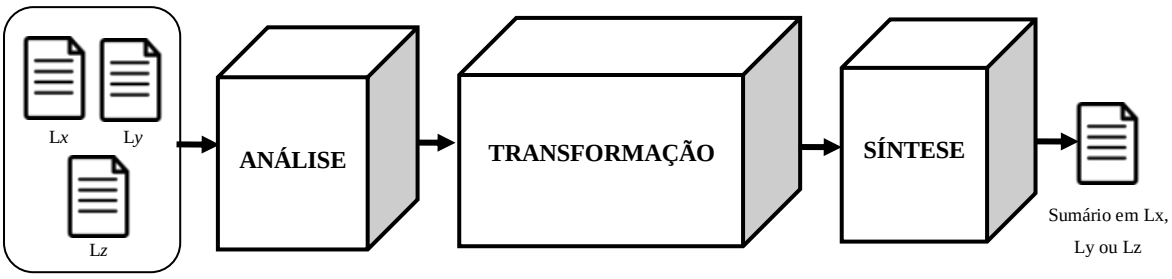
Assim, as sentenças dos textos-fonte são pontuadas e ranqueadas por meio da medida estatística do cosseno, que captura a similaridade lexical entre elas e a consulta do usuário. No caso, as sentenças mais similares à consulta ocupam o topo do ranque. Para compor o sumário, somente as mais bem pontuadas que apresentam pouca similaridade entre si são selecionadas. Dessa forma, pode-se dizer que as sentenças são ranqueadas com base em um critério combinado de relevância e novidade de informação. Outro método *cross-language* multidocumento é o de Bourdin, Huet e Torres-Moreno (2011), em que um sumário em francês é gerado a partir de uma coleção de notícias em inglês. No caso, ao contrário de Orăsan e Chiorean (2008), os textos-fonte em inglês são primeiramente traduzidos para o francês pelo Google Translate e, posteriormente, sumarizados. Assim, a abordagem cross-language é a *early translation*.

Assim como em Wan, Li e Xiao (2010), a seleção de conteúdo em Bourdin, Huet e Torres-Moreno (2011) é baseada na determinação da qualidade de tradução e informatividade das sentenças. Como o método *cross-language* multidocumento de Bourdin, Huet e Torres-Moreno (2011) é superficial baseado em grafo (*graph-based methods*), a qualidade da tradução e a informatividade são critérios inseridos no grafo. Nesses grafos, as sentenças são codificadas em nós e a similaridade entre elas é modelada como arestas entre nós. Assim, as sentenças são pontuadas pela sua conectividade e pela qualidade de tradução.

### *2.3.2 A sumarização multilíngue*

Os métodos/sistemas multilíngues recebem esse nome quando sumarizam uma coleção composta por textos em línguas diferentes (L*x*, L*y* e L*z*, por exemplo), com o objetivo de gerar um sumário em uma das línguas dos textos-fonte (L*x*, L*y* ou L*z*) (p.ex.: EVANS; KLAVANS; MCKEOWN, 2004, ROARK; FISHER, 2005, EVANS; KLAVANS; MCKEOWN, 2005, TOSTA; DI-FELIPPO; PARDO, 2013, TOSTA, 2014), como pode ser observado na Figura 6.

Figura 6 – Arquitetura genérica da SAMM.




Fonte: Tosta (2014).

Nesse grupo, destacam-se os trabalhos de Evans, Klavans e McKeown (2004, 2005) e Roak e Fischer (2005).

Evans, Klavans e McKeown (2004) desenvolveram uma versão multilíngue para o sistema de sumarização *online* denominado *Columbia Newsblaster*. Nela, a SA tem início com uma coleção composta por textos jornalísticos em inglês e, por exemplo, em russo, sobre um mesmo evento. Para uma mesma coleção, realizam-se dois processos de SA. Em um deles, os textos em russo são traduzidos automaticamente para o inglês e, juntamente com os textos originais em inglês, são submetidos ao processo de SAM. No outro, os textos

traduzidos para o inglês e os originais em inglês são submetidos separadamente ao processo de SAM. Nesse último caso, os autores disponibilizam a visualização dos 2 sumários para o usuário familiarizado com a língua inglesa, de tal forma que este pode comparar as diferenças e semelhanças de conteúdo entre os sumários provenientes de textos em inglês e de textos em russo sobre o mesmo assunto.

Para a sumarização dos textos-fonte traduzidos (para o inglês) e dos originais em inglês, o *Columbia Newsblaster* utiliza um de dois sistemas de SAM em função da similaridade dos textos.

Quando os textos-fonte são muito similares, o sistema utilizado é o MultiGen (McKeown *et al*., 1999) que, a partir do agrupamento das sentenças similares em *clusters*, realiza a análise sintática das mesmas e a fusão de informação para a produção do sumário.

Quando os textos são muito distintos, o sistema utilizado no ambiente Columbia Newsblaster é o *Dissimilarity Engine for Multi-document Summarization* (DEMS) (SCHIFFMAN; NENKOVA; MCKEOWN, 2002), que se baseia na utilização de várias estratégias superficiais e profundas para pontuar e ranquear as sentenças de um *cluster*. As estratégias superficiais do DEMS são baseadas em (i) localização, (ii) data de publicação, (iii) tamanho e (iv) *lead words*. As estratégias mais profundas, por sua vez, são baseadas em (i) verbos semanticamente relevantes, segundo a qual sentenças que apresentam verbos plenos que ocorrem associados a sujeitos específicos tendem a transmitir uma informação completa e, por isso, são privilegiadas em detrimento de sentenças que apresentam verbos semanticamente mais vazios, e (ii) conceitos, segundo a qual as sentenças constituídas pelos conceitos mais frequentes da coleção são privilegiadas em detrimento das demais; para identificar os conceitos subjacentes às palavras, os autores utilizam os *synsets* e a relação de hiponímia/hiperonímia da WN.Pr (FELLBAUM, 1998).

Caso sentenças traduzidas sejam selecionadas para compor o sumário, essas são substituídas pelo sistema DEMS, que foi modificado por Evans, Klavans e McKeown (2004) para identificar sentenças originais similares e substituí-las no sumário. Especificamente, o DEMS passou a englobar o sistema de detecção de similaridade denominado SimFinder (HATZIVASSILOGLOU; KLAVANS; HOLCOMBRE, 2001), que se baseia no compartilhamento de: (i) nomes próprios, (ii) itens lexicais morfologicamente relacionados, (iii) itens lexicais sinônimos, (v) itens lexicais com mesmo hiperônimo e (iv) núcleos sintagmáticos.

Em Evans, Klavans e McKeown (2005), o método de SAMM parte de um *corpus* bilíngue formado unicamente por textos em inglês e em árabe com o objetivo de produzir

um sumário em inglês. Os textos em árabe são traduzidos para o inglês e somente as traduções são submetidas ao processo de seleção de conteúdo. Uma vez que a coleção passa a ser monolíngue multidocumento, os autores aplicam o sistema DEMS (SCHIFFMAN; NENKOVA; MCKEOWN, 2002) para selecionar a sentenças que comporão os sumários. Em Evans, Klavans e McKeown (2005), o sistema DEMS simplifica as sentenças (ou seja, quebra as sentenças complexas em menores) antes de pontuá-las e ranqueá-las. Tendo em vista que as sentenças (simplificadas) selecionadas para compor o sumário podiam apresentar problemas de gramaticalidade e/ou inteligibilidade gerados pela TA do árabe para o inglês, os autores utilizaram o sistema SimFinder (HATZIVASSILOGLOU; KLAVANS; HOLCOMBRE, 2001) para identificar sentenças originais em inglês que são similares às traduzidas, as quais são efetivamente levadas ao sumário. Ao final, o sumário é composto apenas por essas sentenças originais. Segundo Evans, Klavans e McKeown (2005), a substituição por sentenças similares originais melhorou em 68% a inteligibilidade do sumário, sem prejudicar a informatividade.

Roak e Fisher (2005) sumarizam coleções compostas por textos traduzidos e originais em inglês. A seleção das sentenças da coleção para compor o extrato (em inglês) é feita com base em nove atributos superficiais (variações de estatísticas lexicais e de posição da sentença no texto-fonte), dando-se preferência às mais bem ranqueadas advindas dos textos originais em inglês. Para evitar a redundância, os autores calculam a sobreposição de bigramas entre a sentença extraída do ranque e as que já compõem o sumário. Se mais da metade dos bigramas que compõem a sentença já estiver presente no sumário, a sentença é descartada. Os autores utilizam 80 coleções disponibilizadas pela DUC como *corpus*, mas não fornecem informações sobre o desempenho do método.

### *2.3.3 A sumarização independente de língua*

No geral, os métodos/sistemas independentes de língua utilizam apenas conhecimento linguístico superficial e/ou conhecimento empírico estatístico e, por isso, processam diferentes línguas, seguindo o pressuposto de que o conhecimento superficial/estatístico é capaz de generalizar fenômenos que são recorrentes na maioria das línguas naturais (p.ex.: COWIE *et al.*, 1998, RADEV *et al.*, 2004, LITVAK; LAST; FRIEDMAN, 2010). Assim sendo, a SA independente de língua pode ser mono ou multidocumento.

Nesse cenário, destaca-se a plataforma *online* MEAD[10], que disponibiliza: (i) métodos de SA independentes de língua, que podem ser aplicados em combinação ou de forma isolada, e (ii) métodos intrínsecos e extrínsecos de avaliação automática de sumários (RADEV *et al.*, 2004). Especificamente, os métodos de SA independentes de língua disponíveis no MEAD são baseados em: (i) centroide; (ii) localização (da sentença no texto-fonte); (iii) similaridade lexical com a primeira sentença do texto-fonte (ou com o título); (iv) tamanho da sentença, (v) palavras-chave, etc (RADEV *et al*., 2004).

Outros métodos independentes de língua podem ser encontrados em Litvak, Last e Friedman. (2010). Especificamente, os autores investigaram 31 métodos, os quais podem ser agrupados em 3 grandes classes: (i) baseados na estrutura (textual) (do inglês, *structure-based methods*); (ii) baseados na representação dos textos em vetores (do inglês, *vector-based methods*), e (iii) baseados na representação dos textos-fonte em grafos (do inglês, *graph-based methods*).

## 2.4 A Sumarização Multidocumento Multilíngue e o Português

Para o português, desconhecem-se trabalhos que focam o desenvolvimento de métodos/sistemas *cross-language* e independentes de língua partindo de coleções com textos em mais de uma língua. Também não há registros de trabalhos em SAMM do português baseados em conhecimento léxico-conceitual com foco em conceitos superordenados.

Quanto aos métodos/sistemas multilíngues, destacam-se os trabalhos de Tosta, Di-Felippo e Pardo (2013) e Tosta (2014). Este último mais recentemente publicado em Di-Felippo, Tosta e Pardo (2016).

Na etapa de análise, os textos-fonte têm suas sentenças segmentadas e as coleções em inglês e espanhol são traduzidas para o português, unificando as coleções para a aplicação das estratégias de sumarização. Assim, na etapa de transformação, testaram-se os 4 métodos, que se diferenciam pelos atributos superficiais adotados. O Método 1 baseia-se na frequência de palavras do texto, pontuando as sentenças conforme a soma da frequência de ocorrência de suas palavras de classe aberta em toda a coleção; o Método 2 leva em conta a localização da sentença no texto, atribuindo etiquetas “início”, “meio” e “fim” às sentenças, priorizando aquelas classificadas como “início”; o Método 3 é baseado na frequência de palavras no texto com tratamento da redundância e dos problemas de TA, consistindo em pontuar as sentenças conforme o Método 1 e tratar a redundância e os problemas de tradução por meio

[10] http://www.summarization.com/mead/

de um cálculo de similaridade lexical baseado na sobreposição de palavras idênticas, denominado *word overlap* — dessa forma, as sentenças consideradas muito semelhantes são descartadas e as que apresentam problemas de agramaticalidade advindo da tradução são substituídas por suas semelhantes no texto original em português; e, por fim, o Método 4 leva em consideração a localização das sentenças nos textos-fonte com tratamento da redundância e dos problemas de TA, que é caracterizado por pontuar as sentenças conforme o Método 2, seguindo os mesmos passos de tratamento do Método 3. Na etapa de síntese, os métodos geram o sumário justapondo as sentenças conforme a ordem em que aparecem nos textos-fonte. Avaliado pela legibilidade, com base nos parâmetros da DUC, o resultado ao que se chegou foi o de que o Método 3, pautado na frequência com tratamento da redundância e da tradução, apresentou um melhor desempenho entre os 4. No entanto, as médias obtidas em cada um dos critérios avaliados evidenciam que os extratos multilíngues gerados por métodos puramente superficiais apresentam problemas de qualidade linguística, que são agravados pela questão da tradução automática.

Tabela 4 – Média das pontuações dos métodos de Tosta et al (2013).

| **Critérios** | **Método 1** | **Método 2** | **Método 3** | **Método 4** |
|---|---|---|---|---|
| Gramaticalidade | 2 | 2,3 | 3 | 2,8 |
| Não-redundância | 2 | 2,8 | 3 | 3 |
| Clareza referencial | 2,8 | 3 | 3,2 | 3 |
| Foco temático | 4 | 3,8 | 4 | 3,8 |
| Coesão e coerência | 2,8 | 2,8 | 2,8 | 2,4 |

*Fonte:* Tosta et al (2013)

Considerando que a SAMM tem como desafio não somente identificar a informação principal da coleção, evitando a redundância, mas também lidar com os problemas de TA decorrentes da tradução integral dos textos-fonte, a investigação de Tosta (2014) traz 2 métodos de SAMM para o português brasileiro, que se distinguem pela estratégia adotada na seleção de conteúdo, como forma de driblar esse problema. Tais métodos são baseados em conhecimento léxico-conceitual, sob o argumento de que uma SA pautada nos conceitos que as palavras veiculam contribui para sumários mais genéricos e informativos, e com menos problemas de redundância.

Para tanto, utilizou-se o CM2News (DI-FELIPPO, 2016), primeiro *corpus* bilíngue (português-inglês) para o português brasileiro. Em ambos os métodos, a etapa de análise consiste em segmentar as sentenças dos textos-fonte e anotar seus substantivos (nomes comuns) com seus respectivos conceitos (isto é, conjunto de *synsets*) da WordNet de

Princeton. Assim, na etapa de transformação, as sentenças são pontuadas em função da frequência de ocorrência de seus conceitos/*synsets* em toda a coleção e é gera-se um ranque com as sentenças mais bem pontuadas. A partir daí, o Método 1 (denominado CFUL, do inglês *concept frequency + user language*) privilegia a língua o usuário — neste caso, o português —, selecionando para compor o sumário apenas as sentenças em português, sob o argumento de que um sumário composto exclusivamente por sentenças originais extraídas dos textos-fonte em português reflete as informações mais relevantes de toda uma coleção bilíngue, já que tais sentenças foram pontuadas e ranqueadas considerando não somente a frequência de ocorrência dos conceitos do texto em português, mas também do texto em inglês. Além disso, esse tipo de estratégia de seleção de conteúdo elimina quaisquer problemas de agramaticalidade, já que não envolve nenhum processo de TA. O Método 2 (denominado CF, do inglês *concep frequency*), por sua vez, seleciona as sentenças mais bem pontuadas do ranque (independentemente da língua) até que se atinja a taxa de compressão; nesse caso, as sentenças em inglês que forem selecionadas são automaticamente traduzidas para o português. O Método 2 é pautado na hipótese de que um sumário composto por sentenças originais em português e sentenças que foram extraídas em inglês e posteriormente traduzidas para o português apresenta menos problemas de agramaticalidade do que se o processo de sumarização fosse aplicado sobre uma coleção com textos originais em português e textos traduzido integralmente para o português, como foi realizado em Tosta et al (2013).

Para garantir que as sentenças selecionadas não sejam redundantes entre si, ambos os métodos, durante a seleção das sentenças no ranque, aplicam um fator de redundância, calculado pela medida *word overlap*, que calcula a similaridade entre um par de sentenças por meio da sobreposição de palavras. Ao fim, na etapa de síntese, tanto o Método 1 quanto o Método 2 juspõem a sentença na ordem em que foram selecionadas e ordena-as pela ordem em que ocorreram nos textos-fonte.

Para avaliar a qualidade linguística e a informatividade, ambos os métodos foram submetidos aos critérios propostos pela DUC'2007 e ao pacote de medidas ROUGE (LIN, 2004); como resultado, observou-se um melhor desempenho do Método 1, que privilegia a língua do usuário, em ambos os quesitos analisados. O Método 1 também se sobressaiu ao Método 3 de Tosta et al (2013), sob paradigma superficial, que propôs uma SAMM baseada na localização das sentenças nos textos-fonte com tratamento da redundância e dos problemas de TA.

Tabela 5 – Comparação dos métodos de Tosta (2014) nos critérios da DUC.

| **Critérios** | **Método 1** | **Método 2** | **Método de Tosta et al (2013)** |
|---|---|---|---|
| Gramaticalidade | 4,3 | 3,5 | 3 |
| Não-redundância | 4,3 | 3,4 | 3 |
| Clareza referencial | 3,7 | 3,3 | 3,2 |
| Foco temático | 4,1 | 3,5 | 4 |
| Coesão e coerência | 3,4 | 2,6 | 2,8 |

*Fonte:* Tosta (2014)

Tabela 6 – Comparação dos métodos de Tosta (2014) da ROUGE.

| Método | Média ROUGE-1 | | | Média ROUGE-2 | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-F** | **Cobertura** | **Precisão** | **Medida-F** |
| **CF** | 0,373642 | 0,3699385 | 0,371175 | 0,174795 | 0,175514 | 0,1748445 |
| **CFUL** | 0,3559595 | 0,32881 | 0,3412685 | 0,155607 | 0,144254 | 0,149433 |

*Fonte:* Tosta (2014)

Como investigações futuras, o trabalho de Tosta (2014) propõe a investigação da influência de conceitos superordinados (em relação de hiponímia e hiperonímia) na informatividade de sumários gerados pela SAMM pautada em conhecimento léxico-conceitual, o que é um dos objetivos do presente trabalho.

Com base na revisão da literatura, os resultados da aplicação de conhecimento léxico-conceitual evidenciam que os métodos de SAMM geram extratos mais informativos, coerentes e coesos. No entanto, há alguns aspectos que podem ser explorados para avançar o estado-da-arte. Um deles diz respeito à pontuação das sentenças, que é feita somente com base na frequência simples dos conceitos na coleção, desconsiderando-se, por exemplo, qualquer tipo de relação entre os conceitos. Outro ponto diz respeito à identificação da redundância, feita com base somente na sobreposição lexical. Tendo em vista que a medida estatística comumente utilizada é a *word overlap*, ressalta-se que esta não captura a similaridade que se dá, por exemplo, entre palavras sinônimas. Assim, nas próximas seções, apresentam-se as etapas de pesquisa deste trabalho, que busca investigar esses dois aspectos ainda não explorados. Na Seção 3, apresentam-se, especificamente, a seleção e a extensão do *corpus* utilizado neste trabalho.

# 3 SELEÇÃO E EXTENSÃO DO *CORPUS*

## 3.1 O *corpus* CM2News

Tendo em vista os objetivos da pesquisa, fez-se necessária a escolha de um *corpus* adequado à investigação, que deveria ser: (i) multidocumento, (ii) multilíngue e (iii) anotado em nível léxico-conceitual. Diante dessas características, selecionou-se o CM2News (DI-FELIPPO, 2016), *corpus* multidocumento composto por 20 coleções bilíngues (português-inglês) de textos jornalísticos. Os 40 textos-fonte do CM2News contabilizam 19.984 palavras. Cada coleção do *corpus* é composta por (i) dois textos-fonte que versam sobre o mesmo assunto, um em inglês e outro em português, (ii) um sumário humano em português, que corresponde a 30% do tamanho do maior texto da coleção (70% de taxa de compressão), e (iii) anotação conceitual dos nomes comuns dos textos-fonte. As coleções dividem-se em 6 domínios: (i) mundo (8 coleções), (ii) saúde (4 coleções), (iii) poder (3 coleções), (iv) ciência (3 coleções), (v) entretenimento (1 coleção) e (vi) meio ambiente (1 coleção).

## 3.2 A inclusão de novas coleções: geração do CM2News (2.0)

Embora o CM2News tenha se mostrado adequado para o desenvolvimento de métodos de SAMM (TOSTA, 2014), esse recurso, segundo as tipologias de *corpus* (cf. SARDINHA, 2004), é relativamente pequeno. Assim, construíram-se 10 novas coleções, contribuindo para o enriquecimento do *corpus* (CAMARGO, DI-FELIPPO, 2019). A construção das novas coleções englobou as tarefas de compilação dos textos-fonte, criação de sumários mutilíngue de referência e anotação conceitual dos nomes dos textos-fonte.

### *3.2.1 A compilação dos textos-fonte*

Para construir as novas coleções, a seleção das notícias seguiu os mesmos critérios de compilação descritos em Di-Felippo (2016). Seguindo as diretrizes da autora, realizou-se a compilação manual das notícias a partir de fontes jornalísticas consideradas confiáveis. No caso, os textos em português foram compilados das versões *online* dos jornais *Folha de São Paulo*[11] e *Uol Notícias*[12], e os textos em inglês, dos portais *online* CNN[13] e BBC[14]. Quanto à atualidade dos textos, compilaram-se notícias publicadas entre abril e outubro de 2018,

[11] Disponível em: https://www.folha.uol.com.br/
[12] Disponível em: https://noticias.uol.com.br/
[13] Disponível em: https://edition.cnn.com/
[14] Disponível em: https://www.bbc.com/

garatindo, assim, que os eventos veiculados por tais notícias fossem atuais (considerando o momento da extensão do *corpus*). Para que a extensão englobasse domínios variados e estes contribuíssem para o balanceamento dos já existentes na primeira versão do CM2News (doravante, CM2News (1.0)), construíram-se: duas coleções ao domínio saúde, uma ao domínio poder, três ao domínio meio ambiente, uma ao domínio ciência e três ao domínio entretenimento. Ao final, os domínios de conhecimento da versão pós-extensão do CM2News (doravante, CM2News (2.0)) estão distribuídos como ilustrado na Tabela 7.

Tabela 7 – Dados estatísticos das diferentes versões do CM2News.

| Coleções/Domínio | CM2News (1.0) (Pré-extensão) | CM2News (2.0) (Pós-extensão) |
|---|---|---|
| Mundo | 8 | 8 |
| Saúde | 4 | 6 |
| Poder | 3 | 4 |
| Ciência | 3 | 4 |
| Entretenimento | 1 | 4 |
| Meio Ambiente | 1 | 4 |
| TOTAL | 20 | 30 |

Fonte: elaborada pela autora.

O tamanho dos textos também foi um parâmetro empregado para a seleção das notícias. Especificamente, para compor uma coleção *x*, buscou-se compilar notícias que tivessem extensão ou tamanho semelhantes. Na coleção C28, por exemplo, que engloba textos sobre "uma baleia morta na Indonésia", tem-se uma notícia em português com 287 palavras e uma notícia em inglês com 359 palavras.

Ademais, para cada nova coleção construiu-se um sumário (multidocumento multilíngue) de referência. Tais sumários foram produzidos por 3 linguistas computacionais com experiência em SA, sendo que cada participante recebeu entre três e quatro coleções. A construção dos sumários de referência foi feita por meio de um formulário *online* durante um período de 7 dias. Para tanto, os participantes seguiram um protocolo básico, que previa (i) leitura de ambos os textos da coleção, (ii) elaboração de um sumário abstrativo, e (iii) aplicação da taxa de compressão de 70% (calculados em número de palavras, sobre o maior texto da coleção).

Na Tabela 8, tem-se os dados quantitativos da versão 2.0 do CM2News. Destaca-se que o referido recurso possui agora 30 coleções bilíngues, totalizando 27.217 palavras. Na Tabela 8, a extensão compreende as coleções de C1 a C30. Na sequência, descreve-se o processo de anotação conceitual das novas coleções do *corpus*.

Tabela 8 – Dados quantitativos do CM2News (2.0).

| Coleção | Domínio | Assunto | Documento | Língua | Publicação | Qt. Palavra | |
|---|---|---|---|---|---|---|---|
| C1 | Mundo | Ataques em Londres | D1_C1_folha | PT | 11/08/2011 – 09:11 | 520 | 1.311 |
| | | | D2_C1_bbc | IN | 11/08/2011 – 11:10 (GMT) | 791 | |
| C2 | Poder | Kit gay | D1_C2_folha | PT | 25/05/2011 – 13:12 | 286 | 516 |
| | | | D2_C2_bbc | IN | 25/05/2011 – 21:07 (GMT) | 230 | |
| C3 | Saúde | Intoxicação alimentar | D1_C3_folha | PT | 30/05/2011 – 18:47 | 716 | 1.419 |
| | | | D2_C3_bbc | IN | 30/05/2011 – 5:43 (GMT) | 703 | |
| C4 | Mundo | Massacre na Noruega | D1_C4_folha | PT | 08/08/2011 – 14h20 | 356 | 911 |
| | | | D2_C4_bbc | IN | 02/08/2011 – 14:52 (GMT) | 555 | |
| C5 | Ambiente | Novo código florestal | D1_C5_folha | PT | 25/05/2011– 00:43 | 670 | 1.217 |
| | | | D2_C5_bbc | IN | 25/05/2011– 09:50 (GMT) | 547 | |
| C6 | Mundo | Conflito na universidade da CA | D1_C6_folha | PT | 20/11/2011– 00:15 | 289 | 645 |
| | | | D2_C6_bbc | IN | 21/11/2011– 23:26 (GMT) | 356 | |
| C7 | Saúde | Proibição do fumo em NY | D1_C7_folha | PT | 24/05/2011– 13:38 | 370 | 887 |
| | | | D2_C7_bbc | IN | 24/05/2011– 18:36 (HKT) | 517 | |
| C8 | Mundo | Terremoto na Nova Zelândia | D1_C8_folha | PT | 05/03/2011– 05:01 | 397 | 948 |
| | | | D2_C8_bbc | IN | 03/03/2011– 04:45 (GMT) | 551 | |
| C9 | Mundo | Terremoto em Missouri | D1_C9_folha | PT | 23/05/2011– 08:04 | 479 | 1.169 |
| | | | D2_C9_bbc | IN | 23/05/2011– 20:21 (GMT) | 690 | |
| C10 | Mundo | Erupção vulcânica na Islândia | D1_C10_folha | PT | 24/05/2011– 12:13 | 770 | 1.476 |
| | | | D2_C10_bbc | IN | 24/05/2011– 15:51 (GMT) | 706 | |
| C11 | Ciência | Patentes genes humanos | D1-C11_bbc | IN | 13/07/2013- 16:34 (GMT) | 454 | 963 |
| | | | D2_C11_folha | PT | 13/06/2013-23:50 | 509 | |
| C12 | Poder | Protestos: transporte | D1_C12_folha | PT | 14/06/2013-07:25 | 290 | 808 |
| | | | D2_C12_bbc | IN | 14/06/2013-12:43 (GMT) | 518 | |
| C13 | Mundo | Eleições do Irã | D1_C13_folha | PT | 15/06/2013 – 17:57 | 581 | 1.266 |
| | | | D2_C13_bbc | IN | 16/06/2013 - 08:38 (GMT) | 685 | |
| C14 | Saúde | Epidemia de dengue no MS | D1_C14_folha | PT | 11/01/2013 1-9:03 | 321 | 534 |
| | | | D2_C14_bbc | IN | 21/01/2013- 00:21 (GMT) | 213 | |
| C15 | Saúde | Mastectomia preventiva | D1_C15_folha | PT | 15/05/2013 – 03:01 | 603 | 1.367 |
| | | | D1_C15_bbc | IN | 14/05/2013 -17:02 (GMT) | 764 | |
| C16 | Ciência | Missão espacial chinesa | D1_C16_folha | PT | 11/06/2013 – 21:06 | 346 | 793 |
| | | | D2_C16_bbc | IN | 11/06/2013-9:38 (GMT) | 447 | |
| C17 | Poder | Protesto: copa das confederações | D1_C17_folha | PT | 15/06/2013 – 14:53 | 640 | 918 |
| | | | D2_C17_bbc | IN | 16/06/2013 -13:19 (GMT) | 278 | |
| C18 | Ciência | Viagra feminino | D1_C18_folha | PT | 16/06/2013 – 03:30 | 670 | 975 |
| | | | D2_C18_bbc | IN | 17/11/2009- 9:35 (GMT) | 305 | |
| C19 | Entreten. | Lançamento: homem de aço | D1_C19_folha | PT | 16/06/2013-13:24 | 441 | 898 |
| | | | D2_C19_bbc | IN | 11/06/2013-10:17(GMT) | 457 | |
| C20 | Mundo | Conflito na Turquia | D1_C20_folha | PT | 17/06/2013 - 09h44 | 515 | 963 |
| | | | D2_C20_bbc | IN | 17/06/2013-13:00(GMT) | 448 | |
| C21 | Poder | Encontro líderes das Coreias | D1_C21_folha | PT | 27/04/2018 - 00:34 | 386 | 770 |
| | | | D2_C21_bbc | IN | 27/04/2018 - 08:06 (GMT) | 384 | |
| C22 | Ciência | Reprodução de camundongos | D1_C22_folha | PT | 11/10/2018 - 12:00 | 578 | 1.240 |
| | | | D2_C22_bbc | IN | 11/10/2018 - 9:46 (GMT) | 662 | |
| C23 | Entreten. | Kanye West na política | D1_C23_uol | PT | 30/10/2018 - 19:49 | 328 | 782 |
| | | | D2_C23_bbc | IN | 31/10/2018 - 7:57 | 454 | |
| C24 | Entreten. | Bebê de Hilary Duff | D1_C24_folha | PT | 30/10/2018 - 11:00 | 182 | 285 |
| | | | D2_C24_cnn | IN | 30/10/2018 - 15:21 (GMT) | 103 | |
| C25 | Entreten. | Acusações ao Stallone | D1_C25_uol | PT | 31/10/2018 - 05:05 | 150 | 280 |
| | | | D2_C25_bbc | IN | 31/10/2018 - 1:45 (GMT) | 130 | |
| C26 | Ambiente | Oleoduto EUA-Canadá | D1_C26_folha | PT | 09/11/2018 - 17:55 | 428 | 973 |
| | | | D2_C26_bbc | IN | 09/11/2018 - 14:32 (GMT) | 545 | |
| C27 | Ambiente | Ataque de leoa | D1_C27_folha | PT | 22/10/2018 - 16:15 | 220 | 419 |
| | | | D2_C27_bbc | IN | 22/10/2018 - 10:11 (GMT) | 199 | |
| C28 | Ambiente | Baleia morta na Indonésia | D1_C28_uol | PT | 21/11/2018 - 11:21 | 287 | 646 |
| | | | D2_C28_cnn | IN | 21/11/2018 - 16:57 (GMT) | 359 | |
| C29 | Saúde | EUA poliomielite | D1_C29_folha | PT | 17/10/2018 - 8:00 | 390 | 782 |
| | | | D2_C29_cnn | IN | 23/10/2018 - 12:27 (GMT) | 392 | |
| C30 | Saúde | Camisinha autolubrificante | D1_C30_uol | PT | 18/10/2018 - 12:17 | 522 | 956 |
| | | | D2_C30_cnn | IN | 19/10/2018 - 15:20 (GMT) | 434 | |
| **TOTAL** | | | | | | | 27.217 |

Fonte: elaborado pela autora.

### *3.2.2 A criação dos sumários de referência*

Uma vez que os textos-fonte das 10 novas coleções foram compilados, procedeu-se à produção dos sumários multilíngue de referência para cada uma dessas coleções.

A produção feita por 3 linguistas computacionais com experiência em sumarização automática, seguindo o mesmo protocolo utilizado por Tosta (2014) para a construção dos sumários de referência das coleções do CM2News (1.0).

Assim, dada uma coleção *x*, produziu-se seu sumário de referência com base nas seguintes diretrizes: (i) ler ambos os textos-fonte (português e inglês) da coleção *x*; (ii) elaborar um sumário abstrativo, isto é, com livre reescrita dos textos-fonte, (iii) elaborar um sumário (abstrativo) informativo, isto é, que contenha a informação central dos textos-fonte a ponto de sua leitura substituir a leitura da coleção de textos; (iv) produzir um sumário abstrativo informativo com tamanho equivalente a 30% (em número de palavras) do maior texto da coleção, já que a taxa de compressão de 70% também fora utilizada no CM2News (1.0); (v) escrever um sumário de referência em português, posto que esta é a língua-alvo da sumarização em questão.

Na sequência, descreve-se a anotação dos textos-fonte das novas coleções do *corpus.*

### *3.2.3 A anotação léxico-conceitual*

A anotação das 10 novas coleções do CM2News (2.0) foi realizada por 1 linguista computacional. Especificamente, o especialista realizou a anotação em um total de 20 dias, em sessões diárias de 60 a 90 minutos.

Seguindo as características da versão prévia do *corpus,* a anotação em questão consistiu na explicitação dos conceitos nominais por meio de rótulos advindos da WN.Pr. A respeito da anotação realizada por Tosta (2014) e seguida neste trabalho, cabem aqui dois destaques, um sobre a classe dos nomes e outro sobre a WN.Pr.

A escolha dos nomes pautou-se no fato de que tal categoria gramatical é, entre as unidades de classe aberta, a mais frequente, atuando na veiculação do conteúdo semântico principal dos textos[15].

Ademais, a escolha da WN.Pr como repositório de conceito se deveu, além do fato de não existir, à época da construção do CM2News (1.0), uma ontologia suficientemente robusta de língua geral em português que fosse computacionalmente tratável, pela sua (i) adequação linguística, uma vez que simula a organização do léxico mental, e (ii) abrangência, posto que é uma das ontologias mais extensas do inglês.

---

[15] De Luca (2019), ainda que de forma incipiente, investigou a combinação da frequência dos nomes e dos verbos na SAM e os resultados evidenciaram que essa união não afeta a informatividade dos extratos.

Em linhas gerais, a WN.Pr consiste em uma rede de palavras e expressões (englobando a categoria dos nomes, verbos, adjetivos e advérbios), cujos conceitos a elas subjacentes são representados por meio de *synsets* (do inglês *synonym sets*), isto é, conjuntos de formas de uma mesma categoria gramatical que podem ser intercambiáveis em determinado contexto sem perda de significado, como *"car, auto, automobile, machine, motorcar"*. Entre os *synsets*, pode-se estabelecer 5 relações lógico-conceituais distintas, a saber: antonímia, hiponímia, meronímia, acarretamento e causa (LYONS, 1979, CRUSE, 1986, FELLBAUM, 1998).

Essas relações podem ser assim definidas: (i) "antonímia" é uma relação que se estabelece com base em diferentes tipos de oposição semântica; (ii) "hiperonímia/hiponímia" é a relação entre um conceito mais genérico (hiperônimo) e um conceito mais específico (hipônimo); (iii) "meronímia/holonímia" é a relação entre um *synset* que expressa um "todo" e outros *synsets* que expressam partes do todo; (iv) "acarretamento" é a relação que se estabelece entre uma ação A1 e uma ação A2, como os conceitos de "correr" e "mover-se" e (v) "causa" é a relação que se estabelece entre uma ação A1 e uma ação A2 quando a ação A1 denotada pelo verbo *x* causa a ação A2 denotada pelo verbo *y*, como "matar" e "morrer".

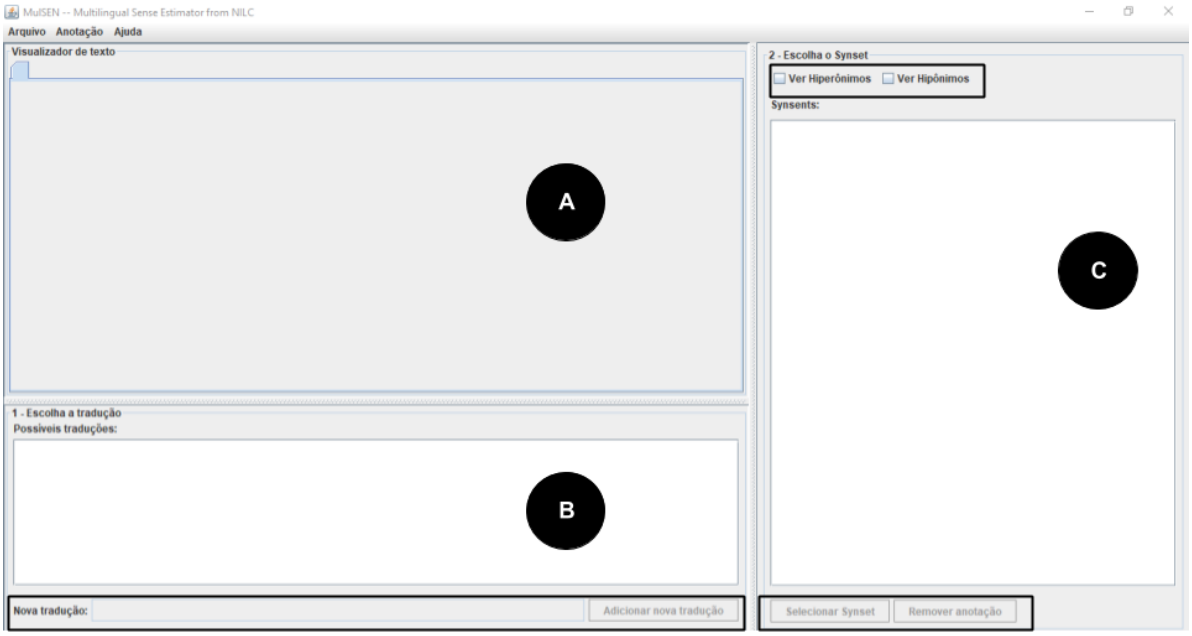
Para anotar os conceitos subjacentes às unidades lexicais dos novos textos do agora CM2News (2.0), utilizaram-se o editor MulSen e as diretrizes de Tosta (2014), os quais foram aplicados na anotação do CM2News (1.0).

*a) O editor MulSen e suas funcionalidades gerais*

O editor MulSen foi desenvolvido como uma versão multilíngue do editor NASP (NÓBREGA, 2013) especificamente para a tarefa de anotação semântica pautada nos conceitos da WN.Pr, em coleções multilíngues. Trata-se de um editor que possibilita a anotação simultânea de até dois textos, contando com uma etapa de pré-processamento que envolve a etiquetação morfossintática automática e a desambiguação lexical de sentido (DLS). Como pode ser observado na Figura 7, a interface é composta por 3 telas principais.

A janela "Visualizador de texto" (A) exibe os textos-fonte da coleção ao anotador. Na abertura de dois ou mais textos-fonte de uma coleção, essa janela compõe-se por duas ou mais abas de exibição, uma para cada texto. A exibição dos textos não é simultânea, no entanto, o acesso a ambos pode ser feito durante todo o processo de anotação, podendo o anotador passar de um texto para o outro sempre que desejar.

Figura 7 – Interface do editor MulSen.




Fonte: elaborada pela autora.

Na janela "Escolha a tradução" (B), apresentam-se as possíveis traduções para o inglês de uma palavra que não seja desse idioma (p.ex.: português). Para a anotação de uma palavra *w* de um texto em inglês, o MulSen exibe a própria palavra *w* na janela B. Caso o editor não sugira traduções adequadas, o anotador pode inserir um equivalente em inglês no campo "Nova tradução" (parte inferior da janela B). Após a escolha de um equivalente em inglês, o MulSen exibe os *synsets* constituídos por esse equivalente e que podem representar o conceito da palavra a ser anotada. As possíveis traduções são provenientes do acesso ao dicionário *online* WordReference[16].

Na janela "Escolha o *synset*" (C), exibem-se todos os *synsets* que possuem o equivalente em inglês como elemento constitutivo. Na parte superior dessa janela, há a opção de exibição dos hiperônimos ("Ver Hiperônimos") e hipônimos ("Ver Hipônimos") dos *synsets* listados. Tal opção é importante para desambiguar sentidos, considerando que os *synsets* da WN.Pr podem não cobrir conceitos ligados à realidade cultural de outras línguas, como o português. Na parte inferior, há os botões que permitem a seleção do *synset* adequado e a efetiva anotação, isto é, a associação do *synset* escolhido ao nome do texto.

Uma vez que a tela principal tenha sido descrita, segue uma descrição geral sobre as funcionalidades de anotação do MulSen.

[16] Disponível em: http://www.wordreference.com/

Ao abrir os textos (português-inglês) de uma coleção, o editor automático realiza 2 tarefas de pré-processamento, a saber: (i) etiquetação morfossintática (em inglês, *part-of-speech* ou PoS *tagging*) dos nomes, e (ii) desambiguação lexical de sentido (DLS, do inglês, *word sense disambiguation*) dos nomes para auxiliar na identificação dos conceitos nominais.

A etiquetação morfossintática consiste na atribuição da categoria gramatical a cada uma das palavras de um texto. Para os textos em português, o MulSen utiliza o etiquetador MXPOST (RATNAPARKHI, 1986) e, para os textos-fonte em inglês, o editor utiliza o *TreeTagger* (SHIMID, 1994).

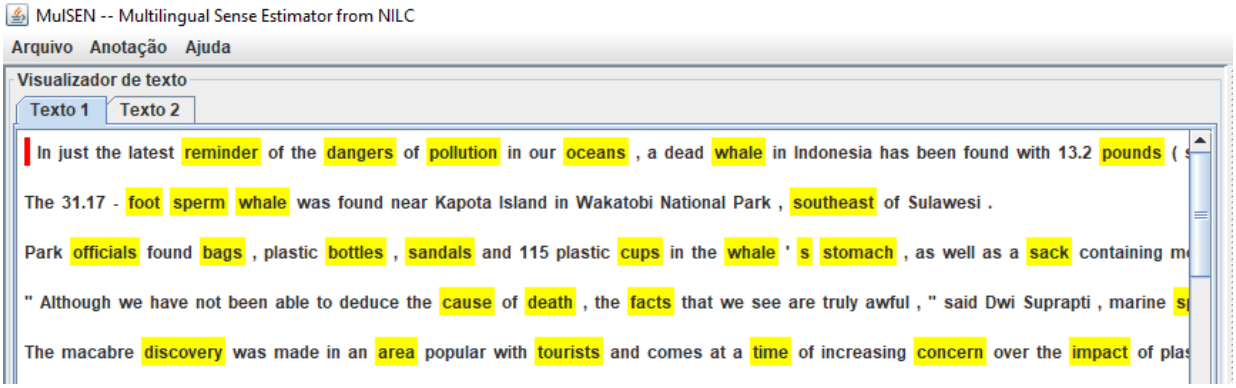
Para a DLS, somente as palavras etiquetadas como "nome" são consideradas. A DLS corresponde à detecção do conceito subjacente a uma palavra, dados o contexto em que ocorre e um repositório de conceitos (AGIRRE; EDMONDS, 2006). No caso do MulSen, a DLS consiste em identificar, para cada nome do texto em português e em inglês, o *synset* da WN.Pr que mais adequadamente codifica o conceito expresso por ele. Para a identificação do conceito mais adequado, o editor emprega uma versão do algoritmo de Lesk (1986) (NÓBREGA, 2013). Com base no algoritmo, o editor atribui somente um conceito a todas as ocorrências de uma mesma palavra em uma coleção de documentos[17].

Na Figura 8, a aba "Texto 1" exibe, por exemplo, o texto-fonte em inglês da C28 (sobre "uma baleia encontrada morta na Indonésia") do CM2News (2.0) após a etiquetação e a DLS. Na aba "Texto 2", tem-se o texto em português, que não está à mostra na Figura. Na aba "Texto 1", os nomes que ocorrem no texto estão destacados em amarelo ou vermelho. As palavras em amarelo foram etiquetadas como nome e desambiguadas, ou seja, reconhecidas como nome pelo *tagger* e posteriormente anotadas pelo método de DLS, sendo, assim, associadas a *synsets* que precisam ser confirmados (ou não) pelo anotador. As palavras em vermelho foram apenas etiquetadas em nível morfossintático e não desambiguadas. Isso significa que a DLS não foi capaz de sugerir *synsets* porque a palavra em questão não está contemplada na WN.Pr[18]. Nesses casos, o anotador pode sugerir um sinônimo para verificar se o conceito/*synset* está armazenado sob esse rótulo e não sob o que ocorreu de fato no texto. Caso contrário, pode-se buscar por um conceito mais genérico.

[17] Essa abordagem resulta da verificação em *corpus* de que as palavras tendem a ocorrer com o mesmo sentido em textos que abordam um mesmo assunto. Somente após as tarefas de etiquetação e DLS, os textos-fonte são exibidos aos anotadores.

[18] No caso dos textos em português, as palavras destacadas em vermelho decorrem sobretudo da não sugestão de um equivalente de tradução pelo acesso ao WordReference.

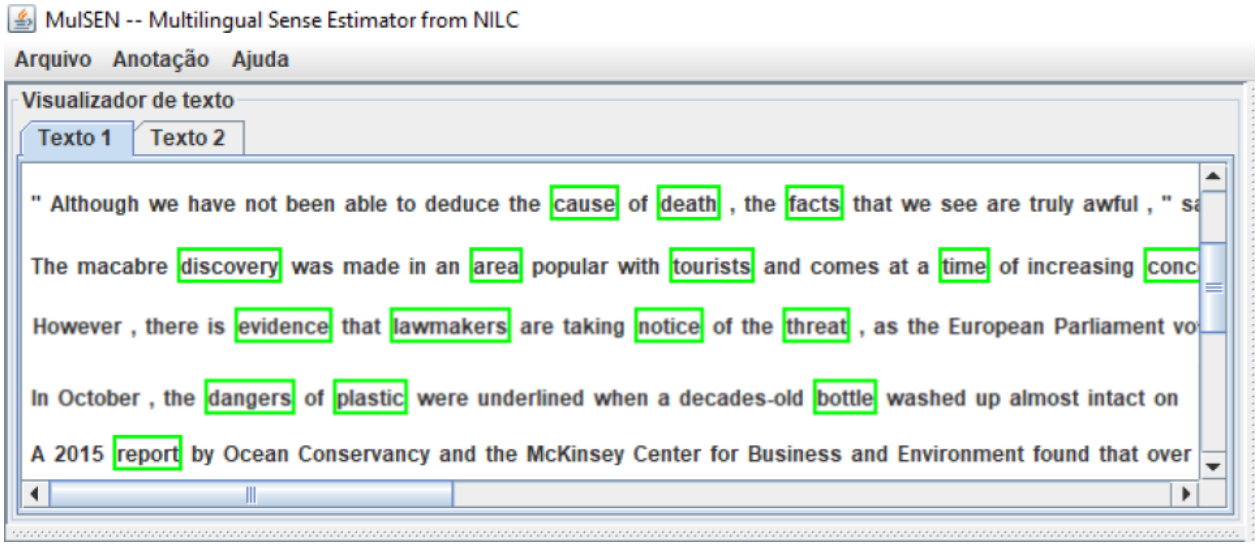
Figura 8 - Exibição do texto-fonte em inglês após a etiquetação e DLS.




Fonte: elaborada pela autora.

Após a seleção do *synset* adequado (seja ele sugerido pelo editor ou escolhido pelo anotador humano) para cada palavra e a confirmação da anotação, as palavras do texto são destacadas em verde, o que indica que a anotação léxico-conceitual foi finalizada (cf. Figura 9).

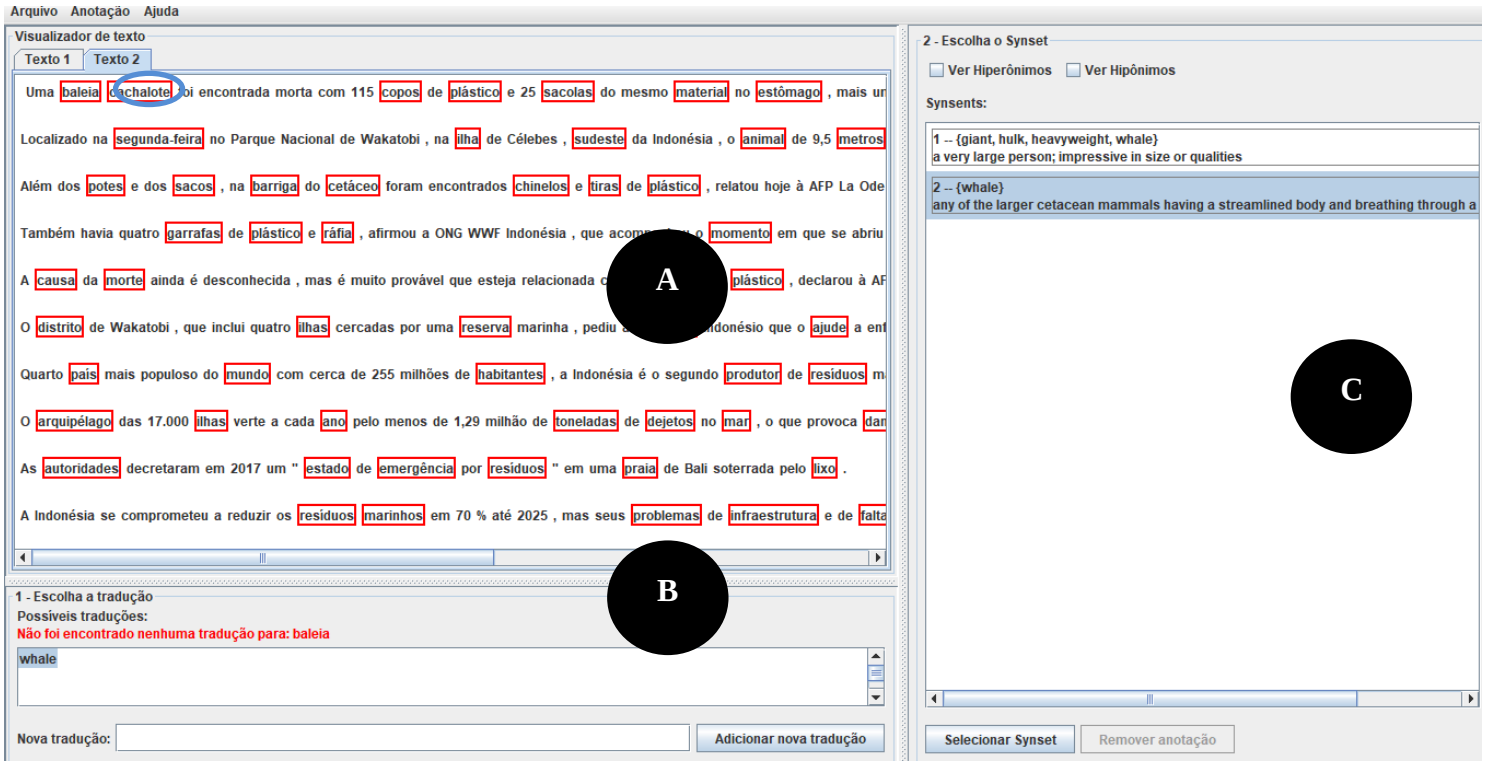
Figura 9 – Exibição do texto-fonte em inglês após anotação.




Fonte: elaborada pela autora.

No caso da anotação das 10 novas coleções que compõem o CM2News (2.0), nem todas as funcionalidades automáticas do MulSen funcionaram. Para exemplificar o processo de anotação da versão 2.0 do *corpus*, considera-se o primeiro nome (círculo azul) do texto em português da Figura 10 (“baleia”). No caso, tal texto é proveniente da coleção C28 do CM2News (2.0), que engloba notícias sobre “uma baleia encontrada morta na Indonésia”.

Figura 10 – Exibição do texto-fonte em português para anotação conceitual.




Fonte: elaborada pela autora.

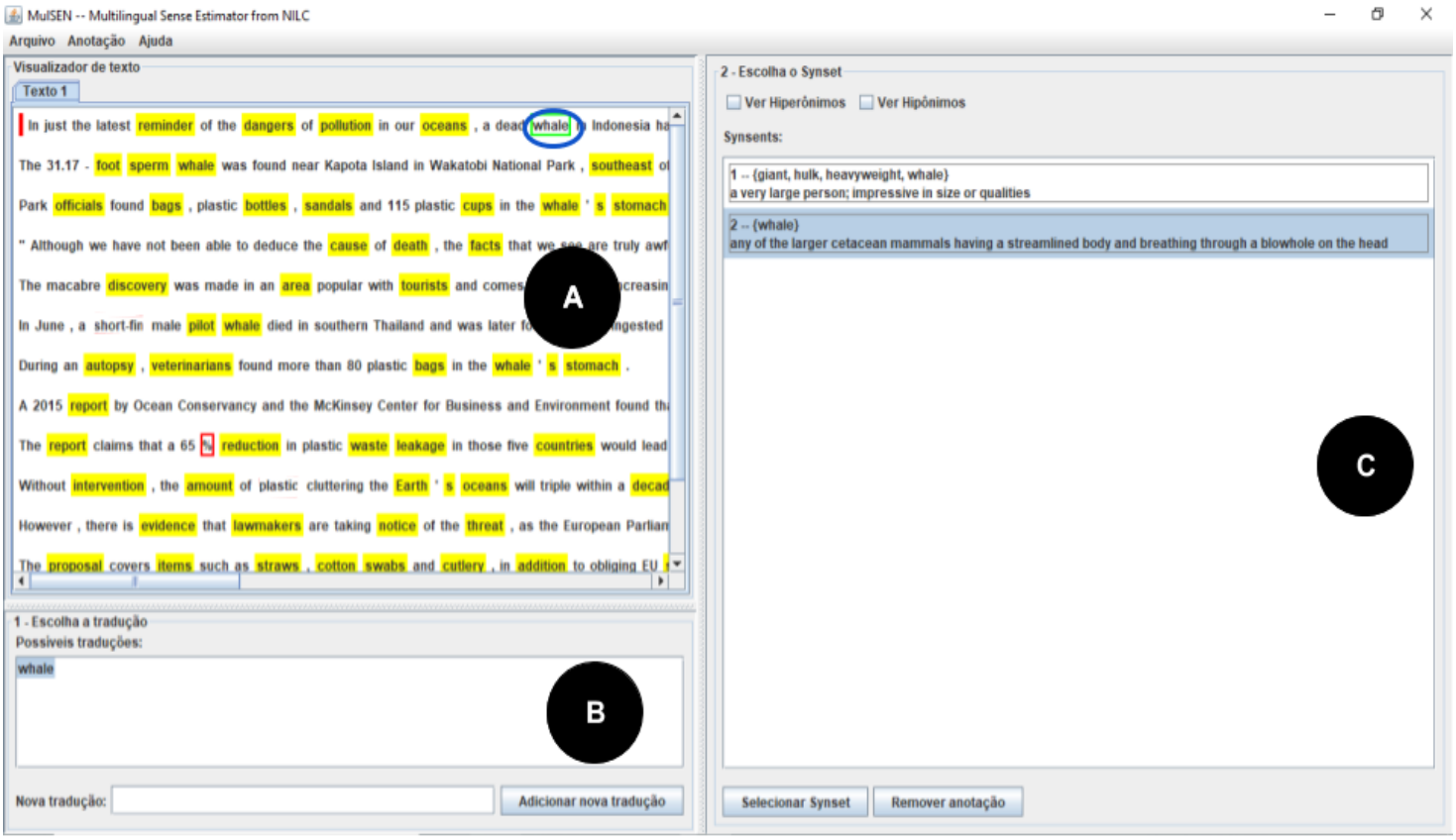
Na janela “Visualizador de texto” (A), tem-se o texto-fonte em português pré-processado automático. Devido a problemas com a API[19] de acesso ao dicionário WordReference, o editor não foi capaz de sugerir equivalentes de tradução para os nomes identificados pelo *tagger*, o que não permitiu, por conseguinte, a aplicação do algoritmo de DLS. Dessa forma, os equivalentes de traduções para cada nome da coleção foram inseridos manualmente no campo “Nova tradução” (B). No caso da anotação de “baleia”, inseriu-se o equivalente “*whale*”, permitindo que o editor recuperasse 2 *synsets* da WN.Pr que possuem tal item como elemento constitutivo. Tais *synsets* estão exibidos na janela C (Figura 10). Analisando as opções, conclui-se que o *synset* 2 representa adequadamente o conceito subjacente ao nome “baleia” do texto em português. Ao clicar em “Selecionar *Synset*” e confirmar a escolha, a palavra “baleia” (assim como todas as suas demais ocorrências) fica associada ao *synset* em questão e graficamente destacada em verde.

Para anotação do texto-fonte em inglês (Figura 11), o pré-processamento automático (isto é, *tagging* e DLS) funcionou como o previsto. Assim, na Figura 11, há palavras destacadas em vermelho e em amarelo. Considerando “*whale*” (circulado em azul), o editor

[19] Sigla para *Application Programming Interface*. Trata-se de um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web.

recuperou os 2 *synsets* constituídos por esse item lexical, os quais estão exibidos na janela C. A DLS sugere o *synset* 2 como sendo a representação mais provável do conceito, destacando-o de azul. Nesse caso, a sugestão foi confirmada pelo anotador humano, o que fez a palavra em questão ser associada a tal *synset*, recebendo o destaque verde no texto.

Figura 11 – Exibição do texto-fonte em inglês para anotação conceitual.




Fonte: elaborada pela autora.

Após a anotação conceitual dos textos de uma coleção, o MulSen gera um arquivo no formato XML (do inglês, *Extensible Markup Language*), um dos mais utilizados para a tarefa de anotação de *corpus.* Na Figura 12, há três blocos de informação que indicam: anotador, texto-fonte e palavra anotada. No bloco que se refere especificamente à anotação de uma palavra (*token*), apresentam-se as seguintes informações: (i) *Word* (unidade lexical anotada, isto é, "*whale*"), (ii) *Tag* (etiqueta sintática, que, no caso, indica núcleo de sintagma nominal) e *MorfoTag* (etiqueta morfossintática, isto é, "substantivo"), (iii) *Lemma* (forma canônica, ou seja, "whale"), (iv) *Translations* (forma de inserção/seleção da tradução e equivalente

especificamente selecionado)[20], e (v) *Synsets* (lista de *synsets* constituídos pelo equivalente de tradução e o *synset* selecionado dentre eles). Na Figura 12, vê-se que, dentre os *synsets* da WN.Pr que possuem "whale" como um de seus elementos constitutivos, o anotador selecionou o *synset* codificado pelo ID 2062744, indicado pelo valor "true".

Figura 12 – Anotação léxico-conceitual em XML gerada pelo MulSen.

```
<?xml version="1.0" encoding="UTF-8"?>
<save>
    <Annotators>
        <Annotator id="1">Anotador 1</Annotator>
    </Annotators>
    <Files>
        <Text language="ENGLISH" name="en_D2_C8_cnn_21-11-2018.tagged">
            <Token>
                <Word>whale</Word>
                <Tag>NN</Tag>
                <MorphoTag>Substantivos</MorphoTag>
                <Lemma>whale</Lemma>
                <Type>ANNOTED</Type>
                <Translations manual_translation="false">
                    <Translate selected="true">whale</Translate>
                </Translations>
                <Synsets>
                    <Synset selected="true">2062744</Synset>
                    <Synset selected="false">10129133</Synset>
                </Synsets>
            </Token>
```

Fonte: elaborada pela autora.

Para manter a coerência entre a anotação conceitual da primeira e da segunda versão do *corpus*, utilizaram-se as mesmas diretrizes de anotação que Tosta (2014) empregou ao construir o CM2News (1.0), as quais são descritas na sequência.

[20] Na anotação da palavra "*whale*", o atributo "*translations manual_translation*" é especificado com o valor "false" porque a inserção de uma tradução não é necessária, já que a palavra original já está em inglês. O valor "true" para "translate selected" somente indica que a palavra original do texto foi utilizada para a recuperação dos *synsets* da WN.Pr.

*b) As diretrizes de anotação de Tosta (2014) aplicadas ao CM2News (2.0)*

Segundo o autor, tais diretrizes estão divididas em 2 grandes grupos, a saber: (i) regras gerais e (ii) regras específicas (Quadro 3).

Quadro 3 – Diretrizes de anotação do *corpus* CM2News (2.0).

| **Regras gerais (RG)** | |
|---|---|
| RG1 | Ler cuidadosamente os textos-fonte de cada coleção |
| RG2 | Iniciar a anotação preferencialmente pelo texto-fonte em inglês da coleção |
| RG3 | Anotar todos os nomes comuns e siglas do *corpus* |
| RG4 | Refinar a anotação morfossintática automática |
| RG5 | Ignorar as palavras anotadas equivocadamente como nome |
| RG6 | Selecionar o mesmo *synset* para anotar diferentes expressões linguísticas do mesmo conceito na coleção |
| **Regras específicas (RE)** | |
| RE1 | Anotar somente o nome nuclear das expressões multipalavras |
| RE2 | Anotar todos os nomes de sintagmas recorrentes livres |
| RE3 | Analisar todas as traduções sugeridas pelo MulSen e os respectivos *synsets* |
| RE4 | Testar diferentes equivalentes antes de adicionar uma tradução ao MulSen |
| RE5 | Selecionar os *synsets* mais adequados para anotar os nomes |
| RE6 | Selecionar *synsets* hiperônimos |
| RE7 | Anotar o núcleo das expressões metafóricas |

Fonte: Tosta (2014).

A RG1 estabelece que anotação deve iniciar com a leitura integral e cuidadosa dos textos-fonte de cada coleção antes de se iniciar o processo de anotação. Tal regra pauta-se na ideia de que uma leitura prévia é util para familiarizar o anotador quanto ao conteúdo da coleção, bem como propiciar uma compreensão global do texto a ser anotado.

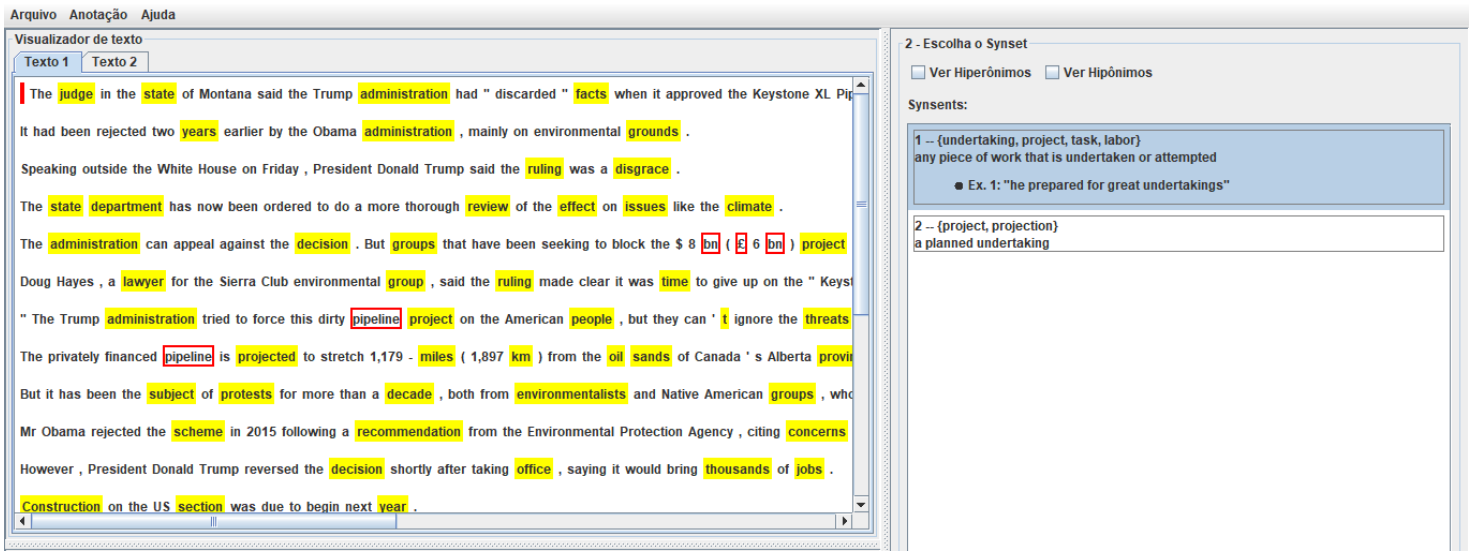
A RG2 estabelece que a anotação seja iniciada pelo texto em inglês da coleção. Tal regra foi delimitada porque o texto em inglês pode fornecer diretamente os equivalentes de tradução para a anotação das palavras em português.

A RG3 estabelece que a anotação deve contemplar os nomes comuns (com exceção dos nomes próprios) e as siglas presentes nos textos-fonte, pois, no gênero jornalístico, as siglas são frequentes, veiculando conteúdo que pode ser relevante na coleção.

A RG4 estabelece que a anotação morfossintática automática deve ser revisada, uma vez que os *taggers* não são ferramentas 100% precisas. Especificamente, essa regra estabelece que os casos de silêncio (ou seja, nomes não etiquetados automaticamente pelos *taggers*) sejam identificados pelo anotador para que o MulSen realize os demais procedimentos que permitem anotação conceitual final.

A RG5 estabelece que os casos de ruídos (ou seja, palavras anotadas equivocadamente como nome) produzidos pelos *taggers* sejam ignorados, não sendo alvo da anotação conceitual. Esse é o caso, por exemplo, do verbo "*projected*" (circulado em azul) da Figura 13. Tal verbo foi erroneamente anotado como nome, ao qual se sugeriu o *synset* {*undertaking, project, task, labor*} ("*any piece of work that is undertaken or attempted*") ("trabalho que seja realizado ou testado"). Com base na RG5, clicou-se no botão "Remover anotação". Uma vez que anotação tenha sido removida, a palavra continua com o destaque em vermelho, mas esta deve ser ignorada como nome.

Figura 13 – Ilustração de um caso de ruído gerado pelo tagging.

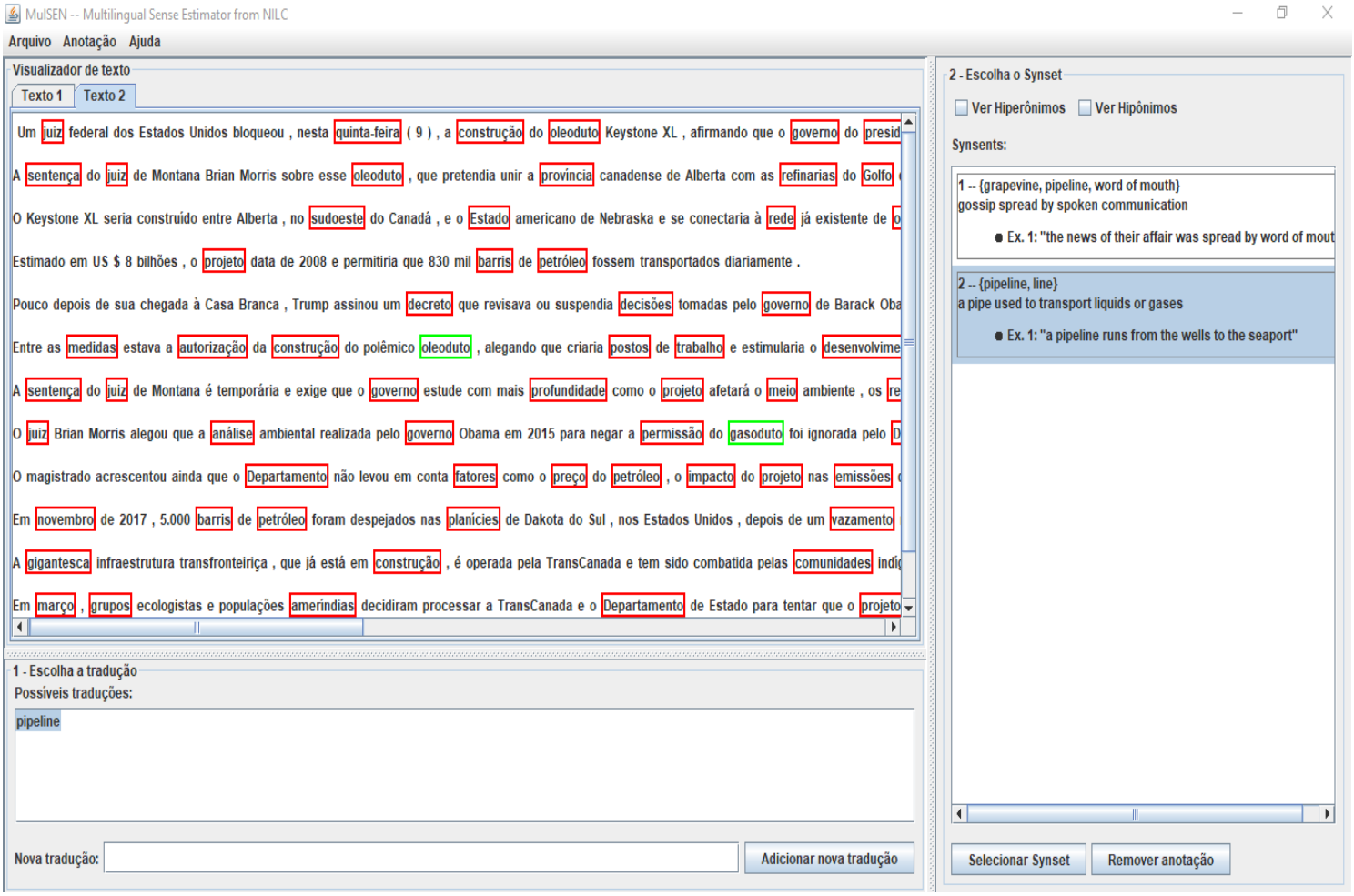



Fonte: elaborada pela autora.

A RG6 estebelece que o anotador deve garantir a seleção do mesmo *synset* para anotar: (i) todas as ocorrências de uma palavra *x*, com o sentido *y* no mesmo texto; (ii) as ocorrências de palavras sinônimas de *x* no mesmo texto, e (iii) as ocorrências dos equivalentes de *x* no outro texto da coleção. Essa regra implica a consulta frequente a ambos os textos por meio

das abas “Texto 1” e “Texto 2” durante a anotação de uma mesma coleção. Seguindo-se a RG6, todas as ocorrências da palavra “oleoduto” com o sentido de “um tubo usado para transportar líquidos ou gases” no texto em português da C26 (Figura 14) foram anotadas com o *synset {pipeline, line}* (“*a pipe used to transport liquids os gases*”). Ademais, “gasoduto”, interpretado como sinônimo de “oleoduto”, foi anotada com o mesmo *synset, {pipeline, line}*. No texto em inglês, a palavra “pipeline” é o equivalente de “oleoduto” e “gasoduto”, sendo, portanto, anotada com o mesmo *synset, {pipeline, line}*.

Figura 14 – Diferentes expressões de um mesmo conceito.




Fonte: elaborada pela autora.

No que diz respeito às regras específicas, a RE1 foi estabelecida porque os *taggers* somente anotam unidades simples (isto é, sequências de caracteres separadas por espaços em branco), não abrangendo expressões multipalavras. Assim, essa regra estabelece que os nomes etiquetados isoladamente em nível morfossintático, mas que são, na verdade, núcleos de expressões multipalavras, sejam anotados com *synsets* que representem os conceitos expressos pelas expressões multipalavras, desde que haja tais *synsets*. Esse é o caso, por

exemplo, do nome “fluidos”, que é núcleo da expressão “fluidos corporais” e foi anotado com o *synset {liquid body substance, bodily fluid, body fluid, humor, humour}* (“*the liquid parts of the body”*)[21], posto que este representa o conceito subjacente à expressão toda. Tal regra foi estabelecida posto que unigramas compõem uma das características de análise de línguas que permitem a lexicalização de grande parte dos conceitos.

A RE2 estabelece que, em casos de sintagmas recorrentes livres[22] (SLRs) (do inglês *recurrent free phrases*), todos os nomes que compõem a expressão devem ser anotados com seus respectivos conceitos. Assim, acredita-se que seja possível codificar o conceito representado pela expressão como um todo. Um exemplo da aplicação dessa regra está na anotação da C28. No texto-fonte em português, identificou-se o SLR “estado de emergência”. Segundo a RE2, ambos os nomes, “estado” e “emergência”, devem ser anotados. Especificamente, “estado” foi anotado com o *synset* {*state*} (“*the way something is with respect to its main attributes*”) (“a maneira como algo é com relação aos seus principais atributos”) e “emergência” foi anotado com o *synset* {*emergency*} (“*a sudden unforeseen crisis (usually involving danger) that requires immediate action*”) (“uma repentina crise imprevista (geralmente envolvendo perigo) que requer ação imediata”).

A RE3 estabelece que todas as traduções sugeridas pelo MulSen (as quais foram recuperadas do WordReference) sejam analisadas antes da seleção definitiva do equivalente de tradução, assim como os *synsets* sugeridos para cada uma delas. Tosta (2014) estabeleceu essa regra com o objetivo de garantir a seleção do equivalente de tradução mais adequado em inglês. Como na anotação do CM2News (2.0) os equivalentes foram inseridos manualmente, a RE3 não foi efetivamente aplicada.

Ao contrário da RE3, a RE4 foi bastante utilizada. Tal regra estabelece que diferentes traduções, quando existentes, sejam testadas antes de se adicionar a expressão em inglês ao editor. O motivo para o estabelecimento da RE4 foi o fato de que, por vezes, uma palavra *y* inserida pelo anotador como equivalente de tradução não consta na WN.Pr. Isso, no entanto, não significa necessariamente que o conceito não está codificado na base, mas sim que a unidade *y* não está armazenada na base de dados. Por conseguinte, a RE4 determina que as várias possibilidades de tradução, quando existentes, sejam testadas para que possíveis *synsets* correspondentes sejam recuperados da WN.Pr. Ressalta-se que, para sugerir um equivalente de traduação, o anotador pode recorrer a recursos externos ao editor MulSen,

[21] “As partes líquidas do corpo” (tradução nossa).
[22] Por SLR, entende-se uma combinação de palavras que, apesar de frequente, apresentam baixo grau de estabilidade e fixação (BENTIVOGLI; PIANTA, 2003).

como dicionários e serviços *online.* Entre os dicionários, destacam-se a versão *online* do Michaelis Moderno Dicionário Inglês & Português[23] e os diferentes dicionários disponíveis no site *Cambridge Dictionaries Online*[24]. Quanto aos serviços *online,* utilizaram-se o *Google translate*[25] e o Linguee[26].

A RE5 estabelece que o anotador deve selecionar o *synset* que mais adequadamente representa o conceito subjacente a um nome *x.* Em outras palavras, a RE5 determina que, uma vez que todos os *synsets* recuperados pelo MulSen (inclusive o sugerido pela DLS) tenham sido analisados, o *synset* mais adequado seja selecionado para a anotação. Essa regra foi formulada principalmente porque a WN.Pr, por vezes, indica que uma palavra representa conceitos muito similares, cuja distinção nem sempre é simples. Esse é o caso, por exemplo, do nome "cerimônia" da C21 (5), que aborda o encontro entre os líderes da Coreia do Sul e da Coreia do Norte.

(5) "[...] O encontro deste 27 de abril é o ápice da distensão iniciada com um discurso de 1º de janeiro por Kim e continuada com a participação de atletas do Norte e de uma equipe mista na Olimpíada de Inverno no Sul, na qual Kim Yo-yong, irmã do ditador, assistiu à **cerimônia** de abertura. [...]"

Uma vez que o equivalente de tradução "*ceremony*" tenha sido inserido, o editor recupera todos os *synsets* que possuem esse item como um de seus elementos constitutivos, os quais estão listados no Quadro 4. Dentre os 3 conceitos expressos por "*ceremony*", o método de DLS sugeriu o *synset {ceremony, ceremonial, ceremonial occasion, observance}*, ("*a formal event performed on a special occasion"*) ("um evento formal realizado em uma ocasião especial"). Analisando-se os demais conceitos/*synsets,* observa-se que os 3 *synsets* codificam conceitos muito similares, cuja distinção é bastante questionável. Com base na RE5, o anotador confirmou a sugestão e confirmou o *synset* 1 para a anotação do nome "cerimônia".

[23] Disponível em: http://michaelis.uol.com.br/
[24] Disponível em: http://dictionary.cambridge.org/
[25] Disponível em: http://translate.google.com.br/
[26] Disponível em: http://www.linguee.com.br/

Quadro 4 – Conceitos subjacentes a “ceremony” e seus respectivos synsets.

| | ***Synset*** | **Glosa/Frase-exemplo (Tradução da glosa)** |
|---|---|---|
| 1 | {*ceremony, ceremonial, ceremonial occasion, observanc*e} | a formal event performed on a special occasion. “a ceremony commemorating Pearl Harbor”<br>(“um evento formal realizado em uma ocasião especial. “uma cerimônia comemorativa de Pearl Harbor””) |
| 2 | {*ceremony*} | any activity that is performed in an especially solemn elaborate or formal way. “the ceremony of smelling the cork and tasting the wine”; “he makes a ceremony of addressing his golf ball”; “he disposed of it without ceremony<br>(“qualquer atividade que seja executada de uma maneira especialmente elaborada ou formal. “a cerimônia de cheirar a rolha e provar o vinho”; “ele faz uma cerimônia de abordar sua bola de golfe”; “ele se desfez sem cerimônia”) |
| 3 | {*ceremony*} | the proper or conventional behavior on some solemn occasion. “an inaugural ceremony”<br>(“o comportamento adequado ou convencional em alguma ocasião solene. Ex. 1: “uma cerimônia inaugural””) |

*Fonte*: adaptado de Fellbaum (1998)

A RE6 estabelece que, diante da inexistência de um *synset* que represente o conceito específico subjacente a uma palavra, o *synset* hiperônimo (ou seja, mais genérico) seja selecionado. Na C21, por exemplo, ocorre o nome “desnuclearização”, que no contexto descrito em (6), deve significar “interdição ou diminuição da utilização de armamento nuclear”. O equivalente de tradução em inglês “*denuclearization*” não está armazenado na WN.Pr, provavelmente por se tratar de um conceito relativamente recente e bastante específico. Assim, com base na RE6, selecionou-se o conceito mais genérico expresso pelo *synset* {*disarming, disarmament*} (“*act of reducing or depriving of arms*”) (“ato de reduzir ou privar de armas”).

(6) “[...] O regime comunista diz ter interesse na **desnuclearização** da península Coreana e se comprometeu no sábado (21) a não realizar mais testes atômicos [...]”

Por fim, a RE7 estabelece que apenas os nomes nucleares em expressões metafóricas sejam anotados com o *synset* correspondente ao conceito da expressão toda. Esse é o caso, por exemplo, da expressão “feixe de lenha” que ocorre na C5 do *corpus* (7), a qual engloba notícias sobre a “aprovação, na Câmara dos Deputados, do texto-base da reforma do Código

Florestal". Nesse texto, o nome "feixe", apesar de etiquetado em isolado, é núcleo da expressão "feixe de lenha". No caso, interpretou-se que "feixe de lenha" foi empregado em sentido metafórico, referindo-se ao "texto final da reforma florestal". Seguindo-se a RE7, apenas "feixe" foi anotado. Para tanto, esse nome foi traduzido para *text* e, por meio dele, selecionou-se o *synset {text, textual matter}* (*"the words of something written*") ("as palavras de algo escrito").

(7) "[...] Como relator, não aguento mais amarrar e desamarrar esse **feixe de lenha** e carregá-lo por mais tempo [...]"

Na Tabela 9, apresentam-se os dados quantitativos da anotação das 10 novas coleções do *corpus* CM2News (2.0), isto é: (i) quantidade de palavras por texto-fonte de cada coleção, (ii) quantidade de palavras por coleção e (iii) quantidade de nomes anotados por coleção.

Tabela 9 – Dados quantitativos da anotação da extensão do *corpus* CM2News (2.0)

| ***Cluster*** | **Domínio** | **Assunto** | **Qt. palavra/doc** | **Qt. palavra/*cluster*** | **Qt. nomes anotados/*cluster*** |
|---|---|---|---|---|---|
| C21 | Poder | Encontro de líderes das Coreias | 386 | 770 | 202 |
| | | | 384 | | |
| C22 | Ciência | Reprodução de camundongos | 578 | 1.240 | 255 |
| | | | 662 | | |
| C23 | Entreten. | Kanye West na política | 328 | 782 | 143 |
| | | | 454 | | |
| C24 | Entreten. | Bebê de Hilary Duff | 182 | 285 | 68 |
| | | | 103 | | |
| C25 | Entreten. | Acusações a Stallone | 150 | 280 | 51 |
| | | | 130 | | |
| C26 | Meio ambiente | Oleoduto EUA-Canadá | 428 | 973 | 163 |
| | | | 545 | | |
| C27 | Meio ambiente | Ataque de leoa em zoológico | 220 | 419 | 87 |
| | | | 199 | | |
| C28 | Meio ambiente | Baleia morta na Indonésia | 287 | 646 | 134 |
| | | | 359 | | |
| C29 | Saúde | EUA poliomielite | 390 | 782 | 250 |
| | | | 392 | | |
| C30 | Saúde | Camisinha autolubrificante | 522 | 1.056 | 240 |

Fonte: elaborada pela autora.

# 4 O MÉTODO PROPOSTO E A SUA APLICAÇÃO NO *CORPUS*

## 4.1 A descrição do CFULHiper

Para investigar os aspectos ainda não explorados na SAMM segundo a revisão da literatura, propôs-se o método denominado CFULHiper, cujo algoritmo, adaptado de Tosta (2014), está descrito no Quadro 5.

Quadro 5 – Algoritmo do método proposto.

| **Método CFULHiper** | |
|---|---|
| Análise | 1. Analisar cada um dos textos da coleção em nível léxico-conceitual, ou seja, anotar os nomes comuns com os conceitos/*synsets* da WN.Pr |
| Transformação | 2. Calcular a taxa de compressão<br>3. Pontuar as sentenças em função da frequência dos *synsets*/conceitos na coleção, privilegiando os conceitos superordenados<br>4. Ranquear as sentenças em função da pontuação dos conceitos<br>5. Selecionar a 1ª sentença do ranque que seja advinda do texto em português<br>6. Caso a taxa de compressão não tenha sido atingida:<br>6.a. Selecionar a próxima sentença em português do ranque<br>6.b. Verificar a redundância da sentença em questão com a já selecionada via sobreposição de conceitos<br>6.c. Eleger a sentença somente se não for redundante<br>7. Repetir o passo 6 até que a taxa de compressão seja atingida |
| Síntese | 8. Justapor as sentenças na ordem em que foram selecionadas<br>9. Ordenar as sentenças pela posição de ocorrência nos textos-fonte |

Fonte: elaborado pela autora.

Pela descrição da fase de análise no algoritmo, vê-se que o CFULHiper, assim como o CFUL, também se baseia nos conceitos nominais de uma coleção para gerar o seu respectivo extrato multilíngue.

As principais diferenças do CFULHiper frente ao CFUL estão na fase de transformação. O CFULHiper caracteriza-se especificamente por selecionar conteúdo (i) privilegiando conceitos (nominais) genéricos e (ii) evitando a redundância pela sobreposição conceitual.

A hipótese para (i) é a de que conceitos superordenados (hiperônimos) podem auxiliar na identificação de conteúdo genérico na coleção, relevante para compor extratos informativos/genéricos. A hipótese para (ii) é a de que a similaridade entre sentenças é mais adequadamente calculada quando se inclui a ocorrência de expressões distintas de um mesmo conceito (sinônimos), o que não ocorre com a sobreposição lexical.

Para privilegiar os conceitos superordenados, utilizou-se o trabalho de Zacarias (2016) como motivação e, a partir dele, o método CFULHiper aplica a frequência acumulada como critério para a pontuação diferenciada dos referidos. A escolha por esse tipo de frequência se deu em decorrência do seu uso eficaz não só na SA, mas também em diferentes aplicações de PLN, como evidenciado por Zacarias (2016). Assim, a frequência acumulada de um conceito hiperônimo *x* é calculada pela soma da sua frequência simples e da frequência simples de todos os seus conceitos subordinados (ou hipônimos) na coleção (caso *x* o tenham) (ZACARIAS, 2016). Assim, o topo do ranque é ocupado pelas sentenças compostas pelos conceitos nominais mais frequentes e genéricos da coleção.

Para o tratamento da redundância, o CFULHiper identifica a similaridade entre sentenças com base na sobreposição de conceitos, que é determinada pela medida *concept overlap*, apresentada em (4). Assim, o valor dessa medida para um par de sentenças (S1, S2) é obtido pela soma dos conceitos em comum entre S1 e S2, dividida pela soma do total de conceitos no par de sentenças, gerando um valor entre 0 e 1, sendo que, quanto mais próximo de 1 mais redundante é o par de sentenças analisado.

$$(4) \qquad \textit{Concept Overlap (S1, S2)} = \frac{\#CommonConcepts}{\#Concepts(S1) + \#Concepts(S2)}$$

Por fim, salienta-se que, assim como o CFUL, o CFULHiper também seleciona apenas as sentenças em português do ranque para a construção do extrato multilíngue em português, pois, segundo Tosta (2014), um extrato composto por sentenças originais na língua-alvo reflete as informações centrais da coleção porque os conceitos do texto em língua estrangeira (no caso, o inglês) também são considerados para pontuar os conceitos e, consequentemente, as sentenças.

## 4.2 A aplicação do CFULHiper ao CM2News (2.0)

Antes da aplicação efetiva do CFULHiper, os textos-fonte anotados de cada coleção do *corpus* CM2News (2.0) foram submetidos a um pré-processamento semiautomático. A partir

do pré-processamento dos arquivos XML (cf. Figura 12, p. 51) gerados pelo editor MulSen, aplicou-se o CFULHiper às 30 coleções de forma automática por meio de uma rotina computacional especificamente criada para a tarefa[27]. Tal rotina realiza dois processos específicos de acordo com o algoritmo do método: (i) pontuação e ranqueamento das sentenças com base na frequência (simples e acumulada) dos conceitos da coleção, e (ii) seleção de conteúdo com tratamento da redundância baseada na medida *concept overlap*.

### *4.2.1 O pré-processamento automático do corpus*

Para a aplicação automática do CFULHiper, cada coleção do CM2News (2.0) foi submetida à: (i) identificação das relações conceituais hierárquicas, necessária ao cálculo da frequência acumulada, e (ii) cálculo da frequência simples e acumulada dos conceitos, o que permitiu construir de fato o ranque das sentenças.

#### *a) A identificação das relações hierárquicas*

Para cada conceito nominal *c* anotado nos textos-fonte de uma coleção, a identificação automática das relações hierárquicas na coleção teve início com a extração dos hiperônimos que ocupam as 3 posições imediatamente superiores a *c* na hierarquia da WN.Pr. A herança somente dos hiperônimos que ocupam os 3 níveis acima de *c* na WN.Pr foi feita com base na observação empírica de que os conceitos que ocupam as posições mais superiores na hieraquia da WN.Pr (chamados *top-concepts*) não ocorrem nos textos-fonte do *corpus*.

Dado um conceito *c*, sabe-se, com base em Fellbaum (1998), que a hierarquia a que *c* pertence tem em média 12 níveis na WN.Pr. O conceito ou *synset {victim}*, por exemplo, que ocorre somente no texto em inglês da coleção C1[28], pertence a uma hierarquia composta por 11 níveis. No sentido *top-down* (isto é, do genérico para o específico), *{victim}* ocupa o nível 9, como é ilustrado de forma simplificada[29] na Figura 15, com os dados obtidos da WN.Pr. A análise empírica (manual) do *corpus* (realizada em 3 coleções de teste) revelou que, dado um conceito *c* do texto, somente conceitos hiperônimos que ocupam os 3 níveis imediatamente acima de *c* na hierarquia tendem a ocorrer nos textos-fonte. Com isso, vê-se

---

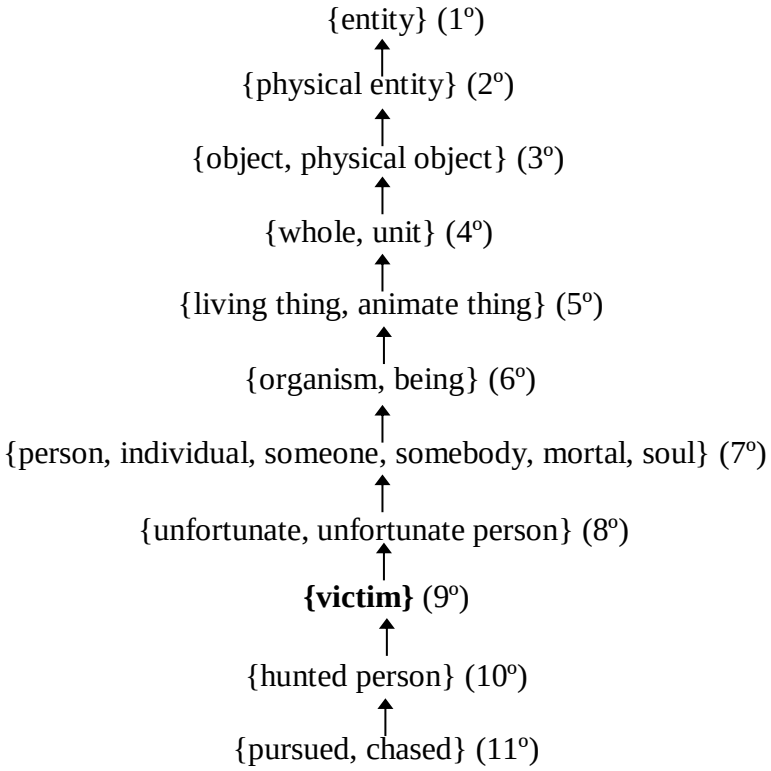
[27] A implementação do método CFULHiper foi feita por meio de um trabalho colaborativo com o cientista da computação **Roney Lira de Sales Santos**, doutorando do ICMC-USP e pesquisador do Núcleo Interinstitucional de Linguística Computacional (NILC).

[28] Os textos-fonte e o sumário de referência da coleção C1 estão disponíveis na íntegra no Anexo 1.

[29] Diz-se "simplificada" porque a Figura 15 ilustra apenas a hierarquia mais profunda tendo como ponto de partida o conceito *{victim}*. Em outras palavras, *{victim}* possui outros hipônimos imediatos (isto é, co-hipônimos de *{hunted person}*) que possuem conceitos subordinados (hipônimos), como é o caso de *{murderee}*. Assim, o conceito *{murderee}* pertence a uma hierarquia com apenas 10 níveis.

os chamados *top*-conceitos, os quais ocupam o topo da hierarquia (como *{entity}*, *{physical entity}* e *{object, physical object}*) não costumam ocorrer. Uma possível explicação para isso pode ser o fato de tais conceitos serem extremamente genéricos.

Figura 15 – Ilustração nos níveis conceituais na hierarquia nominal da WN.Pr.




Fonte: elaborada pela autora.

A partir da extração dos 3 hiperônimos imediatos de todos os conceitos anotados em uma coleção, fez-se a identificação das relações hierárquicas, que constitui em verificar se os hiperônimos herdados haviam sido anotados nos textos-fonte da coleção. Dado o conceito *{victim}*, por exemplo, herdaram-se os hiperônimos *{unfortunate, unfortunate person}*, *{person, individual, someone, somebody, mortal, soul}* e *{organism, being}*. Na sequência, esses 3 conceitos foram comparados a todos os *synsets* que ocorreram (ou foram anotados) na coleção. No caso, identificou-se a ocorrência do conceito *{person, individual, someone, somebody, mortal, soul}* na coleção. Assim, estabeleceu-se que os conceitos *{victim}* e *{person, individual, someone, somebody, mortal, soul}*, ambos com ocorrências na coleção C1, estão em relação hieráquica, na qual *{person, individual, someone, somebody, mortal, soul}* é hiperônimo de *{victim}*. E esse processo foi feito para todos os conceitos anotados.

Assim, dado o conceito *{person, individual, someone, somebody, mortal, soul}*, por exemplo, herdaram-se os conceitos *{organism, being}*, *{living thing, animate thing}* e

*{whole, unit}* por serem os hiperônimos que ocupam os 3 níveis imediatamente acima do conceito em questão e, na sequência, procedeu-se à comparação destes com todos os *synsets* anotados em C1. Dessa forma, todos os conceitos hierarquicamente relacionados puderam ser conectados. Ressalta-se que a maioria dos conceitos nominais que ocorrem no *corpus*, no entanto, não estabelece relação do tipo hieráquica (superordenado – subordinado) com outros da coleção; no entanto, outros tipos de relação não foram verificados.

*b) O cálculo das frequências*

A frequência simples dos conceitos das 20 coleções da versão 1.0 do CM2News (2.0) já constavam do *corpus*. Neste trabalho, procedeu-se ao cálculo automático da frequência simples e da acumulada para as 10 novas coleções do CM2News (2.0).

De acordo com Tosta (2014), a frequência simples de ocorrência de um conceito (ou *synset*) em uma coleção *C* equivale ao número de vezes que o *synset* em questão foi anotado na coleção. Em (8), tem-se trechos de duas sentenças advindas da coleção C1, que engloba notícias sobre "um ataque ocorrido em Londres em 2011". Os conceitos em (8) estão representados pelos seus números identificadores[30] entre os sinais de colchetes angulares e a frequência entre parênteses. Especificamente, observa-se em (8) que: (i) a frequência simples do *synset* <8209687>, indexado às palavras "*Police*" e "polícia" em C1, é 19; (ii) a frequência simples do *synset* <7942152>, indexado à "*people*" em C1, é 6; (iii) a frequência simples do *synset* <7308889>, indexado à "caso" em C1, é 3, e (iv) a frequência simples do *synset* <220522>, indexado "assassinato" em C1, é 1.

(8) a. **Police**< 8209687>(19) have said 922 **people**<7942152>(6) have been arrested [...].
b.A **polícia**<8209687>(19) está tratando o **caso**<7308889>(3) como **assassinato**<220522>(1) [....].

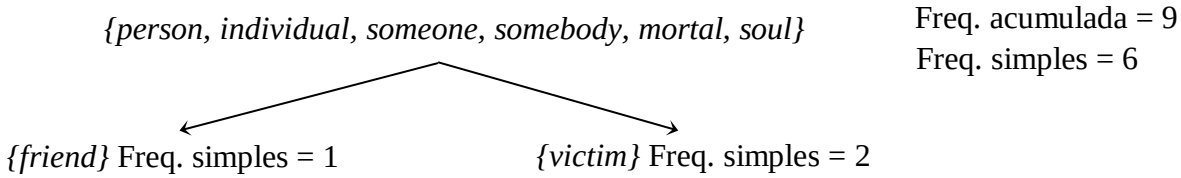
Uma vez que a frequência simples de todos os conceitos/*synset* do *corpus* foi calculada, passou-se para o cálculo da frequência acumulada. Para todo conceito *c* de uma coleção *C* que tenha sido identificado como hiperônimo (*c_hiper*), o cômputo automático da frequência acumulada consistiu em somar a frequência simples de *c_hiper* à frequência simples de todos os seus conceitos hipônimos.

Para ilustrar, considera-se o conceito/*synset {person, individual, someone, somebody, mortal, soul}*<7846>, cuja frequência simples em C1 é 6. Tal conceito, segundo a etapa de

[30] Por uma questão de brevidade de representação do conceito no exemplo (8), explicita-se o *synset* por meio de seu número identificador na WN.Pr (ID) e não pelas unidades lexicais que o constituem.

identificação manual das relações hierárquicas (cf. 4.2.1a, p. 61), não só é hiperônimo de *{victim}*, cuja frequência simples é 2, mas também de *{friend}*, cuja frequência simples é 1. Assim, para o conceito *c_hiper {person, individual, someone, somebody, mortal, soul}* em C1, a frequência acumulada é 9 (6+2+1), como ilustrado na Figura 16. Com isso, dá-se um peso diferenciado para um conceito que, em uma relação de hierarquia em uma coleção de textos-fonte, é o superordenado (ou mais genérico).

Figura 16 – Ilustração da frequência acumulada de um conceito hiperônimo.




Fonte: elaborada pela autora.

A partir da identificação das frequências, cada ocorrência de um conceito *c* em sua respectiva coleção foi associada à sua frequência final, seja ela simples ou acumulada. Tal associação foi feita em um arquivo *txt.* Em (9), tem-se duas sentenças provenientes dos textos de C1. Destaca-se, por exemplo, que toda ocorrência do conceito *{person, individual, someone, somebody, mortal, soul}*, codificado pelo ID <7846>, foi associada à frequência acumulada 9. Da mesma forma, toda ocorrência do conceito/*synset {police, police force, constabulary, law}*, codificado pelo ID <8209687>, foi associada à frequência simples 19. Tais ocorrências estão sublinhadas no exemplo em (9).

(9)

a. Na **cidade**<8524735>(5) de Nottingham não foram registrados **incidentes**<13978033>(4) relevantes, depois que a **polícia**<8209687>(19) aplicou uma **política**<5901508>(1) de **tolerância**<1071090>(1) zero sobre qualquer **pessoa**<7846>(9) que tentasse causar **desordem**<13972797>(8).

b. O **governo**<8050678>(3) também dará à **polícia**<8209687>(19) poderes para exigir que as **pessoas**<7846>(9) retirem **proteções**<4014297>(1) do **rosto**<5600637>(1) (...)

### *4.2.2 A pontuação e o ranqueamento automáticos das sentenças*

Após o cálculo das frequências simples e acumulada para cada conceito de uma coleção *C*, como descrito na seção anterior, procedeu-se à pontuação de cada uma das sentenças-fonte da coleção *C*, o que foi feito pela soma da frequência (simples ou acumulada) de seus conceitos constitutivos.

A sentença em (9a) de C1, por exemplo, é composta por sete conceitos distintos. O conceito codificado pelo ID <7846> é expresso pelo *synset {person, individual, someone, somebody, mortal, soul}*, cuja frequência acumulada em C1 é nove. Os demais conceitos/*synsets* não são hiperônimos na coleção, possuindo, assim, as frequências simples indicadas entre parênteses. Ao somar a frequência de todos os conceitos/*synsets*, tem-se que a pontuação final da sentença em questão é **47** (10).

(10)

a. Na **cidade**<8524735>(5) de Nottingham não foram registrados **incidentes**<13978033>(4) relevantes, depois que a **polícia**<8209687>(19) aplicou uma **política**<5901508>(1) de **tolerância**<1071090>(1) zero sobre qualquer **pessoa**<7846>(9) que tentasse causar **desordem**<13972797>(8). = **47**

Uma vez que todas as sentenças-fonte de uma coleção tenham sido pontuadas como em (9), estas são organizadas em um ranque, em cujo topo devem estar as sentenças constituídas pelos conceitos mais frequentes e genéricos da coleção. Por essa razão, as sentenças do topo são consideradas, segundo o método CFULHiper, as mais relevantes para compor o extrato Na Tabela 10, ilustra-se o ranque de C1, que possui um total de 54 sentenças. Nele, a sentença em (9a), por exemplo, ocupa a 14ª posição[31].

Tabela 10 – Ranque da coleção C1 segundo o método CFULHiper.

| **Posição no ranque** | **Sentença** | **Pontuação** |
|---|---|---|
| S1 | A Hugo Boss store and a bureau de change on Sloane Square were attacked between Monday night and the early hours of Tuesday before the looters targeted shops in Pimlico Road, police said. | 72 |

[31] Por uma questão de comparação, tem-se, no Apêndice 2, o ranque das sentenças-fonte de C1 gerado com base no método CFUL de Tosta (2014). Para a construção desse ranque, as sentenças foram pontuadas exclusivamente, como mencionado, por meio da soma da frequência simples de seus conceitos/*synsets* na coleção. Assim, a sentença em (9a) ocupa a 15ª posição com pontuação especificada em 41.

| | | |
|---|---|---|
| S2 | A maior presença policial e a chuva intensa que caiu na noite desta quarta-feira em algumas partes do país parecem ter evitado um quinto dia de vandalismo em algumas áreas afetadas pelos distúrbios nas últimas noites. | 70 |
| S3 | On Wednesday night the Met made a number of arrests in connection with the attack on the Sony DADC warehouse in Enfield, looting in central London and an arson attack on a furniture shop in Croydon. | 67 |
| S4 | O governo também dará à polícia poderes para exigir que as pessoas retirem proteções do rosto e vai compensar as pessoas cujas casas ou empresas foram depredadas na onda de violência em Londres e outras cidades britânicas esta semana, disse ele. | 65 |
| S5 | The Met has also issued CCTV images of a man suspected of being involved in an attack on a 68 - year-old man in Spring Bridge Road in Ealing, west London, on Monday night. | 54 |
| S6 | Wednesday was a relatively calm night, with the exception of an incident in Eltham, south-east London, where officers were pelted with missiles by a group of people. | 53 |
| S7 | The figures include two boys of 17 and a man of 18 arrested over an arson attack which destroyed a Sony warehouse in Enfield, north London, on Monday. | 51 |
| S8 | Em Leicestershire, os agentes elogiaram o comportamento cidadão e indicaram que não houve casos de desordem nesse condado, graças a uma forte operação policial que produziu 19 detenções, entre elas as de dois adolescentes de 15 anos. | 49 |
| S9 | Police believe Trevor Ellis, of Brixton Hill, and his friends were involved in an altercation with another group of nine people, resulting in a chase involving three cars. | 49 |
| S10 | A polícia de Nottinghamshire indicou hoje que não houve denúncias sobre agrupamentos significativos de jovens e informou que realizou apenas quatro detenções, frente às 86 registradas um dia antes. | 48 |
| S11 | He said: There's ways to get involved and volunteer to put things back to communities through local authorities, through the police service, there's lots of things we can ask you to do which will make our city even safer. | 47 |
| S12 | Two of three teenagers arrested in connection with the fire - one 17 - year-old and a man of 18 - remain in police custody. | 47 |
| S13 | Police have said 922 people have been arrested over violence, disorder and looting in London, with 401 charged. | 47 |
| S14 | Na cidade de Nottingham não foram registrados incidentes relevantes, depois que a polícia aplicou uma política de tolerância zero sobre qualquer pessoa que tentasse causar desordem. | 47 |
| S15 | The number of police officers across the capital was increased from 6,000 to 16,000 on Tuesday after the violence escalated on Monday. | 46 |
| S16 | In the incidents in central London, police arrested two boys aged 17 from Notting Hill and Belgravia on suspicion of burglary. | 42 |
| S17 | Two other boys aged 17 were arrested over looting in Sloane Square and Pimlico, also on Monday night. | 41 |
| S18 | A 26 - year-old man who died after being found with bullet wounds in a car at Duppas Hill Road, Croydon, south London, on Monday night, was shot in the head, it has been revealed. | 39 |
| S19 | É responsabilidade do governo assegurar que qualquer contingência futura seja avaliada, incluindo se há tarefas que o Exército pode assumir que possam liberar mais policiais para a linha de frente, disse Cameron ao parlamento, em sessão emergencial para discutir a violência. | 38 |
| S20 | A noite passada transcorreu de forma pacífica, sem mais focos de desordem, em West Midlands, e os agentes dessa região se centraram em manter uma presença de alta visibilidade, o que ajudou a evitar mais distúrbios. | 37 |
| S21 | Também se manteve a calma em Gloucester, onde a Polícia de Gloucestershire deteve seis pessoas por diferentes incidentes de ordem pública. | 37 |

| | | |
|---|---|---|
| S22 | A polícia está tratando o caso como assassinato, pelo que estão investigando um homem de 32 anos. | 35 |
| S23 | A police spokesman said: This altercation culminated in a vehicle pursuit involving three vehicles which commenced in Scarbrook Road, Croydon, passing along the A 232 flyover into Duppas Hill Road where the victim was shot. | 34 |
| S24 | Juizados municipais em várias cidades inglesas como Londres, Manchester e Solihull, em West Midlands, permaneceram em funcionamento durante a última noite para agilizar os vários casos diante da avalanche de detenções. | 32 |
| S25 | O Reino Unido vai considerar convocar o Exército em distúrbios futuros para liberar policiais para lidar com baderneiros, afirmou o primeiro-ministro, David Cameron, nesta quinta-feira. | 32 |
| S26 | He said: It's a huge drain on both the physical, emotional and practical resources of London's police service. | 31 |
| S27 | Four magistrates'courts in London have been working through the night since Tuesday to try to process the people charged in connection with the riots. | 30 |
| S28 | The increased number of officers will be out in London for a third night before the deployment is reviewed, Deputy Assistant Commissioner Stephen Kavanagh said. | 29 |
| S29 | Another man, 21, arrested over the attack, has been bailed until September. | 29 |
| S30 | The disorder on Monday night began in Hackney and spread to Croydon, Clapham, Camden, Lewisham, Peckham, Newham, East Ham, Enfield, Woolwich, Ealing and Colliers Wood. | 28 |
| S31 | Mais de 1.000 agentes foram destacados para vigiar as ruas, e realizaram na noite de ontem 48 detenções, segundo seus últimos dados. | 28 |
| S32 | Up to 16,000 officers were on duty across the capital on Wednesday night. | 28 |
| S33 | Police raids started on Thursday morning as 100 warrants were issued. | 27 |
| S34 | Na região de West Midlands, onde está Birmingham, os agentes praticaram até o momento mais de 300 detenções, e em Manchester e no subúrbio de Salford foram detidas outras 100 pessoas. | 27 |
| S35 | The defendants include an 11 - year-old boy from Romford, who pleaded guilty to stealing from a Debenhams store, and Alex Bailey, a 31 - year-old learning mentor at a Stockwell primary school, who admitted burglary at an electrical goods store in Croydon. | 26 |
| S36 | Em relação ao debate de emergência, já havia expectativas que o primeiro-ministro, David Cameron, anunciasse novas medidas para enfrentar os recentes distúrbios, assim como detalhes da ajuda econômica que será oferecida aos que perderam suas casas ou negócios durante os ataques. | 24 |
| S37 | The Sony warehouse, which stored CDs, DVDs, Blu-ray discs and games, was gutted in the blaze which was tackled by 40 firefighters on Monday. | 24 |
| S38 | The 15 - year-old boy and a man aged 25 are being held on suspicion of arson with intent to endanger life. | 23 |
| S39 | O custo de 200 milhões de libras (R $ 523,7 milhões) será amparado pelo governo, que também irá assumir os custos de zelar pelas pessoas que perderam suas casas. | 23 |
| S40 | A cidade de Birmingham manteve durante a madrugada uma vigília pelos três homens asiáticos mortos após serem atropelados por um veículo quando tentavam proteger o lugar onde moravam. | 21 |
| S41 | In the early hours of this morning we started knocking on doors to arrest people. | 21 |
| S42 | Two further arrests were made over the fire in The House of Reeves furniture store in Croydon. | 20 |
| S43 | O Parlamento britânico realiza nesta quinta-feira uma sessão extraordiná¡ria sobre a grave onda de violência que já provocou mais de 1.000 detenções desde sábado, a maioria em Londres. | 19 |
| S44 | The Met said the group had been dispersed by 22: 00 BST. | 16 |

| | | |
|---|---|---|
| S45 | Officers rounded up about 150 men. | 15 |
| S46 | If you want to protect communities, come and join us, we've got plenty of space for special constables and volunteers but otherwise join local authorities - but don't become a gang. | 12 |
| S47 | We have got more than 100 warrants which we will be working our way through over the coming hours and days. | 11 |
| S48 | Acting Commissioner Tim Godwin appealed to people not to resort to vigilantism. | 9 |
| S49 | The other boy of 17 was released on bail. | 8 |
| S50 | Some local residents were out on the streets claiming to be defending the area from rioters. | 7 |
| S51 | The victim, whose next of kin are being traced, remains in hospital in a critical condition. | 5 |
| S52 | Det Ch Insp John McFarlane asked the suspect to do the decent thing and give yourself up. | 4 |
| S53 | Mr Ellis was shot during the chase. | 3 |
| S54 | They remain in custody. | 2 |

Fonte*:* elaborada pela autora

Na próxima seção, descrevem-se as etapas de seleção de conteúdo e de tratamento da redundância para a efetiva construção dos extratos.

### *4.2.3 A seleção de conteúdo e tratamento da redundância*

A seleção de conteúdo é uma etapa feita a partir dos ranques sentenciais. Esse processo consiste em selecionar as sentenças mais bem ranqueadas até que a taxa de compressão do extrato seja atingida.

Segundo o método CFULHiper, a seleção é feita de acordo com os passos: (i) selecionar a primeira sentença em português do ranque; (ii) caso a taxa de compressão não tenha sido atingida, selecionar a próxima sentença em português do ranque; (iii) verificar, via *concept overlap*, a redundância da 2ª sentença com a 1ª já selecionada; (iv) eleger a 2ª sentença somente se não for redundante; (v) repetir a seleção de sentença e a verificação da redundância até que a taxa de compressão seja atingida.

No caso, a taxa de compressão empregada foi de 70%, o que indica que os extratos englobam 30% do número de palavras do maior texto da coleção-fonte.

Para o tratamento da redundância, propôs-se identificar a similaridade entre as sentenças com base na sobreposição de conceitos, o que foi feito pela aplicação da medida *concept overlap*, cujo cálculo está ilustrado em (4) (cf. p. 60). Assim, o valor dessa medida para um par de sentenças (S1, S2) é obtido pela soma dos conceitos em comum entre S1 e S2, dividida pela soma do total de conceitos no par de sentenças.

Para a verificação da redundância, estudou-se manualmente os dados do *corpus* com o objetivo de identificar um limiar (do inglês, *threshold*) empírico para a redundância. Dessa

forma, uma sentença candidata é efetivamente selecionada a compor o extrato se sua similaridade com a(s) já selecionada(s) para o extrato for inferior ao *threshold*. Caso contrário, a sentença candidata é descartada e a próxima sentença do ranque é considerada, passando pela mesma verificação da redundância.

O *threshold* adotado para calcular a similaridade entre as sentenças por meio do *concept overlap* no método CFULHiper foi de **0,36**. Tal valor representa a média de sobreposição de conceitos entre sentenças de uma coleção consideradas redundantes entre si. Para calcular o referido limiar, as sentenças da coleção C1 foram organizadas em pares em função do grau de similaridade entre elas, resultando em 3 conjuntos: (i) não-redundantes, (ii) pouco redundantes e (iii) totalmente redundantes. Na sequência, calculou-se a *word overlap* para todos os pares de sentenças dos 3 conjuntos. Desse cálculo, observou-se que a sobreposição de conceitos entre as sentenças totalmente redundantes foi de 0,36 em média. Assim, uma sentença (do ranque) candidata só é seleciona para compor o extrato se o valor da *concept overlap* entre ela e as já selecionadas estiver abaixo do *threshold.*

Para exemplificar o processo de seleção de conteúdo, considera-se o ranque da coleção C1 (Tabela 10). De acordo com o método CFULHiper, esse processo teve início com a seleção da primeira sentença em português do ranque (S2) (2ª posição), que é composta por 36 palavras. Considerando a taxa de compressão de 70%, o extrato em questão deve conter 237 palavras. Assim, selecionou-se a próxima sentença em português do ranque (S4) (4ª posição), que contabiliza 41 palavras. Calculando a similaridade entre S4 e S2, verificou-se que as sentenças não são redundantes, posto que o valor obtido para a *concept overlap* foi 0. Por conseguinte, S4 foi incluída no extrato, o qual possui neste estágio um total de 77 palavras. Como a taxa de compressão ainda não foi atingida, a próxima sentença em português do ranque S8 (8ª posição), que contabiliza 37 palavras, é selecionada como candidata. Os valores de *concept ovelap* obtidos para S8-S2 e S8-S4 foram abaixo do *threshold* (isto é, 0 e 0,05, respectivamente), o que permitiu selecionar S8 para extrato, contabilizando um total de 114 palavras. Como o tamanho de 237 palavras ainda não foi atingido, selecionou-se a próxima sentença em português do ranque, S10 (10ª posição), com 29 palavras. A similaridade entre S10 e as demais já selecionadas (S2, S4 e S8) também é inferior ao *threshold* (no caso, 0, 0,2, 0,13, respectivamente**)** e, portanto, S10 também foi efetivamente selecionada, resultando em um extrato parcial de 143 palavras. O mesmo procedimento foi feito para mais três sentenças em português do ranque, isto é, S14 (14ª posição e 26 palavras), S19 (19ª posição e 41 palavras) e S20 (20ª posição e 36 palavras), as quais também se revelaram não redundantes frente às demais que já compunham o extrato

(S2, S4, S8 e S10). Dessa forma, o conjunto de sentenças S2, S4, S8, S10, S14, S19 e S20 contabiliza 246 palavras para o extrato (isto é, 9 a mais que o desejado). A sentença S20 foi a última a ser selecionada por causa do critério de parada adotado neste trabalho, que foi o da "extensão parcial que mais se aproxima do tamanho desejado". Caso não se inserisse S20, o extrato teria 210 palavras, ou seja, 27 a menos que o desejado.

No Quadro 6, tem-se o conjunto final das sentenças selecionadas pelo CFULHiper.

Quadro 6 – Sentenças selecionadas de C1 pelo CFULHiper (70% de compressão)

| **Posição no ranque** | **Sentenças selecionadas para o extrato** | **Posição no texto-fonte** |
|---|---|---|
| S2 | A maior presença policial e a chuva intensa que caiu na noite desta quarta-feira em algumas partes do país parecem ter evitado um quinto dia de vandalismo em algumas áreas afetadas pelos distúrbios nas últimas noites. | 7 |
| S4 | O governo também dará à polícia poderes para exigir que as pessoas retirem proteções do rosto e vai compensar as pessoas cujas casas ou empresas foram depredadas na onda de violência em Londres e outras cidades britânicas esta semana, disse ele. | 2 |
| S8 | Em Leicestershire, os agentes elogiaram o comportamento cidadão e indicaram que não houve casos de desordem nesse condado, graças a uma forte operação policial que produziu 19 detenções, entre elas as de dois adolescentes de 15 anos. | 15 |
| S10 | A polícia de Nottinghamshire indicou hoje que não houve denúncias sobre agrupamentos significativos de jovens e informou que realizou apenas quatro detenções, frente às 86 registradas um dia antes. | 14 |
| S14 | Na cidade de Nottingham não foram registrados incidentes relevantes, depois que a polícia aplicou uma política de tolerância zero sobre qualquer pessoa que tentasse causar desordem. | 13 |
| S19 | É responsabilidade do governo assegurar que qualquer contingência futura seja avaliada, incluindo se há tarefas que o Exército pode assumir que possam liberar mais policiais para a linha de frente, disse Cameron ao parlamento, em sessão emergencial para discutir a violência. | 3 |
| S20 | A noite passada transcorreu de forma pacífica, sem mais focos de desordem, em West Midlands, e os agentes dessa região se centraram em manter uma presença de alta visibilidade, o que ajudou a evitar mais distúrbios. | 11 |

Fonte: elaborado pela autora.

Por fim, gerou-se o extrato por meio da ordenação (manual) das sentenças efetivamente selecionadas. Na busca por coerência e coesão, optou-se por justapor as sentenças segundo a posição em que ocorrem no seu respectivo textos-fonte. De acordo com o Quadro 6, as sentenças selecionadas do ranque foram assim ordenadas segundo sua posição ou ocorrência

no texto em português de C1: S4 (2ª posição no texto-fonte) > S19 (3ª posição) > S2 (7ª posição)> S20 (11ª posição)> S14 (13ª posição)> S10 (14ª posição) > S8 (15ª posição). No Quadro 7, apresenta-se o extrato multilíngue produzido com base no método CFULHiper para a C1 com base em 70% de compressão.

Quadro 7 – Extrato de C1 gerado pelo CFULHiper (70% de compressão).

| |
|---|
| O governo também dará à polícia poderes para exigir que as pessoas retirem proteções do rosto e vai compensar as pessoas cujas casas ou empresas foram depredadas na onda de violência em Londres e outras cidades britânicas esta semana, disse ele.<br>É responsabilidade do governo assegurar que qualquer contingência futura seja avaliada, incluindo se há tarefas que o Exército pode assumir que possam liberar mais policiais para a linha de frente, disse Cameron ao parlamento, em sessão emergencial para discutir a violência.<br>A maior presença policial e a chuva intensa que caiu na noite desta quarta-feira em algumas partes do país parecem ter evitado um quinto dia de vandalismo em algumas áreas afetadas pelos distúrbios nas últimas noites.<br>A noite passada transcorreu de forma pacífica, sem mais focos de desordem, em West Midlands, e os agentes dessa região se centraram em manter uma presença de alta visibilidade, o que ajudou a evitar mais distúrbios.<br>Na cidade de Nottingham não foram registrados incidentes relevantes, depois que a polícia aplicou uma "política de tolerância zero" sobre qualquer pessoa que tentasse causar desordem.<br>A polícia de Nottinghamshire indicou hoje que não houve denúncias sobre agrupamentos significativos de jovens e informou que realizou apenas quatro detenções, frente às 86 registradas um dia antes.<br>Em Leicestershire, os agentes elogiaram o comportamento cidadão e indicaram que não houve casos de desordem nesse condado, graças a uma "forte operação policial" que produziu 19 detenções, entre elas as de dois adolescentes de 15 anos. |

Fonte: elaborado pela autora.

Uma vez que, para cada uma das 30 coleções do CM2News (2.0), um extrato multilíngue tenha sido gerado conforme descrito nas duas seções imediatamente anteriores, realizou-se o procedimento de avaliação do método CFULHiper, apresentado a seguir, na Seção 5.

# 5 Avaliação do método proposto

Neste trabalho, optou-se por avaliar o método CFULHiper segundo a abordagem intrínseca, dado o interesse particular em se verificar o impacto causado pelas estratégias do CFULHiper para (i) pontuar as sentenças e (ii) tratar a redundância na qualidade linguística e informatividade dos extratos.

## 5.1 A avaliação da qualidade linguística

A avaliação da qualidade linguística foi feita com base nos parâmetros ou critérios da DUC'05 (DANG, 2005): (i) gramaticalidade, (ii) não-redundância, (iii) clareza referencial, (iv) foco temático, e (v) estrutura/coerência. Cada um desses parâmetros foi avaliado com base em uma escala de 1 a 5 (1=péssimo, 2=ruim, 3=regular, 4=bom e 5=excelente).

A avaliação dos 30 extratos gerados pelo método CFULHiper foi realizada manualmente por 6 juízes com experiência em Linguística. Para a realização do procedimento, desenvolveu-se um formulário *online* por meio do qual cada juíz leu, avaliou e submeteu o julgamento individualmente. No caso, o formulário continha: (i) descrição da tarefa, (ii) exemplificação dos critérios de qualidade e regras de pontuação, e (iii) extratos a serem avaliados. Organizaram-se os 30 extratos em 3 grupos de 10, sendo que cada grupo foi avaliado por 4 juízes diferentes. No Quadro 8, apresenta-se a distribuição dos extratos por juíz.

Quadro 8 – Distribuição dos extratos por juízes.

| ***Cluster*/extrato** | **Avaliador** | | | | ***Cluster*/extrato** | **Avaliador** | | | | ***Cluster*/extrato** | **Avaliador** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 1 | 2 | 3 | 4 | C11 | 3 | 4 | 5 | 6 | C21 | 1 | 2 | 5 | 6 |
| C2 | 1 | 2 | 3 | 4 | C12 | 3 | 4 | 5 | 6 | C22 | 1 | 2 | 5 | 6 |
| C3 | 1 | 2 | 3 | 4 | C13 | 3 | 4 | 5 | 6 | C23 | 1 | 2 | 5 | 6 |
| C4 | 1 | 2 | 3 | 4 | C14 | 3 | 4 | 5 | 6 | C24 | 1 | 2 | 5 | 6 |
| C5 | 1 | 2 | 3 | 4 | C15 | 3 | 4 | 5 | 6 | C25 | 1 | 2 | 5 | 6 |
| C6 | 1 | 2 | 3 | 4 | C16 | 3 | 4 | 5 | 6 | C26 | 1 | 2 | 5 | 6 |
| C7 | 1 | 2 | 3 | 4 | C17 | 3 | 4 | 5 | 6 | C27 | 1 | 2 | 5 | 6 |
| C8 | 1 | 2 | 3 | 4 | C18 | 3 | 4 | 5 | 6 | C28 | 1 | 2 | 5 | 6 |
| C9 | 1 | 2 | 3 | 4 | C19 | 3 | 4 | 5 | 6 | C29 | 1 | 2 | 5 | 6 |
| C10 | 1 | 2 | 3 | 4 | C20 | 3 | 4 | 5 | 6 | C30 | 1 | 2 | 5 | 6 |

Fonte: elaborado pela autora.

Na sequência, apresenta-se o resultado obtido na avaliação de cada um dos 5 critérios da DUC´05. Especificamente, para cada critério, explicita-se a média das pontuações dadas pelos 4 juízes para o total de 30 extratos. Uma análise mais detalhada do desempenho do CFULHiper quanto a cada um dos 5 critérios da DUC é feita quando da comparação dos resultados obtidos pelos métodos CFUL (TOSTA, 2014) e o método baseline de Tosta, Di-Felippo e Pardo (2013) (cf. Tabela 16).

Na Tabela, 11, observa-se que, quanto à gramaticalidade, os extratos gerados pelo método CFULHiper foram avaliados 78 vezes (71,8%) como “excelente”, 21 vezes (20,7%) como “bom” e apenas 8 vezes (7,4%) como “regular”, o que resultou em uma média de 4,6. Ressalta-se, neste caso, que as pontuações 1=péssimo e 2=ruim não foram atribuídas a nenhum dos extratos gerados pelo método.

Tabela 11 – Avaliação da gramaticalidade dos extratos do CFULHiper.

| **Gramaticalidade** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Péssimo (1)** | | **Ruim (2)** | | **Regular (3)** | | **Bom (4)** | | **Excelente (5)** | | **Média** |
| 0 | 0% | 0 | 0% | 8 | 7,4% | 21 | 20,7% | 78 | 71,8% | **4,6** |

Fonte: elaborado pela autora.

A Tabela 12 exibe os resultados obtidos pelo CFULHiper quanto à não-redundância. Em particular, destaca-se que, em metade dos casos, esse critério foi avaliado como “excelente” (50,4%) e que em apenas 2 avaliações a não-redundância recebeu pontuação 2=ruim. Os demais casos se distribuem em “bom” (25,2%) e “regular” (22,4). No geral, tal critério teve média de 4,2.

Tabela 12 – Avaliação da não-redundância nos extratos do CFULHiper.

| **Não-redundância** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Péssimo (1)** | | **Ruim (2)** | | **Regular (3)** | | **Bom (4)** | | **Excelente (5)** | | **Média** |
| 0 | 0 | 2 | 1,8% | 24% | 22,4% | 27 | 25,2% | 54 | 50,4% | **4,2** |

Fonte: elaborada pela autora.

Com base na Tabela 13, vê-se que, em apenas 3 avaliações, a pontuação 2=ruim foi atribuída aos extratos quanto ao critério da não-redundância. As demais avaliações se distribuem entre as pontuações “regular” (28%), “bom” 29,9% e sobretudo “excelente” (40,5%), o que resultou na média 4.

Tabela 13 – Avaliação da clareza referencial nos extratos do CFULHiper

| **Clareza referencial** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Péssimo (1)** | | **Ruim (2)** | | **Regular (3)** | | **Bom (4)** | | **Excelente (5)** | | **Média** |
| 0 | 0 | 3 | 2,8% | 30 | 28% | 32 | 29,9% | 42 | 40,5% | **4** |

Fonte: elaborado pela autora.

Na Tabela 14, exibem-se os resultados quanto ao critério foco (temático). Observa-se, com base nos dados da referida Tabela, que os extratos obtiveram média 4,5. Em particular, destaca-se que as pontuações 1=péssimo e 2=ruim não foram atribuídas a nenhum dos 30 extratos e que 3=regular foi atribuída a apenas 10 vezes.

Tabela 14 – Avaliação do foco temático nos extratos do CFULHiper

| **Foco** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Péssimo (1)** | | **Ruim (2)** | | **Regular (3)** | | **Bom (4)** | | **Excelente (5)** | | **Média** |
| 0 | 0 | 0 | 0 | 10 | 9,3% | 32 | 29,9% | 65 | 60,7% | **4,5** |

Fonte: elaborado pela autora.

Quanto à estrutura e coerência dos extratos, observa-se na Tabela 15 que a média obtida foi 4, destacando que a pontuação 1=péssimo não foi atribuída aos extratos e que 2=ruim foi atribuída em apenas 1 ocasião.

Tabela 15 – Avaliação da estrutura e coerência nos extratos do método CFULHiper

| **Estrutura-coerência** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Péssimo (1)** | | **Ruim (2)** | | **Regular (3)** | | **Bom (4)** | | **Excelente (5)** | | **Média** |
| 0 | 0 | 1 | 0,9% | 27 | 25,2% | 47 | 43,9% | 31 | 28,9% | **4** |

Fonte: elaborado pela autora.

Para fim de comparação, sistematizam-se na Tabela 16 os resultados obtidos na avaliação da qualidade linguística dos métodos CFULHiper, CFUL (TOSTA, 2014) e o *baseline* de Tosta, Di-Felippo e Pardo (2013). O *baseline*, no caso, aplica a estratégia *early-translation* e pauta a seleção de conteúdo na localização das sentenças nos textos-fonte.

Tabela 16 – Comparação entre os métodos CFULHiper, CFUL e *baseline.*

| **Critérios** | **CFULHiper** | **CFUL** | ***Baseline*** |
|---|---|---|---|
| Gramaticalidade | **4,6** | 4,3 | 3 |
| Não-redundância | 4,2 | **4,3** | 3 |
| Clareza referencial | **4** | 3,7 | 3,2 |
| Foco temático | **4,5** | 4,1 | 4 |
| Estrutura e coerência | **4** | 3,4 | 2,8 |

Fonte: elaborado pela autora.

Frente ao método profundo CFUL, em especial, observa-se que o CFULHiper apresenta resultados consideravelmente superiores quanto à "clareza referencial" e "estrutura/coerência". Especificamente, as médias do CFUL foram 3,7 e 3,4 e as do CFULHiper foram 4,2 e 4, respectivamente. Tais resultados podem indicar que, embora o ranque das sentenças seja construído com base na pontuação dos conceitos em ambos os textos-fonte (língua-alvo e estrangeira), o CFULHiper, ao selecionar apenas as sentenças em português, justapondo-as na ordem em que ocorrem em seu texto-fonte (no caso, o em português), consegue evitar os problemas relativos à coesão, como a quebra de correferência. O CFUL, no entanto, ordena as sentenças em função da posição das mesmas no ranque.

Quanto à "gramaticalidade", a média 4,6 do CFULHiper, em comparação à média 4,3 do CFUL, representa relativa superioridade do método aqui proposto. Tal resultado parece fortalecer a hipótese de que extratos multilíngues compostos exclusivamente por sentenças selecionadas do texto-fonte na língua-alvo não são prejudicados pela tradução automática de eventuais sentenças selecionadas do texto-fonte em língua estrangeira.

Sobre a "não-redundância", destaca-se que, com base nos dados da Tabela 16, os extratos gerados por ambos os métodos profundos, CFUL e CFULHiper, comumente não apresentam muita informação redundante, posto que as médias de pontuação obtidas foram 4,3 e 4,2, respectivamente. Isso pode ser resultado do fato de que a seleção das sentenças em ambos acaba por ser monodocumento, embora o ranque das sentenças seja construído com base nos conceitos de todos os textos-fonte. Sendo a seleção monodocumento, as sentenças extraídas de um único documento já tendem a não apresentar muita redundância.

O CFULHiper, em particular, obteve pontuação média de 4,2, a qual é ligeiramente menor do que a média obtida pelo CFUL (TOSTA, 2014) (4,3) e significativamente maior do que a obtida pelo *baseline* (TOSTA, DI-FELIPPO e PARDO, 2013) (3,0). Isso permite dizer que o tratamento da redundância baseado na *concept overlap* apresenta bons resultados,

mas não interfere consideralmente na qualidade linguística dos extratos. Uma hipótese para isso é o fato de que os textos-fonte (notícias) de coleções multidocumento apresentam baixa variação lexical[32] e polissemia (DE LUCA, 2019), e, por isso, aplicar uma medida de sobreposição lexical (*word overlap*) ou de sobreposição conceitual (*concept overlap*) não gera diferença na identificação da similaridade entre as sentenças.

## 5.2 Avaliação da informatividade

Para avaliar a informatividade dos extratos gerados pelo método CFULHiper, utilizou-se o pacote de medidas ROUGE (LIN; HOVY, 2003). Especificamente, utilizaram-se a ROUGE-1 e a ROUGE-2, que são as medidas mais aplicadas na sumarização automática. A ROUGE-1 calcula a sobreposição de unigramas entre o extrato automático e o sumário de referência e a ROUGE-2 calcula a sobreposição de bigramas entre o extrato automático e o de referência.

Os resultados são fornecidos em termos de precisão, cobertura e medida-f. Para a aplicação da ROUGE, os extratos gerados pelo CFULHiper para cada uma das 30 coleções do CM2News (2.0) foram automaticamente comparados aos seus respectivos sumários de referência.

Na Tabela 17, tem-se os resultados da ROUGE-1 e 2 por *cluster* e as médias.

Tabela 17 – Avaliaçaõ da informatividade dos extratos via ROUGE.

| ***Clusters*** | **ROUGE-1** | | | **ROUGE-2** | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-f** | **Cobertura** | **Precisão** | **Medida-f** |
| **C1** | 0,39512 | 0,37156,0 | 0,38298 | 0,12438 | 0,11848 | 0,12136 |
| **C2** | 0,2987 | 0,29487 | 0,29677 | 0,12162 | 0,11842 | 0,12 |
| **C3** | 0,42718 | 0,45128 | 0,4389 | 0,12308 | 0,12834 | 0,12565 |
| **C4** | 0,42759 | 0,47328 | 0,44928 | 0,23188 | 0,25397 | 0,24242 |
| **C5** | 0,42614 | 0,41899 | 0,42254 | 0,27108 | 0,26163 | 0,26627 |
| **C6** | 0,28125 | 0,5 | 0,36 | 0,1828 | 0,32075 | 0,23288 |
| **C7** | 0,42553 | 0,42857 | 0,42705 | 0,14599 | 0,14815 | 0,14706 |
| **C8** | 0,37692 | 0,35 | 0,36296 | 0,11111 | 0,1037 | 0,10728 |
| **C9** | 0,33529 | 0,32948 | 0,33236 | 0,10976 | 0,10909 | 0,10942 |
| **C10** | 0,31193 | ,0,40719 | 0,35325 | 0,11848 | 0,15528 | 0,13441 |
| **C11** | 0,54676 | 0,51701 | 0,53147 | 0,31579 | 0,29577 | 0,30545 |
| **C12** | 0,5 | 0,51773 | 0,50871 | 0,31618 | 0,31387 | 0,31502 |
| **C13** | 0,30814 | 0,29282 | 0,30028 | 0,04192 | 0,0407 | 0,0413 |

[32] Ao realizar a anotação conceitual de algumas coleções multidocumento e monolíngue do CSTNews (CARDOSO *et al.* (2011), De Luca (2019) evidencia que as notícias que as compõem apresentam baixa ocorrência de sinonímia e polissemia. E isso parece que pode ser estendido para o *corpus* CMSNews (2.0), mesmo que as coleções sejam multilíngues.

| | | | | | | |
|---|---|---|---|---|---|---|
| **C14** | 0,44944 | 0,50633 | 0,47619 | 0,29412 | 0,32895 | 0,31056 |
| **C15** | 0,46305 | 0,45631 | 0,45966 | 0,18367 | 0,1809 | 0,18228 |
| **C16** | 0,3252 | 0,34188 | 0,33333 | 0,05738 | 0,0614 | 0,05932 |
| **C17** | 0,54658 | 0,5641 | 0,55521 | 0,39873 | 0,42282 | 0,41042 |
| **C18** | 0,34807 | 0,33511 | 0,34146 | 0,05 | 0,04972 | 0,04986 |
| **C19** | 0,46154 | 0,39706 | 0,42688 | 0,2 | 0,17424 | 0,18623 |
| **C20** | 0,44203 | 0,46923 | 0,45522 | 0,25 | 0,26613 | 0,25781 |
| **C21** | 0,45631 | 0,4405 | 0,4424 | 0,2968 | 0,2974 | 0,2896 |
| **C22** | 0,48235 | 0,25076 | 0,32998 | 0,17073 | 0,08889 | 0,11691 |
| **C23** | 0,26496 | 0,2844 | 0,27434 | 0,05172 | 0,0566 | 0,05405 |
| **C24** | 0,2931 | 0,40476 | 0,34 | 0,07273 | 0,1 | 0,08421 |
| **C25** | 0,15217 | 0,17073 | 0,16092 | 0,06977 | 0,075 | 0,07229 |
| **C26** | 0,27083 | ,0,29104 | 0,28058, | 0,07692 | 0,08462 | 0,08059 |
| **C27** | 0,30303 | 0,30769 | 0,30534 | 0,15873 | 0,16393 | 0,16129 |
| **C28** | 0,33333 | 0,38462 | 0,35714 | 0,10112 | 0,12 | 0,10976 |
| **C29** | 0,36893 | 0,38776 | 0,37811 | 0,12745 | 0,1383 | 0,13265 |
| **C30** | 0,29577 | 0,30216 | 0,29893 | 0,04255 | 0,04478 | 0,04364 |
| **MÉDIA** | **0,399544** | **0,371561** | **0,396609** | **0,177687** | **0,180753** | **0,175172** |

Fonte: elaborada pela autora.

As médias de precisão, cobertura e medida-f obtidas pelo CFULHiper foram comparadas somente às médias do método CFUL (TOSTA, 2014), pois a informatividade do método de Tosta, Di-Felippo e Pardo (2013), tido como *baseline*, não foi mensurada pelos autores. Na Tabela 18, tem-se a comparação da informatividade entre os referidos métodos profundos e extrativos de SAMM.

Tabela 18 – Comparação da informatividade entre os métodos CFULHiper e CFUL.

| **Método** | **ROUGE-1** | | | **ROUGE-2** | | |
|---|---|---|---|---|---|---|
| | **Cobertura** | **Precisão** | **Medida-f** | **Cobertura** | **Precisão** | **Medida-f** |
| **CFULHiper** | 0,399544 | 0,371561 | **0,396609** | 0,177687 | 0,180753 | **0,175172** |
| **CFUL** | 0,373642 | 0,369939 | 0,371175 | 0,174795 | 0,175514 | 0,174845 |

Fonte: elaborada pela autora.

Para analisar a relevância do desempenho do método CFULHiper quanto à informatividade, toma-se como parâmetro outra tarefa de sumarização, a SAM (monolíngue), que já foi mais amplamente investigada do que a SAMM. Na geração automática de extratos multidocumento (e monolíngue) em português, por exemplo, o RC-4 (CARDOSO, 2014, CARDOSO; PARDO, 2015, 2016), considerado o estado-da-arte da abordagem profunda, obteve 0.4419 de medida-f para ROUGE-1 e 0.2586 de medida-f para ROUGE-2, e o

RSumm (RIBALDO; CARDOSO; PARDO, 2013, RIBALDO; CARDOSO; PARDO, 2016), considerado o estado-da-arte da abordagem híbrida, obteve 0.4190 de medida-f para ROUGE-1 e 0.3434 de medida-f para ROUGE-2. Ao comparar os valores médios de medida-f obtida pelos dois métodos de SAMM, vê-se que estes são um pouco inferiores, já que CFULHiper obteve 0,396609 de medida-f para ROUGE-1 e 0,175172 de medida-f para ROUGE-2 e CFUL apresenta 0,371175 de medida-f para ROUGE-1 e 0,174845 de medida-f para ROUGE-2. Tais valores, entretando, não são tão distantes do estado-da-arte em SAM.

Sobre a comparação entre os dois métodos profundos e extrativos de SAMM, CFUL e CFULHiper, observa-se que o CFULHiper apresenta médias ligeiramente superiores para ambas as medidas ROUGE em comparação às médias do método CFUL. Isso indica que o CFULHiper captura de forma mais eficiente a informação central da coleção do que o método CFUL. Tendo em vista que o método proposto neste trabalho privilegia os conceitos superordenados das relações hierárquicas identificadas nos textos-fonte, seu desempenho ligeiramente superior parece indicar que a seleção de informação mais genérica contribui para melhorar a informatividade dos extratos.

# 6 CONSIDERAÇÕES FINAIS

Nesta pesquisa, investigaram-se (i) a aplicação, na seleção de conteúdo, de conhecimento léxico-conceitual privilegiando conceitos superordenados em relação de hiponímia nas coleções de textos-fonte e (ii) o tratamento da redundância baseada em conceitos, especificamente por meio da medida denominada *concept overlap.*

A investigação em (i) foi realizada com base na hipótese de que os conceitos superordenados (ou hiperônimos) podem auxiliar na identificação de conteúdo relevante para a construção de extratos na SAMM. Já a investigação em (ii) pautou-se na hipótese de que a similaridade entre sentenças é mais adequadamente calculada pela sobreposição de conceitos, posto que a medida tradicional de *word overlap* penaliza sentenças similares nas quais ocorrem itens lexicais sinônimos.

Para investigar os aspectos descritos em (i) e (ii), propôs-se o método extrativo e profundo de SAMM denominado CFULHiper.

De acordo com a informatividade superior dos extratos gerados pelo método em questão frente ao estado-da-arte em SAMM para o português como língua-alvo (metódo CFUL), pode-se dizer que a hipótese de (i) foi confirmada. Diz-se isso porque o ranqueamento das sentenças resultante da pontuação diferenciada dos conceitos hiperônimos parece contribuir para a seleção da informação central da coleção. Sendo assim, as informações mais genéricas, capturadas pelos conceitos hiperônimos, parecem contribuir para melhorar a informatividade dos extratos.

Quanto ao uso da *concept overlap* para o tratamento da redundância, pode-se dizer que a hipótese para tanto não se confirmou, posto que os resultados obtidos pelos métodos CFUL e CFULHiper na avalição manual do critério da "não-redundância" são muito similares.

Uma possível razão para o emprego da *concept overlap* não ter de fato um impacto no tratamento da redundância é o fato de que as sentenças selecionadas para o extrato são provenientes de um único texto-fonte (em português) e, por isso, elas naturalmente já não apresentam muita redundância entre si.

Ademais, salienta-se que os textos-fonte (notícias) de coleções multidocumento tendem a apresentar baixa sinonímia e polissemia (DE LUCA, 2019) e, por isso, aplicar uma medida de sobreposição lexical (*word overlap*) ou de sobreposição conceitual (*concept overlap*) não gera diferença na identificação da similaridade entre as sentenças.

O presente trabalho trouxe, como contribuição, a extensão e balanceamento do *corpus* CM2News, que está disponível para futuras investigações no cenário de SAMM do português com aplicação de conhecimento léxico-conceitual.

Como trabalho futuro, destaca-se o interesse em aplicar o método CFULHiper ao *corpus* CM3News (NASCIMENTO, 2019). Uma vez composto por 20 coleções trilíngues, cujos textos-fonte (1 notícia em português, 1 em inglês e 1 em alemão) já estão anotados em nível léxico-conceitual como o CM2News (2.0), a aplicação do método ao referido *corpus* permitiria observar se o aumento do número de línguas afeta o desempenho do CFULHiper. Além disso, destaca-se a variação da taxa de compressão como uma possibilidade de trabalho futuro no contexto de pontuação diferenciada a conceitos superordenados, considerando a hipótese de que sumários menores podem ser mais genéricos, já que, dessa forma, selecionam-se, entre as primeiras sentenças do ranque, apenas as mais bem pontuadas entre as melhores sentenças do ranque.

# REFERÊNCIAS BIBLIOGRÁFICAS

AGIRRE, E.; EDMONDS, P.G. Word sense disambiguation: Algorithms and applications. Springer Science-Business Media, 2006.

AKABANE, A.T.; PARDO, T.A.S.; RINO, L.H.M. Explorando medidas de redes complexas para sumarização multidocumento. In: STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2, 2011, Cuiabá/MT. Proceedings…Cuiabá, 2011, p.1-3.

BOURDIN, F. A.; HUET, S.; TORRES-MORENO, JUAN-MANUEL. Graph-based approach to cross-language multi-document summarization. In: CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 12, 2011, Tokyo, Japan. Proceedings… Tokyo: CICLing, 2011, p.113-8.

BENTIVOGLI, L.; PIANTA, E. Beyond lexical units: enriching wordNets with *phrasets*. In: EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 3, 2003, Budapest, Hungary. Proceedings... Budapest, 2003. p. 67-70.

BERBER SARDINHA, T. Linguística de corpus. Barueri: Manole, 2004.

CAMARGO, R. T. Investigação de estratégias de sumarização humana multidocumento. 2013. 133 p. Dissertação (Mestrado, Programa de Pós-Graduação em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2013.

CAMARGO, R. T.; DI FELIPPO, A.; PARDO, T. A. S. On strategies of Human Multi-Document Summarization. In: The 10th Brazilian Symposium in Information and Human Language Technology (STIL), 2015. Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology, 2015. v. 01. p. 01-08.

CAMARGO, Y., V.; DI-FELIPPO, A. Enriquecendo o *corpus* CM2News: construção e anotação de coleções bilíngues de notícias. In: WORKSHOP ON PORTUGUESE DESCRIPTION (JDP-STIL), 6, 2019, Salvador/BA. Proceedings... Salvador/BA, pp. 239-243.

CARDOSO, P.C.F.; MAZIERO, E.G.; CASTRO JORGE, M.L.R.; SENO, E.M.R.; DI-FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. A CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. Proceedings… Cuiabá: UFMT, 2011. p. 88-105.

CARDOSO, P. C. F. Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo. São Carlos, SP 2014, Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC/USP), São Carlos, SP, 182 p., 2014.

CASTRO JORGE, M. L. R. Modelagem gerativa para sumarização automática multidocumento. São Carlos, SP, 2015. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC/USP). São Carlos, SP, 151 p., 2015.

CASTRO JORGE, M. L. R.; AGOSTINI, V.; PARDO, T. A. S. Multi-document summarization using complex and rich features. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, 8, 2011, Natal. Proceedings… Natal, RN, 2011, p. 1-12.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. Experiments with CST-based Multidocument summarization. In: ACL WORKSHOP TEXTGRAPHS-5: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010, Uppsala, Sweden. Proceedings…Uppsala, 2010. p. 74-82.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. A Generative approach for multidocument summarization using the Noisy Channel model. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. Proceedings… Cuiabá, MT, 2011, p. 75-87.

COWIE, J., MAHESH, K., NIRENBURG, S., AND ZAJAZ, R. MINDS - multilingual interactive document summarization. In: AAAI SPRING SYMPOSIUM ON INTELLIGENT TEXT SUMMARIZATION, 1998, Menlo Park, CA. Proceedings…, Menlo, 1998. p. 131-2.

DANG, H. T. Overview of DUC 2005. In: Proceedings of the Document Understanding Conference, 2005.

DE LUCA, R. C. Aplicação de conhecimento léxico-conceitual na Sumarização Automática Multidocumento. 2019. 112 p. Dissertação (Mestrado, Programa de Pós-Graduação em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2019.

DI-FELIPPO, A. CM2News: Towards a Corpus for Multilingual Multi-document Summarization. In: CORPORA AND TOOLS FOR PROCESSING CORPORA WORKSHOP (CTPC), 2016, Tomar, Portugal. Proceedings... 2016. v. 01. p. 01-08.

DI-FELIPPO, A.; TOSTA, F.E.S.; PARDO, T.A.S. Applying Lexical-Conceptual Knowledge for Multilingual Multi-document Summarization. In: INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE – PROPOR, 12, 2016, Tomar/Portugal. Proceedings... Tomar: LNAI 9727, p. 38-49, 2016.

ENDRES-NIGGEMEYER, B. Summarization Information. Berlin: Springer, 1998. 374 p.

EVANS, D.K.; KLAVANS, J.L.; MCKEOWN, K.R. Columbia NewsBlaster: multilingual news summarization on the web. In: NORTH AMERICAN CHAPTER OF THE ACL: HUMAN LANGUAGE TECHNOLOGIES, 2004, Boston. Proceedings… Boston, 2004, p. 1-4.

EVANS, D. K.; KLAVANS, J.L.; MCKEOWN, K.R. Similarity-based multilingual multi-document summarization. Technical Report CUCS-014-05, Columbia University, 2005. 8p.

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

GANTZ, J.; REINSEL, D. The Digitalization of the Word: From Edge to Core. International Data Corporation, 2018.

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.
HALTEREN, H.V.; TEUFEL, S. Examining the consensus between human summaries: initial experiments with factoid analysis. In HLT/NAACL-2003 Workshop on Automatic Summarization, 2003.

HATZIVASSILOGLOU, J. L.; KLAVANS J.L.; HOLCOMBE, M. Simfinder: a flexible clustering tool for summarization. In: NAACL AUTOMATIC SUMMARIZATION WORKSHOP, 2001. Pittsburgh, PA, USA. **Proceedings**… Pittsburgh, 2001, p.9.

JURAFSKY, D; MARTIN, J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2001. 1024p.

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

LESK, M. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." *In: Proceedings os the 5th annual international conference on Systems documentation*, pp.24-26. ACM, 1986.

LI, L.; WANG, D; SHEN, C; LI, T. Ontology-enriched multi-document summarization in disaster management. In: ACM SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL (SIGIR), 2010, Geneva. Proceedings… Geneva, Switzerland, 2010. p. 819-820.

LIN, C-Y.; HOVY, E. H. Automatic evaluation of summaries using n-gram cooccurrence statistics. In: THE 2003 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOG, 2003, Edmonton, Canada. Proceedings… Edmonton, 2003.p.71-8

LITVAK, M.; LAST, M.; FRIEDMAN, M. A New approach to improving multilingual summarization using a genetic algorithm. In: THE ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48, 2010, Stroudsburg, PA, USA. Proceedings… Stroudsburg, 2010.p. 927-936.

LIN, C. Y. ROUGE: a Package for Automatic Evaluation of Summaries. In: Workshop ON TEXT SUMMARIZATION BRANCHES OUT (WAS 2004), 8, 2004, Barcelona, Spain. Proceedings… Barcelona, 2004, p. 74-81.

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

MARCUSCHI, L. A. Produção textual, análise de gêneros e compreensão. São Paulo: Parábola Editorial, 2008.

MANI, I. Automatic summarization. Amsterdam: John Benjamins Publishing Co., 2001, 286 p.

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

MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: a theory of text organization. 1987. (Technical Report ISI/RS-87-190).

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

MCKEOWN, K.; KLAVANS, J.; HATZIVASSILOGLOU, V.; BARZILAY, R; ESKIN, E. Towards multi-document summarization by reformulation: Progress and prospects. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 16, 1999, Florida, USA. Proceedings... Florida 1999. p. 453–460.

NASCIMENTO, D. X. Explorando a avaliação de sumários automáticos multidocumento multilíngues. 2019. 84 p. Qualificação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos (2019).

NENKOVA, A.; PASSONNEAU, R. Evaluating content selection in summarization: The pyramid method. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (HLT/NAACL), 2004, Boston. Proceedings... Boston, MA, 2004, 1-8 p.

NENKOVA, A.; MCKEOWN. K. Automatic summarization. Foundations and Trends in Information Retrieval, 5(2-3):103–233, 2011.

NÓBREGA, F. A. A. Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. 2013. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2013.

ORĂSAN, C. Automatic summarization in the informational age. In: INTERNATIONAL CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING, 7, 2009, Borovets, Bulgaria. Proceedings... Borovets: 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, 6, 2008, Marrakesh. Proceedings… Marrakesh, Morocco, 2008. p.6.

PARDO, T.A.S. GistSumm – GIST SUMMarizer: extensões e novas funcionalidades. São Carlos: ICMC-USP, 2005. 8p. (Série de Relatórios do NILC. NILC-TR-05-05).

PARDO, T. A. S.; ALEIXO, P. CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (*Cross-document Structure Theory*). São Carlos: NILC-ICMC, 2008. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional).

RADEV, D. R. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: ACL SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1, 2000, Hong Kong. Proceedings…Hong Kong, 2000. p. 74-83.

RADEV, D.; ALLISON T.; BLAIR-GOLDENSOHN, S.; BLITZER, J.; CELEBI, A.; DIMITROV, S.; DRABEK, E.; HAKIM, A.; LAM, W.; LIU, D.; OTTERBACHER, J.; QI, H.; SAGGION, H.; TEUFEL S.; TOPPER, M; WINKEL, A.; ZHANG, Z. MEAD - a platform for multi-document multilingual text summarization. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4, 2004, Lisbon, Portugal. Proceedings… Lisbon, 2004.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 1996, Philadelphia, PA. Proceedings… Philadelphia, 1996. p. 133-142.

RIBALDO, R.; RINO, L.H.M.; PARDO, T.A.S. Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 10., 2012, Coimbra. Proceedings...Coimbra: Universidade de Coimbra, 2012. p. 260-271.

RIBALDO, R; CARDOSO, P. C. F; PARDO, T.A.S. Investigação de métodos de segmentação e agrupamento de subtópicos para sumarização multidocumento. Anais.. Porto Alegre: SBC, 2013.

RIBALDO, R.; CARDOSO; P.C.F; PARDO, T.A.S. Exploring the subtopic-based relationship map strategy for multi-document summarization. Revista de Informática Teórica e Aplicada: RITA, v. 23, p. 183-211, 2016.

RIBALDO, R.; CARDOSO; P.C.F; PARDO, T.A.S. Exploring the subtopic-based relationship map strategy for multi-document summarization. Revista de Informática Teórica e Aplicada: RITA, v. 23, p. 183-211, 2016.

ROARK, B., FISHER, S.: OGI OHSU baseline multilingual multi-document summarization system. In: MULTILINGUAL SUMMARIZATION EVALUATION (MSE). Michigan, USA, 2005.

SAGGION, H; LAPALME, G. Summary Generation and Evaluation in SumUM, 2000.

SALTON, G., BUCKLEY, C. Weighting approaches in automatic text retrieval. Information Processing and Management, 513–523, 1988.

SCHIFFMAN, B.; NENKOVA, A.; MCKEOWN, K. Experiments in multi-document summarization. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2, 2002, San Francisco. Proceedings… San Francisco, CA, USA, 2002, p. 52-58.

SHANNON, C. E. A Mathematical Theory of Communication. Bell System Technical Journal, 27, p. 379–423, 1948.

SHIMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: INTERNATIONAL CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING, 1994, Manchester/UK. Proceedings... Manchester.

SPARCK JONES, K. Discourse modeling for automatic summarisation. Tech. Report No. 290. University of Cambridge. UK, February, 1993.

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. Automatic summarizing: factors and directions. In: MANI, I.; MAYBURY, M. T. (Eds.). Advances in automatic text summarization. Massachusetts: MIT Press, 1999. p. 1-14.

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

TOSTA, F.E.S.: Aplicação de conhecimento léxico-conceitual na Sumarização Multidocumento Multilíngue. 2013. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos (2014).

TOSTA, F.E.S., DI-FELIPPO, A., PARDO, T.A.S.: Estudo de métodos clássicos de sumarização automática no cenário multidocumento multilíngue. In: WORKSHOP DE INICIAÇÃO CIENTÍFCA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 4, 2013, Fortaleza, Brazil. Proceedings... Fortaleza, 2013, p. 34-36.

VIEIRA, F. E., FARACO, C.A. Escrever na Universidade 1 – Fundamentos. São Paulo: Parábola Editorial, 2019.

WAN, X; LI, H.; XIAO, J. Cross-Language document summarization based on Machine translation quality prediction. In: ANNUAL MEETING OF ASSOCIATION FOR COMPUTACIONAL LINGUISTICS (ACL), 48, 2010, Uppsala, Sweeden. Proceedings… Uppsala, 2010, p. 917-926.

WHITE, M.; KORELSKY, T.; CARDIE, C.; NG, V.; PIERCE, D.; & WAGSTAFF, K. Multidocument summarization via information extraction. In Proceedings of the 1st International Conference on Human Language Technology Research, 2000.

ZACARIAS, A. C. I. Investigação De Métodos De Sumarização Automática Multidocumento baseados em hierarquias conceituais. 2016. 141 p. Dissertação (Mestrado, Programa de Pós-Graduação em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2016.

# ANEXO 1 – TEXTOS-FONTE E SUMÁRIO DE REFERÊNCIA DE C1

Texto-fonte em Português de C1

O Reino Unido vai considerar convocar o Exército em distúrbios futuros para liberar policiais para lidar com baderneiros, afirmou o primeiro-ministro, David Cameron, nesta quinta-feira.

O governo também dará à polícia poderes para exigir que as pessoas retirem proteções do rosto e vai compensar as pessoas cujas casas ou empresas foram depredadas na onda de violência em Londres e outras cidades britânicas esta semana, disse ele.

"É responsabilidade do governo assegurar que qualquer contingência futura seja avaliada, incluindo se há tarefas que o Exército pode assumir que possam liberar mais policiais para a linha de frente", disse Cameron ao parlamento, em sessão emergencial para discutir a violência.

O custo de 200 milhões de libras (R$ 523,7 milhões) será amparado pelo governo, que também irá assumir os custos de zelar pelas pessoas que perderam suas casas.

O Parlamento britânico realiza nesta quinta-feira uma sessão extraordinária sobre a grave onda de violência que já provocou mais de 1.000 detenções desde sábado, a maioria em Londres.

Em relação ao debate de emergência, já havia expectativas que o primeiro-ministro, David Cameron, anunciasse novas medidas para enfrentar os recentes distúrbios, assim como detalhes da ajuda econômica que será oferecida aos que perderam suas casas ou negócios durante os ataques.

A maior presença policial e a chuva intensa que caiu na noite desta quarta-feira em algumas partes do país parecem ter evitado um quinto dia de vandalismo em algumas áreas afetadas pelos distúrbios nas últimas noites.

Juizados municipais em várias cidades inglesas como Londres, Manchester e Solihull, em West Midlands, permaneceram em funcionamento durante a última noite para agilizar os vários casos diante da avalanche de detenções.

Na região de West Midlands, onde está Birmingham, os agentes praticaram até o momento mais de 300 detenções, e em Manchester e no subúrbio de Salford foram detidas outras 100 pessoas.

A cidade de Birmingham manteve durante a madrugada uma vigília pelos três homens asiáticos mortos após serem atropelados por um veículo quando tentavam proteger o lugar onde moravam.

A polícia está tratando o caso como assassinato, pelo que está investigando um homem de 32 anos.

A noite passada transcorreu "de forma pacífica, sem mais focos de desordem", em West Midlands, e os agentes dessa região se centraram em manter "uma presença de alta visibilidade", o que ajudou a evitar mais distúrbios.

Mais de 1.000 agentes foram destacados para vigiar as ruas, e realizaram na noite de ontem 48 detenções, segundo seus últimos dados.

Na cidade de Nottingham não foram registrados incidentes relevantes, depois que a polícia aplicou uma "política de tolerância zero" sobre qualquer pessoa que tentasse causar desordem.

A polícia de Nottinghamshire indicou hoje que não houve denúncias sobre agrupamentos significativos de jovens e informou que realizou apenas quatro detenções, frente às 86 registradas um dia antes.

Em Leicestershire, os agentes elogiaram o comportamento cidadão e indicaram que não houve casos de desordem nesse condado, graças a uma "forte operação policial" que produziu 19 detenções, entre elas as de dois adolescentes de 15 anos.

Também se manteve a calma em Gloucester, onde a Polícia de Gloucestershire deteve seis pessoas por diferentes incidentes de ordem pública.

Texto-fonte em Inglês de C1

Police have said 922 people have been arrested over violence, disorder and looting in London, with 401 charged.
The figures include two boys of 17 and a man of 18 arrested over an arson attack which destroyed a Sony warehouse in Enfield, north London, on Monday.
Two other boys aged 17 were arrested over looting in Sloane Square and Pimlico, also on Monday night.
Police raids started on Thursday morning as 100 warrants were issued.
Up to 16,000 officers were on duty across the capital on Wednesday night.
Wednesday was a relatively calm night, with the exception of an incident in Eltham, south-east London, where officers were pelted with missiles by a group of people. Officers rounded up about 150 men.
The Met said the group had been dispersed by 22:00 BST.
Some local residents were out on the streets claiming to be defending the area from rioters.
Acting Commissioner Tim Godwin appealed to people not to resort to vigilantism.
He said: "There's ways to get involved and volunteer to put things back to communities through local authorities, through the police service, there's lots of things we can ask you to do which will make our city even safer.
"If you want to protect communities, come and join us, we've got plenty of space for special constables and volunteers but otherwise join local authorities - but don't become a gang."
The number of police officers across the capital was increased from 6,000 to 16,000 on Tuesday after the violence escalated on Monday.
'Huge drain'
The increased number of officers will be out in London for a third night before the deployment is reviewed, Deputy Assistant Commissioner Stephen Kavanagh said.
He said: "It's a huge drain on both the physical, emotional and practical resources of London's police service.
"In the early hours of this morning we started knocking on doors to arrest people.
"We have got more than 100 warrants which we will be working our way through over the coming hours and days."
A 26-year-old man who died after being found with bullet wounds in a car at Duppas Hill Road, Croydon, south London, on Monday night, was shot in the head, it has been revealed.
Police believe Trevor Ellis, of Brixton Hill, and his friends were involved in an altercation with another group of nine people, resulting in a chase involving three cars. Mr Ellis was shot during the chase.
A police spokesman said: "This altercation culminated in a vehicle pursuit involving three vehicles which commenced in Scarbrook Road, Croydon, passing along the A232 flyover into Duppas Hill Road where the victim was shot."
The Met has also issued CCTV images of a man suspected of being involved in an attack on a 68-year-old man in Spring Bridge Road in Ealing, west London, on Monday night.
The victim, whose next of kin are being traced, remains in hospital in a critical condition. Det Ch Insp John McFarlane asked the suspect to "do the decent thing and give yourself up".
Teenagers arrested
On Wednesday night the Met made a number of arrests in connection with the attack on the Sony DADC warehouse in Enfield, looting in central London and an arson attack on a furniture shop in Croydon.
The Sony warehouse, which stored CDs, DVDs, Blu-ray discs and games, was gutted in the blaze which was tackled by 40 firefighters on Monday.

Two of three teenagers arrested in connection with the fire - one 17-year-old and a man of 18 - remain in police custody. The other boy of 17 was released on bail.
In the incidents in central London, police arrested two boys aged 17 from Notting Hill and Belgravia on suspicion of burglary. They remain in custody.
A Hugo Boss store and a bureau de change on Sloane Square were attacked between Monday night and the early hours of Tuesday before the looters targeted shops in Pimlico Road, police said.
Two further arrests were made over the fire in The House of Reeves furniture store in Croydon. The 15-year-old boy and a man aged 25 are being held on suspicion of arson with intent to endanger life. Another man, 21, arrested over the attack, has been bailed until September.
The disorder on Monday night began in Hackney and spread to Croydon, Clapham, Camden, Lewisham, Peckham, Newham, East Ham, Enfield, Woolwich, Ealing and Colliers Wood.
Four magistrates' courts in London have been working through the night since Tuesday to try to process the people charged in connection with the riots.
The defendants include an 11-year-old boy from Romford, who pleaded guilty to stealing from a Debenhams store, and Alex Bailey, a 31-year-old learning mentor at a Stockwell primary school, who admitted burglary at an electrical goods store in Croydon.

Sumário de referência de C1

O Parlamento Britânico realiza nesta quinta feira uma sessão extraordinária sobre a grave onda de violência que está ocorrendo na Inglaterra.
Já havia expectativas que o primeiro ministro, David Cameron, anunciasse novas medidas para agir em relação aos novos distúrbios. De acordo com ele, o governo irá considerar convocar o exército, caso haja distúrbios futuros, dará a polícia poderes para exigir que as pessoas retirem proteções de seus rostos, e vai compensar os cidadãos cujas casas ou empresas foram depredadas na onda de violência desta semana. O valor do custo que será amparado pelo governo é de 200 milhões de libras (R$ 523,7 milhões).
Os distúrbios já provocaram mais de 1000 detenções desde sábado, a maioria em Londres. Na região de West Midlands, onde está Birmigham, agentes praticaram até o momento mais de 300 detenções, e em Manchester e no subúrbio de Salford foram detidas outras 100 pessoas. Entre os detidos estão homens, na maioria com menos de 18 anos, que destruíram comércios e praticaram furtos e incêndios.
Os conflitos provocaram a morte de um homem de 26 anos, que foi encontrado baleado em um carro, em Croydon, no sul de Londres, e de três homens asiáticos, que foram atropelados enquanto tentavam proteger o lugar onde moravam. Além disso, um idoso, de 68 anos, suspeito de estar envolvido em um ataque, foi deixado em estado grave em um hospital.

# ANEXO 2 – RANQUE DE C1 GERADO PELO MÉTODO CFUL.

| Posição no ranque | Sentença | Pontuação |
|---|---|---|
| S1 | A Hugo Boss store and a bureau de change on Sloane Square were attacked between Monday night and the early hours of Tuesday before the looters targeted shops in Pimlico Road, police said. | 72 |
| S2 | A maior presença policial e a chuva intensa que caiu na noite desta quarta-feira em algumas partes do país parecem ter evitado um quinto dia de vandalismo em algumas áreas afetadas pelos distúrbios nas últimas noites. | 70 |
| S3 | On Wednesday night the Met made a number of arrests in connection with the attack on the Sony DADC warehouse in Enfield, looting in central London and an arson attack on a furniture shop in Croydon. | 67 |
| S4 | The Met has also issued CCTV images of a man suspected of being involved in an attack on a 68 - year-old man in Spring Bridge Road in Ealing, west London, on Monday night. | 54 |
| S5 | O governo também dará à polícia poderes para exigir que as pessoas retirem proteções do rosto e vai compensar as pessoas cujas casas ou empresas foram depredadas na onda de violência em Londres e outras cidades britânicas esta semana, disse ele. | 53 |
| S6 | The figures include two boys of 17 and a man of 18 arrested over an arson attack which destroyed a Sony warehouse in Enfield, north London, on Monday. | 51 |
| S7 | Two of three teenagers arrested in connection with the fire - one 17 - year-old and a man of 18 - remain in police custody. | 47 |
| S8 | Police have said 922 people have been arrested over violence, disorder and looting in London, with 401 charged. | 47 |
| S9 | He said: There's ways to get involved and volunteer to put things back to communities through local authorities, through the police service, there's lots of things we can ask you to do which will make our city even safer. | 47 |
| S10 | The number of police officers across the capital was increased from 6,000 to 16,000 on Tuesday after the violence escalated on Monday. | 46 |
| S11 | Em Leicestershire, os agentes elogiaram o comportamento cidadão e indicaram que não houve casos de desordem nesse condado, graças a uma forte operação policial que produziu 19 detenções, entre elas as de dois adolescentes de 15 anos. | 43 |
| S12 | In the incidents in central London, police arrested two boys aged 17 from Notting Hill and Belgravia on suspicion of burglary. | 42 |
| S13 | Wednesday was a relatively calm night, with the exception of an incident in Eltham, south-east London, where officers were pelted with missiles by a group of people. | 41 |
| S14 | Two other boys aged 17 were arrested over looting in Sloane Square and Pimlico, also on Monday night. | 41 |
| S15 | Na cidade de Nottingham não foram registrados incidentes relevantes, depois que a polícia aplicou uma política de tolerância zero sobre qualquer pessoa que tentasse causar desordem. | 41 |
| S16 | A 26 - year-old man who died after being found with bullet wounds in a car at Duppas Hill Road, Croydon, south London, on Monday night, was shot in the head, it has been revealed. | 39 |
| S17 | É responsabilidade do governo assegurar que qualquer contingência futura seja avaliada, incluindo se há tarefas que o Exército pode assumir que possam liberar mais policiais para a linha de frente, disse Cameron ao parlamento, em sessão emergencial para discutir a violência. | 38 |
| S18 | Police believe Trevor Ellis, of Brixton Hill, and his friends were involved in an altercation with another group of nine people, resulting in a chase involving three cars. | 37 |

| | | |
|---|---|---|
| S19 | A noite passada transcorreu de forma pacífica, sem mais focos de desordem, em West Midlands, e os agentes dessa região se centraram em manter uma presença de alta visibilidade, o que ajudou a evitar mais distúrbios. | 37 |
| S20 | A polícia de Nottinghamshire indicou hoje que não houve denúncias sobre agrupamentos significativos de jovens e informou que realizou apenas quatro detenções, frente às 86 registradas um dia antes. | 36 |
| S21 | A polícia está tratando o caso como assassinato, pelo que estão investigando um homem de 32 anos. | 35 |
| S22 | A police spokesman said: This altercation culminated in a vehicle pursuit involving three vehicles which commenced in Scarbrook Road, Croydon, passing along the A 232 flyover into Duppas Hill Road where the victim was shot. | 34 |
| S23 | Juizados municipais em várias cidades inglesas como Londres, Manchester e Solihull, em West Midlands, permaneceram em funcionamento durante a última noite para agilizar os vários casos diante da avalanche de detenções. | 32 |
| S24 | O Reino Unido vai considerar convocar o Exército em distúrbios futuros para liberar policiais para lidar com baderneiros, afirmou o primeiro-ministro, David Cameron, nesta quinta-feira. | 32 |
| S25 | Também se manteve a calma em Gloucester, onde a Polícia de Gloucestershire deteve seis pessoas por diferentes incidentes de ordem pública. | 31 |
| S26 | He said: It's a huge drain on both the physical, emotional and practical resources of London's police service. | 31 |
| S27 | Four magistrates'courts in London have been working through the night since Tuesday to try to process the people charged in connection with the riots. | 30 |
| S28 | The increased number of officers will be out in London for a third night before the deployment is reviewed, Deputy Assistant Commissioner Stephen Kavanagh said. | 29 |
| S29 | Another man, 21, arrested over the attack, has been bailed until September. | 29 |
| S30 | Up to 16,000 officers were on duty across the capital on Wednesday night. | 28 |
| S31 | The disorder on Monday night began in Hackney and spread to Croydon, Clapham, Camden, Lewisham, Peckham, Newham, East Ham, Enfield, Woolwich, Ealing and Colliers Wood. | 28 |
| S32 | Mais de 1.000 agentes foram destacados para vigiar as ruas, e realizaram na noite de ontem 48 detenções, segundo seus últimos dados. | 28 |
| S33 | Police raids started on Thursday morning as 100 warrants were issued. | 27 |
| S34 | The defendants include an 11 - year-old boy from Romford, who pleaded guilty to stealing from a Debenhams store, and Alex Bailey, a 31 - year-old learning mentor at a Stockwell primary school, who admitted burglary at an electrical goods store in Croydon. | 26 |
| S35 | The Sony warehouse, which stored CDs, DVDs, Blu-ray discs and games, was gutted in the blaze which was tackled by 40 firefighters on Monday. | 24 |
| S36 | Em relação ao debate de emergência, já havia expectativas que o primeiro-ministro, David Cameron, anunciasse novas medidas para enfrentar os recentes distúrbios, assim como detalhes da ajuda econômica que será oferecida aos que perderam suas casas ou negócios durante os ataques. | 24 |
| S37 | The 15 - year-old boy and a man aged 25 are being held on suspicion of arson with intent to endanger life. | 23 |
| S38 | Na região de West Midlands, onde está Birmingham, os agentes praticaram até o momento mais de 300 detenções, e em Manchester e no subúrbio de Salford foram detidas outras 100 pessoas. | 21 |
| S39 | In the early hours of this morning we started knocking on doors to arrest people. | 21 |
| S40 | A cidade de Birmingham manteve durante a madrugada uma vigília pelos três homens asiáticos mortos após serem atropelados por um veículo quando tentavam proteger o lugar onde moravam. | 21 |
| S41 | Two further arrests were made over the fire in The House of Reeves furniture store in Croydon. | 20 |

| | | |
|---|---|---|
| S42 | O Parlamento britânico realiza nesta quinta-feira uma sessão extraordinária sobre a grave onda de violência que já provocou mais de 1.000 detenções desde sábado, a maioria em Londres. | 19 |
| S43 | O custo de 200 milhões de libras (R $ 523,7 milhões) será amparado pelo governo, que também irá assumir os custos de zelar pelas pessoas que perderam suas casas. | 17 |
| S44 | Officers rounded up about 150 men. | 15 |
| S45 | If you want to protect communities, come and join us, we've got plenty of space for special constables and volunteers but otherwise join local authorities - but don't become a gang. | 12 |
| S46 | We have got more than 100 warrants which we will be working our way through over the coming hours and days. | 11 |
| S47 | Acting Commissioner Tim Godwin appealed to people not to resort to vigilantism. | 9 |
| S48 | The other boy of 17 was released on bail. | 8 |
| S49 | Some local residents were out on the streets claiming to be defending the area from rioters. | 7 |
| S50 | The victim, whose next of kin are being traced, remains in hospital in a critical condition. | 5 |
| S51 | The Met said the group had been dispersed by 22:00 BST. | 4 |
| S52 | Det Ch Insp John McFarlane asked the suspect to do the decent thing and give yourself up. | 4 |
| S53 | Mr Ellis was shot during the chase. | 3 |
| S54 | They remain in custody. | 2 |