Programa de Pós-Graduação em Linguística

PPGL

EXPLORANDO A AVALIAÇÃO DE SUMÁRIOS AUTOMÁTICOS MULTIDOCUMENTO MULTILÍNGUES

DARLAN XAVIER NASCIMENTO

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

EXPLORANDO A AVALIAÇÃO DE SUMÁRIOS AUTOMÁTICOS MULTIDOCUMENTO MULTILÍNGUES

Darlan Xavier Nascimento
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 – São Paulo – Brasil
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 do candidato Darlan Xavier Nascimento, realizada em 12/03/2020:

Profa. Dra. Ariani Di Felippo
UFSCar

Prof. Dr. Thiago Alexandre Salgueiro Pardo
USP

Prof. Dr. Jackson Wilke da Cruz Souza
UNIFAL

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Jackson Wilke da Cruz Souza 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

Nascimento, Darlan Xavier

Explorando a avaliação de sumários automáticos multidocumento multilíngues / Darlan Xavier Nascimento. -- 2020.
101 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador: Ariani Di Felippo
Banca examinadora: Thiago Alexandre Salgueiro Pardo, Jackson Wilke da Cruz Souza
Bibliografia

1. Sumarização automática. 2. Linguística computacional. 3. Avaliação de sumários. I. Orientador. II. Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).
DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário(a) Responsável: Ronildo Santos Prado – CRB/8 7325

Dedico esta dissertação
à minha mãe, que sempre batalhou para que
eu tivesse um futuro melhor, sabendo que a educação
é o único caminho, e ao meu pai, que sei que
está olhando por mim de algum lugar.

AGRADECIMENTOS

Quero agradecer, primeiramente, à minha mãe, Rosa Maria, e a minhas irmãs e sobrinhos, Adriele, Arieli, Filipe e Lariza, por compreenderem que, se não estamos juntos todos os dias, é porque estou em constante busca de me superar e dar meu melhor.

À minha orientadora, Ariani, pelo estímulo e por estar sempre disponível para que este trabalho se concretizasse.

Aos amigos e amigas de sempre e mais recentes, por entenderem quando não pude sair "porque preciso escrever" ou "porque tenho que mexer no *software*", rs.

Às minhas companheiras de pós-graduação, pelos momentos de desabafo, ajuda mútua e de celebração em diferentes regiões do Brasil, e aos novos colegas que fiz nessa jornada, seja para conversar, estudar, escrever *scripts*, sair para comer (e, no meu caso, principalmente beber) ou para ver a Ferroviária e o São Carlos jogarem.

Aos demais professores que tive a oportunidade de conhecer, agradeço por tudo o que pude aprender com vocês.

De modo geral, ao NILC e ao NEA, por facilitarem (e, às vezes, confundirem) minha vida de linguista em meio a tantas tarefas computacionais.

A todos os que participaram deste trabalho, produzindo e avaliando sumários, agradeço pelo tempo e pelo conhecimento que dedicaram.

À CAPES, pelo suporte financeiro durante os últimos 24 meses.

Obrigado a todos! ♥

RESUMO

A Sumarização Automática Multidocumento Multilíngue (SAMM) é uma aplicação computacional por meio da qual se produz um sumário em uma língua de interesse a partir de uma coleção de pelo menos dois textos de conteúdo equivalente e redigidos em idiomas diferentes. Verificou-se, na literatura científica, que poucas pesquisas se concentraram em métodos que geram sumários em português. Tendo como base os métodos CF e CFUL, esta dissertação apresenta o desenvolvimento de um estudo no qual se pretendeu refinar a avaliação da qualidade dos sumários produzidos, variando (i) a língua materna dos produtores dos sumários de referência, isto é, sumários escritos por humanos a partir da leitura dos textos-fonte correspondentes e que são necessários ao cálculo automático da informatividade, e (ii) a taxa de compressão (tamanho desejado do sumário). Além disso, ampliou-se o *corpus* utilizado nos estudos originais desses métodos (que continha material em português e inglês) por meio da inclusão de textos em língua alemã e produziram-se quatro extratos para cada uma das vinte coleções do *corpus*. Os resultados mostram que os sumários de referência apresentam leve interferência da língua materna de quem os redigiu, embora outros fatores possam ser considerados, como a extensão de cada texto-fonte e a compatibilidade de conteúdo. Com relação aos métodos investigados, identificou-se que os extratos com menor taxa de compressão tiveram melhor desempenho na avaliação automática da informatividade, mas pior desempenho em termos de qualidade linguística.

PALAVRAS-CHAVE: Sumarização automática; Linguística computacional; Avaliação de sumários.

ABSTRACT

Multilingual Multi-Document Automatic Summarization (MMDS) is a computational task through which a summary is produced in a target language from a collection of at least two news stories which address the same subject, one in the user's language and the other(s) in foreign language(s). The scientific literature shows that not many researches approach methods which generate summaries in Portuguese. Based on the CF and CFUL summarization methods, the present thesis describes the development of a study whose goal was to refine the summary quality evaluation, by varying (i) the native language of the producers of the reference summaries, that is, summaries written by human subjects after reading the corresponding source texts and which are necessary for the automatic calculation of informativeness, and (ii) the compression rate (desired summary size). Furthermore, this thesis outlines the enlargement of the corpus used for the investigation of these methods through the addition of texts in German (the original corpus included content in Portuguese and English) and the production of four extracts for each of the twenty clusters. The results show that the reference summaries are slightly impacted by their writer's native language, even though additional factors might be taken into account, such as the size of each source text and the content compatibility. Regarding the summarization methods, this study found that extracts with a lower compression rate performed better when it came to the automatic evaluation of informativeness and worse in the assessment of linguistic quality.

KEYWORDS: Automatic summarization; Computational linguistics; Summary evaluation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

CAPÍTULO 1 – Introdução

1.1 Contextualização

Sumarizar é uma tarefa relativamente comum na comunicação humana. Muitas vezes, não se deseja ou não é conveniente fazer uma descrição exaustiva dos acontecimentos do mundo real, o que leva as pessoas a selecionarem aquilo que se percebe como mais relevante, sem prejudicar o que se deseja transmitir. Fora do contexto meramente conversacional, há vários exemplos diretos de "resumos" dessa natureza, como a sinopse de um filme ou de uma obra literária. Porém, também é possível identificar elementos da sumarização em reportagens jornalísticas ou até mesmo em aulas, partindo do pressuposto de que elas não esgotam absolutamente o tema abordado e contêm, essencialmente, o conteúdo que cabe no momento, dadas as condições e os objetivos comunicativos.

A Sumarização Automática (SA) é uma aplicação computacional que visa à automação da referida tarefa manual. Em linhas gerais, a SA produz um sumário (coeso e coerente) a partir de um texto ou um conjunto de textos, buscando reduzir a extensão do material-fonte pela identificação e seleção das informações mais importantes desse material.

Os motivos pelos quais se poderia desejar a produção automática de sumários são diversos. De um lado, observa-se a explosão textual recente, consequência da maior presença das ferramentas digitais no dia a dia das pessoas. Rogers *et al.* (2013) chamam esse fenômeno de "infobesidade", um neologismo que associa a dificuldade do processamento da grande quantidade de informação disponível no mundo digital à obstrução de artérias pelo colesterol e à redução da capacidade (do processamento ou do corpo humano) de ter seu desempenho máximo. Do outro lado, observa-se uma necessidade que emerge do cotidiano multitarefa. A sumarização tem por objetivo reduzir a extensão de um conteúdo, agilizando a identificação do conteúdo que importa em dado contexto, o que pode ser do interesse tanto de pessoas quanto de organizações, como empresas que dispõem de amplos bancos de dados.

Tendo em vista que muitas das informações em formato textual disponíveis na *Web* circulam em diferentes línguas, reconhece-se a necessidade do desenvolvimento de aplicações de SA que lidem não apenas com o volume de informação repetida, mas

também com a multiplicidade de idiomas, permitindo o acesso às informações na língua do usuário ou em uma língua na qual seja proficiente. Dessa necessidade, métodos/sistemas de Sumarização Automática Multidocumento Multilíngue (SAMM) têm sido desenvolvidos.

As aplicações de SAMM partem obrigatoriamente de um conjunto composto pelo menos por um texto em uma língua L_x e outro texto em uma língua L_y que abordam o mesmo assunto e geram o sumário correspondente a essa coleção em uma dessas línguas-fonte (L_x ou L_y) (p.ex.: EVANS *et al.*, 2004).

Assim definida, a SAMM precisa lidar com (i) os problemas clássicos da SA, como a busca por coesão e coerência nos sumários, (ii) as questões características da multiplicidade de textos-fonte, como a ocorrência de informações contraditórias, redundantes e complementares no material-fonte, e (iii) o agravante da multiplicidade das línguas-fonte. O item (iii) é normalmente contornado pela realização de uma fase de tradução automática (TA) do material-fonte (p.ex.: EVANS; McKEOWN; KLAVANS, 2005) ou pelo emprego de métodos profundos que não necessitam da tradução integral dos textos-fonte (p.ex.: TOSTA, 2014).

Para a SAMM que tem o português como língua de interesse, o melhor método é o CFUL (*concept frequency* + *user language*), seguido pelo CF (*concept frequency*) (TOSTA, 2014; DI-FELIPPO *et al.*, 2016). Tais métodos realizam a SAMM em 4 etapas: (i) identificação dos conceitos nominais dos textos-fonte e cômputo de sua frequência na coleção, (ii) ranqueamento das sentenças pela soma da frequência de seus conceitos nominais, (iii) seleção das sentenças mais bem pontuadas e não redundantes entre si até que se atinja a taxa de compressão (tamanho desejado do sumário) e (iv) justaposição das sentenças selecionadas para compor o sumário. Neste caso, o sumário obtido também pode ser chamado de extrato, pois é composto exclusivamente por sentenças presentes nos textos-fonte, sem reescrita de seu conteúdo.

A diferença entre os métodos reside na etapa (iii). O CFUL seleciona apenas as sentenças mais bem classificadas em português, enquanto o método CF seleciona as sentenças de maior pontuação, independentemente da língua-fonte, e traduz automaticamente as eventuais sentenças em língua estrangeira para o português. Esses dois métodos superaram o *baseline* (isto é, um método simples para comparação) (TOSTA; DI-FELIPPO; PARDO, 2013) que realiza a TA dos textos-fonte em língua estrangeira para o português, ranqueia as sentenças em função da posição que ocorrem

nos textos-fonte e seleciona as primeiras sentenças para compor o sumário, tratando a redundância entre elas e substituindo as que apresentam problemas de TA por similares advindas dos textos em português.

Os métodos foram avaliados intrinsecamente (SPARCK JONES; GALLIERS, 1996; MANI, 2001), ou seja, tal verificação concentrou-se na qualidade do sistema em si, e não em sua capacidade de atender às necessidades de outras ferramentas computacionais. Nesse caso, houve uma avaliação intrínseca por meio da análise manual da qualidade linguística e da análise automática da informatividade dos sumários. Para tanto, cada um dos referidos métodos gerou um sumário (com taxa de compressão[1] de 70%) para cada uma das 20 coleções bilíngues (inglês-português) do *corpus* CM2News (TOSTA, 2014).

Quanto à qualidade, os sumários automáticos de cada coleção foram analisados em função dos seguintes parâmetros (DANG, 2005): gramaticalidade, não-redundância, clareza referencial, foco e estrutura/coerência. A informatividade foi analisada pela métrica ROUGE (LIN; HOVY, 2003), que a determina pelo cômputo do número de *n*-gramas em comum entre um sumário automático e um ou mais sumários de referência e a expressa pelas medidas de precisão, cobertura e medida-f. Assim, os sumários gerados para as coleções do *corpus* foram comparados aos seus respectivos sumários de referência em português, produzidos por falantes nativos do português de forma abstrativa (ou seja, com reescrita do material original) e com base na mesma taxa de compressão dos sumários automáticos. Como resultado, verificou-se que o CFUL, que seleciona apenas as sentenças em português mais bem ranqueadas, gera sumários mais informativos e com menos desvios gramaticais.

Com base no melhor desempenho do método CFUL, surgiram dois questionamentos: (i) A língua materna dos redatores dos sumários de referência tem alguma influência sobre a produção desses textos a ponto de eles conterem mais informação proveniente do texto-fonte da respectiva língua materna, afetando, por consequência, o desempenho dos métodos? (ii) O desempenho dos métodos é o mesmo quando da geração de sumários com diferentes taxas de compressão?

---

[1] A compressão é um valor tipicamente expresso em porcentagem que indica a taxa de redução da extensão de um texto para a produção de um sumário. Um sumário com 70% de compressão, por exemplo, contém 30% do tamanho do texto original. Tradicionalmente, no caso da sumarização multidocumento para o português, a taxa se baseia na quantidade de palavras do maior texto-fonte.

Desse modo, propôs-se refinar a avaliação dos métodos profundos de SAMM que têm o português como língua de interesse, explorando os aspectos que permitem responder a essas questões, com o intuito de contribuir para o avanço das pesquisas em SAMM.

1.2 Objetivos e hipóteses

Esta pesquisa teve como objetivo analisar sistematicamente elementos relacionados à avaliação de sumários produzidos pelos métodos CF e CFUL. Especificamente, pretendeu-se:

a) investigar se a língua materna dos produtores dos sumários multilíngues de referência influencia a produção desses textos, a ponto de os mesmos conterem mais conteúdo proveniente do texto-fonte do respectivo idioma materno, sob a hipótese de que o melhor desempenho do CFUL se deve ao fato de que os sumários manuais do *corpus* CM2News foram produzidos exclusivamente por falantes nativos do português, contendo preferencialmente informações advindas dos textos-fonte nessa língua;
b) avaliar o desempenho dos métodos quando da geração de sumários com taxas de compressão diferentes, sob a hipótese de que extratos menores gerados pelo método CF, por exemplo, podem apresentar poucos problemas de qualidade linguística, sobretudo aqueles que resultam da TA dos textos-fonte.

1.3 Metodologia

Para alcançar os objetivos e verificar a validade das hipóteses, o trabalho foi equacionado nas seguintes etapas metodológicas:

Tarefa 1: Revisão da literatura

Consistiu na leitura de textos acadêmicos que abordam conceitos básicos de SA, incluindo as estratégias de avaliação dos sistemas e, sobretudo, os principais trabalhos sobre SAMM.

Tarefa 2: Seleção dos métodos de SAMM

Nesta etapa, realizou-se a seleção/estudo dos métodos profundos de SAMM de Tosta (2014) (CFUL e CF), cuja avaliação é objeto de exploração neste trabalho.

Tarefa 3: Seleção e extensão de um *corpus* multidocumento multilíngue

A Tarefa 3 consistiu na seleção de um *corpus* adequado à pesquisa. No caso, selecionou-se o CM2News (1.0) (TOSTA, 2014), *corpus* jornalístico que possui 20 coleções bilíngues (português-inglês). A Tarefa 3 englobou a extensão do CM2News (1.0) pela (i) inclusão de um novo texto-fonte a cada coleção (em uma língua estrangeira distinta das já cobertas pelo recurso) e (ii) anotação léxico-conceitual desses novos textos segundo as diretrizes e a ferramenta de anotação de Tosta (2014), o que deu origem a um *corpus* estendido, denominado CM3News.

Tarefa 4: Geração dos extratos automáticos

A Tarefa 4 consistiu em submeter cada coleção trilíngue do CM3News aos métodos CF e CFUL para que estes gerassem extratos em português de acordo com diferentes taxas de compressão (30% e 70%), que são estipuladas em função da quantidade média de palavras dos textos-fonte da coleção. Desse modo, cada coleção do CM3News possui (i) um extrato com 30% de compressão produzido pelo método CF, (ii) um extrato com 70% de compressão produzido pelo método CF, (iii) um extrato com 30% de compressão produzido pelo método CFUL e (iv) um extrato com 70% de compressão produzido pelo método CFUL.

Tarefa 5: Produção dos sumários de referência

No âmbito dessa Tarefa, houve a produção de sumários de referência para algumas coleções do CM3News por falantes nativos de uma das línguas estrangeiras que compõem o *corpus* (além do português, considerada língua-alvo). Ao contrário dos extratos automáticos, esses sumários de referência são do tipo *abstract* informativo, pois houve reescrita do conteúdo dos textos-fonte. Além disso, eles foram produzidos com base nas mesmas taxas de compressão que os extratos automáticos resultantes da Tarefa 4 (isto é, 30% e 70%).

Tarefa 6: Exploração da avaliação dos extratos automáticos

Para investigar a influência da língua materna dos participantes humanos no processo de seleção de conteúdo a compor um sumário de referência multilíngue, decidiu-se analisar a origem das informações contidas nos sumários de referência por meio de alinhamentos entre tais sumários e os textos-fonte. Quanto à influência da variação da taxa de compressão no desempenho dos métodos, compararam-se os extratos com 30% e 70% de compressão gerados pelos métodos CF e CFUL em termos da informatividade (via ROUGE) e da qualidade linguística (segundo os parâmetros da DUC'05).

1.4. Estrutura da dissertação

Este texto está estruturado em seis capítulos. No Capítulo 2, apresenta-se a revisão da literatura. No Capítulo 3, apresenta-se a seleção dos métodos de SAMM e do *corpus* utilizados nesta pesquisa, com destaque para a extensão do *corpus* selecionado, que englobou a produção dos diferentes sumários de referência. No Capítulo 4, discorre-se sobre a aplicação do método selecionado ao *corpus* estendido para a geração dos extratos para avaliação. No Capítulo 5, descreve-se especificamente a investigação sobre a avaliação dos extratos multilíngues. No Capítulo 6, tecem-se algumas considerações sobre o trabalho realizado e apontam-se possíveis desdobramentos da pesquisa ora descrita.

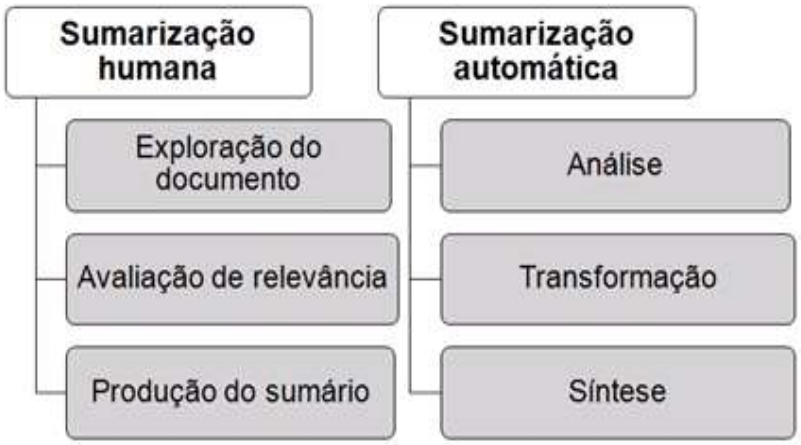
CAPÍTULO 2 – A Sumarização Automática Multidocumento Multilíngue

2.1 Conceitos básicos de Sumarização Automática

Como mencionado, a Sumarização Automática (SA) é a subárea do Processamento de Língua Natural (PLN) na qual se busca automatizar a produção de sumários (ou resumos) principalmente a partir de textos (MANI, 2001). Uma de suas motivações mais mencionadas é a enorme quantidade de informação disponível, sobretudo no meio digital, gerando dificuldades para que as pessoas a assimilem de forma efetiva.

Os sistemas que realizam essa tarefa de PLN são denominados sumarizadores automáticos. Sparck Jones (1993) sintetiza o mecanismo de funcionamento dessas ferramentas em três processos ideais: (i) análise, (ii) transformação e (iii) síntese. Posteriormente, Cremmins (1996) e Endres-Niegemeyer (1998) estabeleceram uma comparação entre os processos automáticos e as etapas da sumarização realizada por humanos, a saber, exploração do documento, avaliação de relevância e produção do sumário. A Figura 1 mostra a correlação entre os procedimentos mencionados.

Figura 1 – Etapas de sumarização humana e automática



Fonte: Sparck Jones (1993) com adaptação de Endres-Niegemeyer (1998)

Na análise, o sistema de SA interpreta os textos-fonte e extrai sua representação formal. Na transformação, com base na representação formal dos textos-fonte, o sistema seleciona o conteúdo mais relevante, de modo que o resultado possa ser uma versão reduzida do material original. Na síntese, o sistema emprega a representação interna produzida na etapa anterior para montar um sumário em língua natural. Isso pode ser

obtido de diversas formas, conforme Sparck Jones (1993), incluindo métodos de justaposição, ordenação, fusão e correferenciação do conteúdo selecionado.

Além do procedimento básico mencionado acima, a geração automática de sumários pode variar conforme diferentes fatores. Um dos mais evidentes é a extensão do sumário, regida pelo que se chama de taxa de compressão. Geralmente expressa em porcentagem, essa taxa indica a diferença entre os tamanhos do texto-fonte (fixado em 100%) e do sumário. Portanto, se o sumário tiver apenas 40% da extensão do texto-fonte, diz-se que ele foi gerado com uma taxa de compressão de 60%.

Outro fator relevante durante a produção de sumários é o tipo de conteúdo a ser veiculado com base em seu público-alvo. Principalmente quando o objetivo é sumarizar um texto muito extenso ou uma coleção de documentos sobre um tema, pode-se optar por um sumário que seja mais detalhado e que, portanto, contenha explicações mais pormenorizadas sobre termos ou expressões do texto. Essa opção se mostra mais vantajosa quando se sabe que a audiência não dispõe de vasto conhecimento sobre o assunto e que um sumário que não mencione informações contextuais pode não ser muito útil. Por outro lado, se o público-alvo do referido sumário for especializado e tiver conhecimento sobre o tema a ser abordado, podem-se dispensar tais elementos de contextualização. Esses dois casos refletem sumários focados nos interesses dos usuários, mas os criadores de um sistema de sumarização podem dar preferência à geração de sumários mais genéricos, que não façam distinção do tipo de leitor.

Um sumário também pode ser classificado conforme sua função. Diz-se que um sumário é indicativo quando seu papel é meramente identificar o tema central de um texto/conjunto de textos ou indicar os pontos que o leitor deve consultar para ter acesso a uma informação específica desejada. O sumário é considerado informativo quando apresenta o conteúdo principal do material original de forma coerente e coesa. Desse modo, entende-se que o sumário informativo pode servir de substituto ao(s) texto(s)-fonte, algo que o sumário indicativo não é capaz de realizar. Por fim, há os sumários críticos, que apresentam não apenas as informações centrais dos textos-fonte, mas também avaliações sobre elas. No caso, pode-se considerar a resenha de um livro como um tipo de sumário crítico, pois sua função também é veicular uma apreciação ou uma análise sobre a obra.

Outra forma de categorizar sumários está diretamente ligada à etapa de transformação do sistema de sumarização. Quando o sumário é composto exclusivamente por segmentos textuais literalmente presentes no(s) texto(s)-fonte, diz-se que ele é um extrato. Por outro lado, quando o sumário não apresenta trechos literais do(s) texto(s)-fonte, diz-se que ele é um *abstract*. Nesse caso, o sistema de sumarização precisa ser mais avançado, pois deve dispor de recursos que lhe permitam reformular trechos do conteúdo original e/ou produzir novo conteúdo. Ou seja, tais sistemas que geram *abstracts* não apenas atenderão às exigências de qualquer sumarizador comum, como também precisarão se ater às problemáticas presentes em sistemas de geração de língua.

A natureza do(s) texto(s)-fonte também é um fator relevante na configuração de sistemas de sumarização. Quando produzem sumários com base em um único texto-fonte, eles são considerados sistemas de SA monodocumento. Naturalmente, se pelo menos dois textos-fonte servirem de base para a geração do sumário, diz-se que há um sistema de SA multidocumento.

Nesse último caso, os textos-fonte podem estar no mesmo idioma ou não, acarretando uma ramificação dessa categoria em três modalidades de processamento. O primeiro tipo é chamado de *cross-language,* em que, a partir de um ou mais textos em uma língua L_x, produz-se um sumário em uma língua L_y. Segundo Wan *et al.* (2010), a forma mais produtiva de gerar sumários dessa natureza é sumarizar o(s) texto(s)-fonte e, na sequência, traduzi-lo(s), de modo a diminuir o volume de texto a ser traduzido e reduzir o impacto dos erros ainda causados por tradutores automáticos. Utilizando recursos de tradução neural e um *corpus* de notícias jornalísticas de língua inglesa, Ouyang *et al.* (2019) apresentaram um sistema robusto que produz sumários abstrativos em idiomas da África e da Ásia que não dispunham, até então, de *corpora* de sumarização.

A segunda modalidade de SA envolvendo mais de uma língua é chamada de multilíngue, em que, a partir de uma coleção de textos em diferentes idiomas, produz-se um sumário em uma dessas línguas. Um exemplo de trabalho nessa direção foi apresentado por Litvak *et al.* (2016), que desenvolveram uma plataforma de sumarização que demonstrou excelentes resultados ao gerar sumários em inglês, árabe e hebraico, mas que também foi testado em outras seis línguas. A arquitetura da ferramenta realiza a sumarização nas seguintes etapas: (i) pré-processamento

linguístico, incluindo segmentação de sentenças e palavras, remoção de *stopwords* e classificação morfossintática, (ii) treino do algoritmo supervisionado, gerando vetores para uma combinação linear dos *features* escolhidos, (iii) classificação das sentenças ou de partes delas, (iv) compressão ou extração de sentenças, (v) pós-processamento, incluindo resolução de anáforas e nomeação de entidades, e (vi) apresentação do resultado, que pode ser na forma de destaques nos textos-fonte ou produção de uma lista com as sentenças extraídas.

Por fim, outra modalidade é a SA independente de língua, como demonstra Orăsan (2009). Nesse modelo, os idiomas presentes importam pouco para a efetivação da tarefa. Por esse motivo, os procedimentos mais comuns baseiam-se em elementos puramente estatísticos com menor grau de conhecimento linguístico, a exemplo da frequência de termos, posição de um termo ou sentença no texto ou o tamanho desses elementos textuais.

A propósito desse assunto, vale a pena referir que os sistemas de SA podem ser categorizados com base na quantidade de conhecimento linguístico empregado. Segundo Mani (2001), a SA pode ter uma abordagem superficial quando é realizada com base em pouco ou nenhum conhecimento linguístico, recorrendo, na maior parte dos casos, a medidas estatísticas como as mencionadas no parágrafo anterior. Por isso, tais sistemas costumam produzir sumários extrativos que, muitas vezes, podem apresentar problemas de ordem linguística ou textual, como falta de coerência ou coesão. Por outro lado, tais sistemas são mais fáceis de construir e conseguem lidar com uma série de problemas, como a presença de elementos inesperados.

Quando o SA é feita com base em conhecimento linguístico codificado em gramáticas, repositórios semânticos e modelos de discurso, por exemplo, diz-se que é de abordagem profunda. Da mesma forma que os sistemas superficiais, os profundos apresentam vantagens e desvantagens: embora sejam capazes de produzir resultados mais satisfatórios linguisticamente, sua implementação é mais custosa e complexa.

Ainda, existem sistemas de SA que fazem uso de recursos linguísticos, mas que também levam em conta aspectos meramente estatísticos dos textos-fonte. Nesse caso, diz-se que tais sistemas apresentam uma abordagem híbrida.

2.2 Estratégias de avaliação em SA

Com o objetivo de avançar o "estado da arte" das aplicações de sumarização, diversas conferências internacionais foram realizadas, como a SUMMAC[2] (*Text Summarization Evaluation Conference*), a DUC[3] (*Document Understanding Conference*) e TAC (*Text Analysis Conference*). Nessas conferências, a importância e as dificuldades presentes na avaliação dos sistemas de SA ficaram evidentes.

De um modo geral, a avaliação desses sistemas pode ser classificada como intrínseca ou extrínseca. Na primeira, avalia-se o desempenho dos sistemas por meio da análise de seus resultados (neste caso, os próprios sumários). Na segunda, avalia-se a utilidade desses resultados para alguma tarefa principal, como a recuperação de informações (SPARCK JONES; GALLIERS, 1996).

Reconhece-se que a avaliação extrínseca é uma tarefa demorada, cara e de planejamento cuidadoso (VAN-HALTEREN; TEUFEL, 2003) e que a intrínseca deve focar a qualidade linguística e a informatividade dos sumários (MANI, 2001). A avaliação intrínseca, aliás, é a mais frequentemente realizada nos trabalhos de SA.

Inicialmente, a avaliação da qualidade dos sumários automáticos tinha uma tendência a ser feita manualmente por sujeitos humanos, uma vez que aspectos textuais como coesão e coerência nem sempre eram captados automaticamente de forma eficiente. Embora esses elementos sejam estudados por especialistas há pelo menos 15 anos, verifica-se, nos anos mais recentes, um crescimento no número de estudos dedicados à análise automática desses aspectos, a exemplo do trabalho realizado por Crossley *et al.* (2016), no qual os autores apresentam uma ferramenta de análise de coesão textual baseada em mais de 150 índices clássicos e recentes, divididos em coesão local, global e geral. Contudo, devido ao extenso repertório de recursos linguísticos necessário para viabilizar a avaliação automática de coesão e coerência, muitos dos estudos presentes na literatura ainda se concentram nos idiomas que dispõem de mais material, como o inglês e o mandarim.

Para avaliar a qualidade dos sumários na SAM, por exemplo, Dang (2005), no contexto da DUC'05, propôs que a qualidade linguística dos sumários seja avaliada em

---

[2] http://www-nlpir.nist.gov/related_projects/tipster_summac/

[3] Essa conferência foi organizada até 2007 pelo *National Institute of Standards and Technology* (NIST). A partir de 2008, a DUC passou a ser a *track* ("trilha") de SA da *Text Analysis Conference* (TAC), sendo realizada anualmente de 2008 a 2011. A última edição da TAC que promoveu a *track* de SA foi em 2014. Mais informações em: http://www.nist.gov/tac/about/index.html

função dos seguintes critérios: *gramaticalidade*, *não-redundância*, *clareza referencial*, *foco* e *estrutura e coerência*. Pelo critério *gramaticalidade*, o sumário não deve apresentar erros de ortografia, pontuação e sintaxe e nem problemas de formatação ou ainda a existência de erros que prejudiquem a legibilidade do texto (p.ex., sentenças agramaticais). O atributo linguístico *não-redundância* estabelece que o sumário não deve conter informações repetitivas, por exemplo, a repetição de fatos, nomes, sintagmas nominais ou até sentenças inteiras. Quanto à *clareza referencial*, o sumário deve fornecer a identificação clara de uma pessoa ou entidade sobre a qual os pronomes e sintagmas nominais se referem. Além disso, o sumário deve conter um *foco* temático que seja identificável por meio de informações inter-relacionadas, ou seja, as sentenças devem conter informações que se relacionem com as informações do sumário como um todo. Com relação ao atributo *estrutura/coerência*, o sumário deve apresentar estrutura e organização adequadas de forma a garantir que o encadeamento das sentenças construa uma estrutura informativa coerente sobre um tópico.

A avaliação da informatividade consiste em identificar o quanto de informação relevante dos textos-fonte o sumário automático incorpora. Essa identificação é feita pela comparação automática entre os sumários automáticos e os sumários humanos, também chamados "sumários de referência". Para tanto, utiliza-se com frequência o pacote de medidas denominado *Recall-Oriented Understudy of Gisting Evaluation* (ROUGE), que calcula a informatividade por meio da coocorrência de *n*-gramas entre os sumários automáticos e os humanos (LIN; HOVY, 2003).

Há várias métricas disponíveis no pacote: (i) ROUGE-N[4], que calcula a sobreposição de *n*-gramas, (ii) ROUGE-L, que retorna dados estatísticos ligados à subsequência comum mais comprida, (iii) ROUGE-W, que se assemelha à anterior, mas atribui um peso maior a subsequências comuns cujos elementos apresentam a mesma ordenação sequencial, e (iv) ROUGE-S, que avalia a coocorrência de bigramas do tipo *skip*, isto é, todos os bigramas possíveis em uma sentença, independentemente do posicionamento das palavras, considerando até termos não consecutivos.

Os dados referentes à informatividade de um extrato automático produzidos pelo pacote ROUGE são expressos em termos de precisão (P) (*precision*), cobertura (C) (*recall*) e medida-*f* (F) (*f-measure*). A precisão é obtida pelo quociente resultante da

[4] O índice N diz respeito ao comprimento do *grama* utilizado para a comparação. Assim, ROUGE-1 se refere à sobreposição de unigramas (sequência única de caracteres separados por espaços em branco), ROUGE-2 se refere à sobreposição de bigramas (sequências de dois unigramas) etc.

divisão do número de *n*-gramas em comum com o sumário de referência (que se chamará de n_c) pelo número total de *n*-gramas do sumário automático (n_{sa}). A cobertura, por sua vez, é obtida pela divisão do número de *n*-gramas em comum com o sumário de referência pelo número total de *n*-gramas do sumário de referência (n_{sr}). Em outras palavras, a precisão diz respeito ao teor de elementos identificados corretamente no cenário que inclui todos os elementos identificados pelo sistema (corretamente ou não), enquanto a cobertura mensura o teor de elementos identificados corretamente no cenário que inclui todos os elementos que deveriam ser originalmente identificados. A partir desses dois valores, tem-se a medida-*f*, que é, em suma, uma média harmônica entre a precisão e a cobertura. Em (1), (2) e (3), descrevem-se as fórmulas pelas quais tais medidas são calculadas. Esses cálculos geram valores entre 0 e 1, sendo que resultados mais próximos de 1 indicam melhores desempenhos em termos de informatividade.

(1) $P = \frac{n_c}{n_{sa}}$ (2) $C = \frac{n_c}{n_{sr}}$ (3) $F = 2 \cdot \frac{(P \cdot C)}{(P+C)}$

Dado que a ROUGE se baseia na sobreposição de *n-gramas*, sabe-se que esta apresenta certa limitação quanto à captura da similaridade de conteúdo. Isso acontece porque a comparação é feita com base em *tokens*, de modo que fenômenos como sinonímia e paráfrase não são considerados pelas medidas. Schluter (2017), aliás, pontua que os principais problemas da ROUGE estão ligados aos fatos de que (i) é difícil obter pontuações perfeitas para sumários extrativos, (ii) é impossível obter pontuações perfeitas para conjuntos de dados de alta qualidade, (iii) o próprio conceito de pontuação perfeita pode diferenciar de caso para caso e (iv) a sumarização automática no estado da arte é não-supervisionada. Diante de tais críticas, há outras formas de avaliação da qualidade de sumários automáticos.

Saggion *et al.* (2002), por exemplo, propuseram três métodos de avaliação baseados em conteúdo que medem a similaridade entre os sumários: (i) similaridade do cosseno, (ii) sobreposição de unidades lexicais (unigrama ou bigrama) e (iii) sobreposição da maior subsequência de unidades lexicais.

Van-Halteren e Teufel (2003) especificaram uma abordagem que combina dois aspectos: (i) comparação entre sumário automático e sumário de referência por meio de *factoids*, que são uma representação pseudossemântica das unidades de informação presentes nos textos-fonte (jornalísticos) e (ii) uso de um sumário consensual de referência, baseado em 50 *abstracts* de um mesmo texto.

Nenkova e Passonneau (2004) propuseram o “método da pirâmide”, no qual, a partir de conjunto de sumários de referência, extraem-se manualmente as “unidades de conteúdo do sumário” (*Summarization Content Units* ou SCU). As SCU são ponderadas em função do número de sumários de referência nos quais ocorre, de tal forma que aquelas mencionadas somente em um sumário são menos importantes do que aquelas mencionadas em vários sumários. Com base nos pesos das SCU, a pirâmide é construída, sendo que no topo ficam as SCU mais relevantes e, na base, as unidades menos relevantes. O sumário automático ideal deve conter as SCU que ocupam as posições superiores da pirâmide.

Louis e Nenkova (2013) apresentam três métodos de avaliação de sumários com o objetivo de reduzir a influência da subjetividade humana. O primeiro método mede a similaridade entre textos-fonte e sumários automáticos, ou seja, considera que, quanto maior a similaridade entre o sumário e seus textos-fonte, melhor é o seu conteúdo. No segundo, pseudomodelos (ou seja, sumários automáticos escolhidos por humanos) são adicionados a um conjunto de sumários de referência (humanos). Dessa forma, a avaliação final se dá pela comparação entre os sumários automáticos e o conjunto de referência expandido. No terceiro, utilizam-se apenas sumários automáticos como referência. Por meio de um cálculo probabilístico das palavras do conjunto de sumários automáticos (referência), obtém-se a distribuição global das palavras nesse conjunto, sendo que tal distribuição indica as informações mais importantes. Assim, a avaliação de um sumário automático é feita pela comparação de seu conjunto de palavras à distribuição global das palavras do conjunto de referência, pois se assume que um bom sumário automático tende a ter propriedades semelhantes à distribuição global.

2.3 A SAMM e a língua portuguesa

Quanto à SAMM que tem o português como língua-alvo, destacam-se Tosta, Di-Felippo e Pardo (2013), Tosta (2014), Di-Felippo, Tosta e Pardo (2016) e Camargo (2019).

Seguindo Evans, McKeown e Klavans (2004, 2005) e Roak e Fisher (2005), Tosta, Di-Felippo e Pardo (2013) propuseram métodos extrativos superficiais para a geração de sumários (informativos e genéricos) em português a partir de coleções multilíngues. Neles, a SAMM ocorre em quatro etapas gerais: (i) tradução automática dos textos-fonte em língua estrangeira para o português, (ii) ranqueamento das

sentenças com base em um atributo superficial de relevância, (iii) seleção das sentenças mais bem pontuadas, que devem expressar a informação central da coleção e (iv) justaposição das sentenças selecionadas. A diferença entre os métodos reside no critério de pontuação e ranqueamento das sentenças e no tratamento da redundância/tradução.

Assim especificados, os métodos investigados por Tosta, Di-Felippo e Pardo (2013) caracterizam-se por englobar a etapa de TA integral dos textos-fonte para o português antes do processo de seleção de conteúdo. Dessa forma, tais métodos seguem a abordagem *early-translation* e baseiam-se em conhecimento linguístico superficial para a seleção do conteúdo a compor o sumário multidocumento. No caso, a pontuação e o ranqueamento das sentenças dos textos-fonte são feitos com base em métodos clássicos de SA comumente utilizados no cenário multidocumento. A seguir, descrevem-se os dois métodos superficiais de melhor desempenho.

- **Método 1:** com base no critério da localização, as sentenças são caracterizadas em função de sua posição no texto-fonte da coleção. As sentenças contidas no primeiro parágrafo de cada um dos três textos são especificadas com o atributo localização="inicio", as sentenças localizadas no último parágrafo com o atributo localização="fim" e as demais, com o atributo localização="meio". Assim, o topo do ranque é ocupado pelas sentenças "início", seguidas pelas sentenças "meio" e, por fim, pelas sentenças "fim". A partir do ranque, a seleção manual de conteúdo no Método 1 consiste em (i) selecionar a sentença de maior pontuação do ranque para iniciar o sumário, (ii) selecionar a próxima sentença do ranque, (iii) calcular a redundância entre a nova sentença candidata e a sentença já selecionada para o sumário, (iv) selecionar a sentença candidata para compor o sumário, caso ela apresente pouca similaridade com a sentença inicialmente selecionada e não contenha problemas de TA, (v) substituir a sentença selecionada não-redundante com problemas de tradução por uma similar proveniente do texto-fonte original em português e (vi) repetir os passos para as demais sentenças do ranque até que a taxa de compressão de 70% fosse atingida. A similaridade, tanto para eliminar a redundância como para substituir sentenças traduzidas agramaticais por originais em português, é calculada de forma automática com base na medida estatística *word overlap* que se baseia na sobreposição das palavras de classe aberta idênticas (JURAFSKY; MARTIN, 2007). O cálculo *word overlap* entre sentenças é feito por

meio da aplicação da fórmula, descrita em (4). A sobreposição de palavras entre um par de sentenças (S1 e S2) é obtida pela divisão entre o número total de palavras em comum entre as sentenças (*CommonWords*) e o número total de palavras em ambas as sentenças (*Words(S1) + Words(S2)*), excluindo-se as palavras de classe fechada (como artigos e preposições), números e símbolos. O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo de 1 for a *Wol,* mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante. A produção dos extratos foi manual pela justaposição das sentenças na ordem em que foram selecionadas.

(4) $$Wol\ (S1, S2) = \frac{\#CommonWords}{\#Words(S1) + \#Words\ (S2)}$$

- **Método 2:** dada uma coleção, as sentenças dos textos-fonte recebem uma pontuação resultante da soma da frequência de ocorrência na coleção de suas palavras de classe aberta, a partir da qual são ranqueadas em ordem decrescente. Assim, o topo do ranque é ocupado pelas sentenças compostas pelas palavras mais frequentes. A pontuação e o ranqueamento são feitos por uma funcionalidade do sumarizador GistSumm (PARDO, 2005). Com base no ranque, a seleção manual de conteúdo no Método 2 segue os mesmos passos do Método 1, já que engloba o tratamento da redundância e dos problemas gerados pela TA. A produção dos extratos também é manual pela justaposição das sentenças na ordem em que foram selecionadas.

Para testar os métodos, os autores utilizaram cinco coleções trilíngues. Cada coleção possuía três notícias sobre um mesmo assunto (uma em português, uma em inglês e uma em espanhol), as quais foram compiladas manualmente das versões online dos jornais *A Folha de São Paulo*, *BBC News* e *El País*, respectivamente. Os textos em inglês e em espanhol foram traduzidos para o português via *Google Translator*[5]. Especificamente, os extratos gerados pelos Métodos 1 e 2 foram avaliados intrinsecamente quanto à qualidade linguística. Para tanto, um especialista avaliou cada extrato em função dos cinco parâmetros da DUC'05: gramaticalidade, não-redundância, clareza referencial, foco temático, e estrutura/coerência. Na Tabela 1, esquematiza-se a média obtida por cada método em uma escala de 1 a 5.

[5] Disponível em https://translate.google.com/

Tabela 1 – Avaliação da qualidade linguística: métodos superficiais de SAMM

Critério	Método	
	1	2
Gramaticalidade	3	2,8
Não-redundância	3	3
Clareza referencial	3,2	3
Foco temático	4	3,8
Estrutura e coerência	2,8	2,4

Fonte: Tosta, Di-Felippo e Pardo (2013).

Com base na Tabela 1, o Método 1, pautado na localização com tratamento da redundância e da tradução, obteve, em média, as mais altas pontuações quanto aos cinco parâmetros. embora a diferença seja discreta. Além disso, constatou-se que, apesar da aplicação da similaridade para a substituição das sentenças traduzidas por originais em português, os sumários ainda apresentam problemas de gramaticalidade. Uma possível explicação reside no fato de que alguns sumários apresentam algumas sentenças traduzidas que não eram redundantes, mas que possuíam problemas de tradução.

Tosta (2014), mais recentemente publicado em Di-Felippo, Tosta e Pardo (2016), desenvolveu dois métodos profundos de SAMM:

- ***Concept frequency method* (CF):** método que produz um extrato em português a partir de um texto em português e um em inglês com base na seleção das sentenças mais bem ranqueadas segundo a frequência de seus conceitos constitutivos em toda a coleção bilíngue de textos-fonte.
- ***Concept frequency* + *user language method* (CFUL):** método que produz um extrato em português a partir de um texto em português e um em inglês com base na seleção das sentenças exclusivamente em português mais bem ranqueadas segundo a frequência de seus conceitos constitutivos em toda a coleção bilíngue de textos-fonte.

Tais métodos realizam a SAMM em quatro etapas: (i) identificação dos conceitos nominais dos textos-fonte e cômputo de sua frequência na coleção, (ii) ranqueamento das sentenças pela soma da frequência de seus conceitos nominais, (iii) seleção das sentenças mais bem pontuadas e não redundantes entre si até que se atinja a extensão desejada e (iv) justaposição das sentenças para compor na ordem em que ocorrem no textos-fonte.

A diferença entre os métodos CF e CFUL reside na etapa (iii), pois o CFUL seleciona apenas as sentenças originalmente em língua portuguesa que estão mais bem ranqueadas. Caso o método CF selecione sentenças em língua inglesa, estas são traduzidas automaticamente para o português.

A avaliação desses métodos foi realizada com base no CM2News (1.0) (*Corpus* Multidocumento Bilíngue de Textos Jornalísticos) (TOSTA, 2014). Ele contém 40 notícias jornalísticas, totalizando 19.984 palavras e divididas em 20 coleções bilíngues (português e inglês) de assuntos diversos.

Para cada coleção, produziram-se manualmente um extrato conforme o método CF e um extrato conforme o Método CFUL. A taxa de compressão foi de 70% em relação ao texto mais extenso e aplicaram-se medidas de sobreposição de palavras (*word overlap*) para evitar redundância entre as sentenças. Quanto ao método CF, utilizou-se o Microsoft Bing[6] para a tradução ao português das sentenças selecionadas para o sumário que estavam em língua inglesa. A avaliação se baseou na informatividade e na qualidade linguística dos extratos. Os métodos foram comparados ao melhor *baseline* de Tosta, Di-Felippo e Pardo (2013), a saber, o método de posição da sentença com tratamento da redundância (Método 1 da Tabela 1).

Para determinar a qualidade dos extratos, utilizaram-se os cinco critérios da DUC'05, os quais foram analisados manualmente por quinze linguistas computacionais. As vinte coleções do *corpus* foram divididas em cinco grupos, cada um dos quais continha os extratos gerados pelos dois métodos de sumarização, totalizando oito extratos por grupo. Cada grupo de extratos foi avaliado por três juízes, que atribuíram pontuações de 1 a 5 a cada uma das cinco propriedades textuais, em que 1=muito ruim, 2=ruim, 3=aceitável, 4=bom e 5=muito bom. Os resultados estão na Tabela 2.

Tabela 2 – Avaliação da qualidade linguística dos métodos CF e CFUL no CM2News

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

Fonte: Di-Felippo, Tosta e Pardo (2016).

[6] Disponível em https://www.bing.com/translator

Tendo em conta os valores médios descritos na Tabela 2, observa-se que o método CFUL teve um desempenho superior ao CF e ao *baseline* em todos os critérios investigados, o que indica que a seleção de conteúdo com base no conhecimento conceitual e na língua do usuário funciona melhor no tratamento dos fatores textuais nos sumários. De certa forma, já se esperava esse desempenho superior, uma vez que as sentenças do método CFUL advêm de um único texto-fonte.

Quanto à avaliação da informatividade, utilizou-se a métrica automática ROUGE, tida como um padrão na área. Em particular, Di-Felippo, Tosta e Pardo (2016) empregaram a ROUGE-1, que determina a quantidade de sobreposições de unigramas entre sumários de referência e sumários automáticos, e a ROUGE-2, que determina a sobreposição de bigramas. Na Tabela 3, apresentam-se os resultados médios obtidos pela ROUGE-1 e pela ROUGE-2 em termos de cobertura, precisão e medida-f.

Tabela 3 – Avaliação da informatividade via ROUGE: métodos profundos de SAMM

Método	**ROUGE-1**			**ROUGE-2**		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
CF	0,355	0,328	0,341	0,155	0,144	0,149
CFUL	0,373	0,369	**0,371**	0,174	0,175	**0,174**
Baseline	0,313	0,271	0,285	0,038	0,032	0,034

Fonte: Di-Felippo, Tosta e Pardo (2016).

Conforme a Tabela 3, o método CFUL tem um desempenho superior ao CF e ao *baseline* em ambas as métricas. Mais uma vez, esses dados dão indícios de que extratos construídos apenas com sentenças originais do texto na língua-alvo veiculam as informações centrais da coleção.

Camargo (2019) tem investigado uma refinação do método CFUL ao (i) atribuir uma pontuação diferenciada aos conceitos superordenados que estão em relação hierárquica a outros na coleção, sob a hipótese de que veiculam informações mais genéricas e, portanto, relevantes para extratos informativos e (ii) tratar a redundância com base na sobreposição de conceitos, buscando capturar mais adequadamente a similaridade de conteúdo entre as sentenças. Para tanto, Camargo e Di-Felippo (2019) adicionaram dez novas coleções bilíngues ao *corpus* CM2News (1.0) e anotaram os nomes dos textos-fonte segundo as diretrizes de Tosta (2014), resultando em uma nova versão do referido recurso linguístico-computacional, o CM2News (2.0).

CAPÍTULO 3 – A seleção dos métodos de SAMM e do *corpus*

3.1 Os métodos profundos CF e CFUL

Tendo em vista a revisão da literatura sobre SAMM em que o português é a língua de interesse, selecionaram-se os métodos de melhor desempenho para a exploração da avaliação aqui proposta. No caso, tratam-se dos métodos profundos CFUL e CF, desenvolvidos por Tosta (2014), os quais serão descritos em detalhes na sequência.

Para apresentar as estratégias de SAMM, consideram-se as fases tradicionais da sumarização: análise, transformação e síntese (SPARCK JONES, 1993). A análise corresponde à interpretação dos textos-fonte, gerando uma representação interna de seu conteúdo. A transformação realiza operações de sumarização, produzindo a representação interna do sumário. Na etapa de síntese, a representação interna do sumário é linguisticamente concretizada, resultando no sumário final.

Nos métodos CF e CFUL, a análise consiste em identificar os conceitos expressos por nomes comuns, que compõem a classe morfossintática mais frequente e cobrem parte do conteúdo principal dos textos. A fim de identificar os conceitos nominais, os métodos empregam a WordNet de Princeton[7] (FELLBAUM, 1998) como o repositório conceitual. Embora a granularidade do inventário de conceitos seja uma vantagem para essa tarefa, verifica-se, às vezes, que tal granularidade pode ser excessiva, dificultando a identificação do *synset* que melhor represente um conceito a ser anotado. Apesar disso, optou-se pelo uso da WN.Pr devido (i) a seu uso generalizado na área de sumarização e em outras aplicações do PLN, (ii) ao fato de ter sido produzido manualmente e (iii) ao fato de que tais recursos para a língua portuguesa ainda são parciais. Considerando que um conceito é codificado na WN.Pr através de um conjunto de sinônimos (um *synset*) em língua inglesa, a anotação dos nomes em textos em outras línguas apresenta um novo desafio: sua tradução para o inglês. Na próxima

[7] A WN.Pr é uma rede em que as palavras e expressões do inglês, pertencentes às categorias dos nomes, verbos, adjetivos e advérbios, organizam-se sob a forma de *synsets* (*synonym sets*). Assim, o *synset* é um conjunto de formas (*word forms*) de uma mesma categoria gramatical que podem ser intercambiáveis em determinado contexto, como {*bicycle, bike, wheel, cycle*}. O *synset* é construído de modo a codificar um único conceito lexicalizado por suas formas constituintes. Entre os *synsets*, codificam-se cinco principais relações lógico-conceituais: antonímia, hiponímia, meronímia, acarretamento e causa. Entre os conceitos nominais, a relação de hiponímia é a mais proeminente. Essa relação ocorre entre um conceito específico (hipônimo) e um conceito genérico (hiperônimo) O *synset* {*car, auto, automobile, machine, motor car*}, por exemplo, é hipônimo de {*motor vehicle, automotive vehicle*}. (FELLBAUM, 1998).

Seção, descreve-se o *corpus* utilizado pelos métodos, bem como o processo de anotação léxico-conceitual desse material, o qual foi necessário para a aplicação dos métodos CF e CFUL.

A transformação corresponde à seleção do conteúdo. Para selecionar as sentenças a comporem o sumário, os métodos executam quatro etapas: (i) calcular a taxa de compressão, (ii) calcular a frequência de cada conceito nominal na coleção, (iii) atribuir pontuações às sentenças em função da frequência de ocorrência de seus conceitos nominais na coleção e (iv) ordenar as sentenças com base em suas pontuações. Quanto à etapa (ii), o cômputo da frequência conceitual agrupa a ocorrência de diferentes palavras na mesma língua que expressam o mesmo conceito, bem como equivalências, ou seja, expressões de um mesmo conceito em diferentes idiomas.

A título de exemplificação, podem-se observar as duas sentenças da Figura 2, que pertencem à mesma coleção e passaram pela anotação de seus conceitos expressos por nomes. Os números dentro dos parênteses angulares indicam o código identificador do *synset* do conceito nominal, enquanto os números entre parênteses indicam a frequência de cada conceito/*synset* na coleção. Os nomes "manifestante" e "*protester*", por exemplo, expressam o mesmo conceito (isto é, "uma pessoa que discorda de uma norma estabelecida"), codificado pelas palavras {*dissenter, dissident, protester, objector, contestant*}. A frequência do conceito nessa coleção foi 16, e tal valor é associado a todas as ocorrências de nomes que lexicalizam o conceito em questão.

Após o cálculo da frequência de todos os conceitos, as sentenças são ordenadas em função da soma da frequência dos conceitos que as constituem. A sentença em português obteve pontuação 51, ocupando a 1ª posição no ranque, enquanto a sentença em inglês, com pontuação 28, ocupa a 12ª posição. Contendo os conceitos mais frequentes, as sentenças mais bem pontuadas veiculam o conteúdo principal da coleção. Assim, sentenças bem pontuadas são as mais adequadas para o sumário.

Figura 2 – Exemplo de pontuação/ranqueamento sentencial nos métodos CF e CFUL

Sentences	Score	Rank
Um grupo<31264>(6) de **manifestantes<10002760>(16)** conseguiu furar o bloqueio<8376948>(2) da Polícia Militar e chegar ao estádio<4295881>(14) Mané Guarrincha neste sábado<15164570>(4), horas<15227846>(2) antes do jogo<7470671(5) de abertura<7452699>(2) da Copa das Confederações.[10]	51	1rst
Brazil's <9379111>(4) opening<74522699>(2) Confederations Cup match<7470671>(5) was affected by **protesters<10002760>(16)** that left 39 people<7942152>(1) injured.	28	12th

Fonte: Di-Felippo, Tosta e Pardo (2016).

Com o ranque montado, o método CF realiza a seleção de sentenças exclusivamente com base na classificação, independentemente da língua-fonte. Especificamente, o CF seleciona inicialmente a sentença com a maior pontuação para compor o sumário (em português) e, caso essa sentença (ou qualquer outra) esteja em inglês, ela é automaticamente traduzida para o português. Se a taxa de compressão, calculada após a tradução, não for atingida após a primeira seleção, mais conteúdo precisará ser selecionado (por exemplo, a segunda sentença mais bem pontuada). Como o material-fonte é uma coleção multidocumento, é preciso verificar a redundância entre a nova sentença candidata e aquela que já foi selecionada, pois o sumário deve refletir os diferentes tópicos da coleção sem redundância. Para evitar que isso ocorra, assumiu-se um limite que a nova sentença deve respeitar em relação a qualquer uma das sentenças anteriormente selecionadas. Assim, se tal limite for ultrapassado, a nova sentença é considerada redundante e não entra para o sumário, de modo que o processo de sumarização segue com a sentença seguinte. Caso contrário, a sentença é incluída no sumário. Se duas sentenças tiverem a mesma pontuação no ranque, o método CF seleciona a mais curta. Esse processo se repete até que se atinja o comprimento desejado para o sumário. O método CF foi proposto sob a hipótese de que uma estratégia de *late-translation*, em que a TA só é utilizada para traduzir as sentenças selecionadas, minimiza os problemas causados pela TA integral dos textos-fonte.

O outro método, CFUL, orienta-se pelo idioma do usuário. Ele seleciona exclusivamente as sentenças mais bem pontuadas que estejam no texto-fonte em português, evitando também a redundância. Assim como o CF, se duas sentenças tiverem a mesma pontuação no ranque, o CFUL seleciona a mais curta. Por consequência, o sumário final conterá apenas sentenças na língua de interesse. Essa abordagem se baseia na hipótese de que um sumário composto apenas por sentenças originalmente em português refletirá as informações mais relevantes da coleção, uma vez que os conceitos presentes no texto em inglês também são considerados para a classificação das sentenças.

Por fim, na etapa de síntese, os métodos geram os extratos, assim como a vasta maioria dos trabalhos em SA atualmente. Para tanto, os métodos CF e CFUL simplesmente justapõem as sentenças selecionadas do ranque, ordenando-as conforme suas posições nos respectivos textos-fonte.

Com base na escolha dos métodos CF e CFUL, fez-se necessário o uso de um *corpus* anotado em nível conceitual. Assim, optou-se pelo CM2News (TOSTA, 2014), cujas características iniciais e posterior extensão são detalhadas a seguir.

3.2 O *corpus* CM2News

Para as pesquisas em SAMM envolvendo o português como língua de interesse, tinha-se como recurso principal à época da seleção do *corpus* para esta pesquisa, o CM2News (TOSTA, 2014), que é um *corpus* multidocumento bilíngue de textos jornalísticos. Destaca-se que o *corpus* serviu de base não só para Tosta (2014), mas também para que Chaud (2015) investigasse métricas estatísticas e conhecimento conceitual para captar a relevância do conteúdo em coleções multilíngues. A Tabela 4 especifica o conteúdo do CM2News.

Com base na Tabela 4, verifica-se que o CM2News contém 20 coleções bilíngues (português-inglês). Especificamente, cada coleção é formada por duas notícias sobre o mesmo acontecimento ou evento, sendo uma em português e outra em inglês. Segundo Tosta (2014), a escolha do português e do inglês como línguas constitutivas do *corpus* foi feita com o objetivo de produzir sumários (multilíngues) em português a partir de textos nessa mesma língua e em inglês, que é o idioma em que há mais informações disponíveis na *Web.* A escolha pelo gênero jornalístico foi feita em função da tradição dos trabalhos em SAM, que comumente focam esse gênero, e devido à facilidade de obtenção de textos desse gênero que versam sobre um mesmo assunto a partir de fontes distintas e em diferentes línguas.

Destaca-se também que as coleções do CM2News abrangem seis domínios distintos, a saber, mundo, poder, saúde, ciência, ambiente e entretenimento. Tais domínios buscavam cobrir eventos variados e atuais à época da construção do *corpus* (2011 a 2013). Todos os textos em português foram extraídos do jornal *Folha de São Paulo*[8] e os textos em inglês foram selecionados do portal *BBC*[9] devido ao grau de confiabilidade das notícias e da qualidade linguística dos textos jornalísticos. As notícias foram compiladas com base em seu tamanho e originalidade. Tosta (2014) buscou compilar textos de tamanho (em número de palavras) similar. Quanto à

[8] Disponível em http://www.folha.uol.com.br/
[9] Disponível em http://www.bbc.co.uk/news/

originalidade dos textos, o autor preocupou-se em selecionar textos que versassem sobre um mesmo assunto ou tema, mas que não fossem traduções um do outro.

Tabela 4 – Descrição do CM2News

Coleção	Domínio	Assunto/Tema	Documento	Língua	Publicação (data/hora)	Qt. pal.
C1	Mundo	Ataques em Londres	D1_C1_folha	PT	11/08/2011 – 09:11	1.311
			D2_C1_bbc	IN	11/08/2011 – 11:10 (GMT)	
C2	Poder	Kit gay	D1_C2_folha	PT	25/05/2011 – 13:12	516
			D2_C2_bbc	IN	25/05/2011 – 21:07 (GMT)	
C3	Saúde	Intoxicação alimentar	D1_C3_folha	PT	30/05/2011 – 18:47	1.419
			D2_C3_bbc	IN	30/05/2011 – 5:43 (GMT)	
C4	Mundo	Massacre na Noruega	D1_C4_folha	PT	08/08/2011 – 14h20	911
			D2_C4_bbc	IN	02/08/2011 – 14:52 (GMT)	
C5	Ambiente	Novo código florestal	D1_C5_folha	PT	25/05/2011– 00:43	1.217
			D2_C5_bbc	IN	25/05/2011– 09:50 (GMT)	
C6	Mundo	Conflito na universidade da CA	D1_C6_folha	PT	20/11/2011– 00:15	645
			D2_C6_bbc	IN	21/11/2011– 23:26 (GMT)	
C7	Saúde	Proibição do fumo em NY	D1_C7_folha	PT	24/05/2011– 13:38	887
			D2_C7_bbc	IN	24/05/2011– 18:36 (HKT)	
C8	Mundo	Terremoto na Nova Zelândia	D1_C8_folha	PT	05/03/2011– 05:01	948
			D2_C8_bbc	IN	03/03/2011– 04:45 (GMT)	
C9	Mundo	Terremoto em Missouri	D1_C9_folha	PT	23/05/2011– 08:04	1.169
			D2_C9_bbc	IN	23/05/2011– 20:21 (GMT)	
C10	Mundo	Erupção vulcânica na Islândia	D1_C10_folha	PT	24/05/2011– 12:13	1.476
			D2_C10_bbc	IN	24/05/2011– 15:51 (GMT)	
C11	Ciência	Patentes genes humanos	D1_C11_bbc	PT	13/07/2013- 16:34 (GMT)	963
			D2_C11_folha	IN	13/06/2013-23:50	
C12	Poder	Protestos: transporte	D1_C12_folha	PT	14/06/2013-07:25	808
			D2_C12_bbc	IN	14/06/2013-12:43 (GMT)	
C13	Mundo	Eleições do Irã	D1_C13_folha	PT	15/06/2013 – 17:57	1.266
			D2_C13_bbc	IN	16/06/2013 - 08:38 (GMT)	
C14	Saúde	Epidemia de dengue no MS	D1_C14_folha	PT	11/01/2013 1-9:03	534
			D2_C14_bbc	IN	21/01/2013- 00:21 (GMT)	
C15	Saúde	Mastectomia preventiva	D1_C15_folha	PT	15/05/2013 – 03:01	1.367
			D1_C15_bbc	IN	14/05/2013 -17:02 (GMT)	
C16	Ciência	Missão espacial chinesa	D1_C16_folha	PT	11/06/2013 – 21:06	793
			D2_C16_bbc	IN	11/06/2013-9:38 (GMT)	
C17	Poder	Protesto: Copa das Confederações	D1_C17_folha	PT	15/06/2013 – 14:53	918
			D2_C17_bbc	IN	16/06/2013 -13:19 (GMT)	
C18	Ciência	Viagra feminino	D1_C18_folha	PT	16/06/2013 – 03:30	975
			D2_C18_bbc	IN	17/11/2009- 9:35 (GMT)	
C19	Entreten.	Lançamento: Homem de Aço	D1_C19_folha	PT	16/06/2013-13:24	898
			D2_C19_bbc	IN	11/06/2013-10:17(GMT)	
C20	Mundo	Conflito na Turquia	D1_C20_folha	PT	17/06/2013 - 09h44	963
			D2_C20_bbc	IN	17/06/2013-13:00(GMT)	
Total de palavras						19.984

Fonte: Tosta (2014).

Além disso, salienta-se que cada coleção do CM2News contém (i) um sumário humano de referência (*abstract*) produzido por falantes nativos da língua portuguesa com base no conteúdo de ambos os textos-fonte da coleção, (ii) um extrato automático em português gerado pelo melhor *baseline* de Tosta, Di-Felippo e Pardo (2013) (Método 1), (iii) dois extratos automáticos em português, sendo um gerado pelo método profundo CF e um pelo CFUL, e (iv) anotação léxico-conceitual dos nomes de ambos os textos-fonte. A taxa de compressão de todos esses os sumários (manual e automáticos) foi de 70% (equivalente a 30% do tamanho do texto mais extenso da coleção).

Ainda com base na Tabela 4, vê-se que as quarenta notícias do CM2News totalizam quase 20 mil palavras. Para a exploração da avaliação dos extratos multilíngues aqui proposta, que inclui a variação (i) da taxa de compressão dos extratos automáticos (isto é, a extensão desejada) e (ii) da língua nativa dos produtores dos sumários de referência, optou-se pela ampliação do *corpus*.

3.3 A extensão do CM2News: construção do CM3News

A extensão do CM2News consistiu no acréscimo de mais uma língua estrangeira ao *corpus* que, originalmente, era composto por duas línguas, sendo o português a língua-alvo e o inglês a língua estrangeira. Como a extensão, o *corpus* passou a englobar três idiomas, motivando sua renomeação para CM3News (*Corpus* Multidocumento Trilíngue de Textos Jornalísticos). Como as demais línguas do *corpus*, o alemão também está entre as mais utilizadas atualmente na *Web*, apesar de ter um número significativamente inferior de falantes nativos no planeta, os quais ficam quase totalmente restritos ao continente europeu. Na Tabela 5, tem-se o ranque ocupado pelas línguas que compõem o CM3News quanto ao número de falantes. Na Tabela 6, apresenta-se a representatividade dessas línguas na *Web*.

Tabela 5 – A representatividade das línguas do *corpus*: número mundial de falantes

Língua	**1ª língua**	**2ª língua**	**Total (Posição)**
	Falantes (Posição)	Falantes (Posição)	
Inglês	379 milhões (3º)	753 milhões (1º)	1,13 bilhão (1º)
Português	221 milhões (6º)	13 milhões (15º)	234 milhões (9º)
Alemão	76 milhões (16º)	56 milhões (11º)	132 milhões (12º)

Fonte: Eberhard *et al.* (2019)

Tabela 6 – A representatividade das línguas do CM3News em *websites*

Posição	Língua	Uso (julho de 2018)	Uso (julho de 2019)
1	Inglês	52,5%	53,9%
2	Russo	6,2%	6,1%
3	Alemão	6,3%	5,7%
4	Espanhol	5,1%	5,0%
5	Francês	4,1%	3,9%
6	Japonês	4,0%	3,5%
7	Português	2,9%	2,9%

Fonte: W3TECHS (2019)

Dos 20 *clusters* originais do CM2News, salienta-se que apenas 19 deles compõem o CM3News. A coleção C14 do CM2News não foi integrada ao CM3News por não ter sido possível encontrar uma notícia jornalística em alemão sobre o evento coberto pela coleção. Portanto, criou-se a C21 em substituição à C14 para que o novo *corpus* tivesse o total de 20 coleções.

Para compilar os textos em alemão (e os textos em português e em inglês da C21), empregaram-se os mesmos critérios aplicados aos textos já presentes no CM2News: (i) tamanho das notícias e (ii) confiabilidade das fontes. Quanto ao tamanho, buscou-se, dada uma coleção Cx, compilar um texto em alemão que tivesse tamanho compatível aos demais textos de Cx. Tendo em vista que as notícias que integram as coleções iniciais do CM3News (advindas do CM2News) foram compiladas entre 2011 e 2013, a identificação de notícias em alemão de tamanho similar nem sempre foi possível devido à escassez de material em alemão sobre tais eventos disponível na *Web* em 2018/2019. Aliás, a busca por satisfazer a diretriz (i) levou à utilização de diversas fonte de notícias, todas elas consideradas confiáveis, como a revista alemã Der Spiegel[10], o jornal Die Welt[11], entre outras[12].

Na Tabela 7, tem-se a quantidade de palavras de cada texto-fonte do CM3News.

[10] Disponível em https://www.spiegel.de
[11] Disponível em https://www.welt.de
[12] Revistas Stern (https://www.stern.de) e Queer (https://www.queer.de), os jornais Rheinische Post (https://rp-online.de), Frankfurter Allgemeine Zeitung (https://www.faz.net), Die Tageszeitung (https://taz.de) e Bild (https://www.bild.de), os jornais suíços 20 Minuten (https://www.20min.ch) e Neue Zürcher Zeitung (https://www.nzz.ch), e a empresa pública de radiodifusão Deutsche Welle (https://www.dw.com/de).

Tabela 7 – Quantidade de palavras por texto-fonte no CM3News

Coleção	Tema/Assunto	Quantidade de palavras			
		Português	Inglês	Alemão	TOTAL
C1	Ataques na Inglaterra	518	788	910	2.216
C2	Kit gay	287	231	393	911
C3	Intoxicação alimentar	716	700	631	2.047
C4	Massacre na Noruega	357	557	174	1.088
C5	Novo Código Florestal	706	588	835	2.129
C6	Conflito em universidade nos EUA	291	358	573	1.222
C7	Proibição do fumo nos EUA	373	511	236	1.120
C8	Terremoto na Nova Zelândia	394	550	340	1.284
C9	Terremoto nos EUA	544	750	521	1.815
C10	Erupção vulcânica na Islândia	844	905	814	2.563
C11	Patentes de genes humanos	518	466	474	1.458
C12	Protestos sobre transporte no Brasil	521	289	460	1.270
C13	Eleições no Irã	589	682	468	1.739
C15	Mastectomia preventiva	604	767	652	2.023
C16	Missão espacial chinesa	348	446	436	1.230
C17	Protestos sobre a Copa das Confederações	638	280	661	1.579
C18	Viagra feminino	674	304	297	1.275
C19	Lançamento de filme	449	466	536	1.451
C20	Conflitos na Turquia	515	447	412	1.374
C21	Queda de ponte na Itália	486	575	700	1.761

Fonte: Elaborado pelo autor.

Na Tabela 8, apresentam-se os dados sobre a comparação da extensão (em número de palavras) dos textos-fonte em cada coleção. Na segunda coluna da tabela, evidencia-se a média de palavras dos textos da coleção. Nas três colunas seguintes, tem-se a comparação entre a extensão de cada texto e a média da coleção. Para tanto, dividiu-se o número de palavras de cada texto (português, inglês ou alemão) pela respectiva média da coleção. Assim, quanto mais próximo de 100%, mais o tamanho do texto se assemelha à média da coleção.

Embora essa não seja uma forma tradicional de se analisar o balanceamento de um *corpus*, as porcentagens como as da Tabela 8 permitiram observar rapidamente (i) se as extensões dos textos eram muito diferentes e (ii) se um eventual desbalanceamento estava vinculado a alguma língua específica.

Tabela 8 – Proporção de texto-fonte por coleção no CM3News

Coleção	Tamanho médio do texto (da coleção)	Variação de cada texto-fonte em relação à média		
		Português	Inglês	Alemão
C1	739	-29,9%	6,7%	23,2%
C2	304	-5,5%	-23,9%	29,4%
C3	682	4,9%	2,6%	-7,5%
C4	363	-1,6%	53,6%	-52,0%
C5	710	-0,5%	-17,1%	17,7%
C6	407	-28,6%	-12,1%	40,7%
C7	373	-0,1%	36,9%	-36,8%
C8	428	-7,9%	28,5%	-20,6%
C9	605	-10,1%	24,0%	-13,9%
C10	854	-1,2%	5,9%	-4,7%
C11	486	6,6%	-4,1%	-2,5%
C12	423	23,1%	-31,7%	8,7%
C13	580	1,6%	17,7%	-19,3%
C15	674	-10,4%	13,7%	-3,3%
C16	410	-15,1%	8,8%	6,3%
C17	526	21,2%	-46,8%	25,6%
C18	425	58,6%	-28,5%	-30,1%
C19	484	-7,2%	-3,7%	10,8%
C20	458	12,4%	-2,4%	-10,0%
C21	587	-17,2%	-2,0%	19,3%

Fonte: Elaborado pelo autor.

Considerando como aceitável um desvio de até 15 pontos percentuais em relação à média da coleção (isto é, textos com pontuações entre -15% e 15% na Tabela 8), destaca-se que 32 dos 60 textos (53,3%) são condizentes com a média da coleção (estando, portanto, balanceados). Se esse desvio for ampliado para até 20 pontos percentuais (textos com pontuações entre -20% e 20%), o número de textos condizentes com a média aumente de 32 para 39 (dos 60 textos) (65%). Observa-se, além disso, que 10 das 20 coleções (50%) satisfazem o desvio de até 20 pontos percentuais nos três textos.

Embora os valores ora mencionados não sejam excepcionalmente significativos para a constatação de que as coleções sejam balanceadas, a variação das médias de cada língua para o conjunto das 20 coleções parece mais promissora: 0% para o português, 1% para o inglês e -1% para o alemão. Em quantidade de palavras, isso significa que o

CM3News tem, no total, 10.372 palavras em português, 10.660 em inglês e 10.523 em alemão.

Ao final, os 60 textos-fonte do CM3News totalizam 31.555 palavras. Na Tabela 9, descrevem-se as coleções finais do referido *corpus*. Os horários de publicação de cada matéria correspondem ao respectivo fuso local. As línguas são indicadas por seus códigos internacionais: PT para o português, EN para o inglês e DE para o alemão.

Tabela 9 – Descrição do CM3News

Coleção	Domínio	Assunto/Tema	Documento	Língua	Publicação (data/hora)	Qt. pal.
C1	Mundo	Ataques em Londres	D1_C1_folha	PT	11/08/2011 – 09:11	2.216
			D2_C1_bbc	EN	11/08/2011 – 11:10	
			D3_C1_stern	DE	11/08/2011 – 19:44	
C2	Poder	Kit gay	D1_C2_folha	PT	25/05/2011 – 13:12	911
			D2_C2_bbc	EN	25/05/2011 – 21:07	
			D3_C2_queer	DE	26/05/2011 – 00:00	
C3	Saúde	Intoxicação alimentar	D1_C3_folha	PT	30/05/2011 – 18:47	2.047
			D2_C3_bbc	EN	30/05/2011 – 05:43	
			D3_C3_rp	DE	25/05/2011 – 21:21	
C4	Mundo	Massacre na Noruega	D1_C4_folha	PT	08/08/2011 – 14:20	1.088
			D2_C4_bbc	EN	02/08/2011 – 14:52	
			D3_C4_presse	DE	07/08/2011 – 11:28	
C5	Ambiente	Novo Código Florestal	D1_C5_folha	PT	25/05/2011 – 00:43	2.129
			D2_C5_bbc	EN	25/05/2011 – 09:50	
			D3_C5_dw	DE	26/05/2011 – 00:00	
C6	Mundo	Conflito na universidade da CA	D1_C6_folha	PT	20/11/2011 – 00:15	1.222
			D2_C6_bbc	EN	21/11/2011 – 23:26	
			D3_C6_spiegel	DE	22/11/2011 – 16:18	
C7	Saúde	Proibição do fumo em NY	D1_C7_folha	PT	24/05/2011 – 13:38	1.120
			D2_C7_bbc	EN	24/05/2011 – 18:36	
			D3_C7_spiegel	DE	23/05/2011 – 15:54	
C8	Mundo	Terremoto na Nova Zelândia	D1_C8_folha	PT	05/03/2011 – 05:01	1.284
			D2_C8_bbc	EN	03/03/2011 – 04:45	
			D3_C8_spiegel	DE	04/03/2011 – 19:52	
C9	Mundo	Terremoto em Missouri	D1_C9_folha	PT	23/05/2011 – 08:04	1.815
			D2_C9_bbc	EN	23/05/2011 – 20:21	
			D3_C9_welt	DE	23/05/2011 – 00:00	
C10	Mundo	Erupção vulcânica na Islândia	D1_C10_folha	PT	24/05/2011 – 12:13	2.563
			D2_C10_bbc	EN	24/05/2011 – 15:51	
			D3_C10_spiegel	DE	24/05/2011 – 11:54	
C11	Ciência	Patentes genes	D1_C11_folha	PT	13/06/2013 – 16:34	1.458

		humanos	D2_C11_bbc	EN	13/06/2013 – 23:50	
			D3_C11_faz	DE	13/06/2013 – 17:35	
C12	Poder	Protestos: Transporte	D1_C12_folha	PT	14/06/2013 – 07:25	1.270
			D2_C12_bbc	EN	14/06/2013 – 12:43	
			D3_C12_spiegel	DE	14/06/2013 – 20:20	
C13	Mundo	Eleições do Irã	D1_C13_folha	PT	15/06/2013 – 17:57	1.739
			D2_C13_bbc	EN	16/06/2013 – 08:38	
			D3_C13_welt	DE	15/06/2013 – 00:00	
C15	Saúde	Mastectomia preventiva	D1_C15_folha	PT	15/05/2013 – 03:01	2.023
			D2_C15_bbc	EN	14/05/2013 – 17:02	
			D3_C15_spiegel	DE	14/05/2013 – 14:39	
C16	Ciência	Missão espacial chinesa	D1_C16_folha	PT	11/06/2013 – 21:06	1.230
			D2_C16_bbc	EN	11/06/2013 – 09:38	
			D3_C16_20min	DE	11/06/2013 – 12:18	
C17	Poder	Protesto na Copa das Confederações	D1_C17_folha	PT	15/06/2013 – 14:53	1.579
			D2_C17_bbc	EN	16/06/2013 – 13:19	
			D3_C17_nzz	DE	17/06/2013 – 14:26	
C18	Ciência	Viagra feminino	D1_C18_folha	PT	16/06/2013 – 03:30	1.275
			D2_C18_bbc	EN	17/11/2009 – 09:35	
			D3_C18_taz	DE	05/06/2015 – 00:00	
C19	Entreten.	Lançamento: Homem de Aço	D1_C19_folha	PT	16/06/2013 – 13:24	1.451
			D2_C19_bbc	EN	11/06/2013 – 10:17	
			D3_C19_stern	DE	17/06/2013 – 12:50	
C20	Mundo	Conflito na Turquia	D1_C20_folha	PT	17/06/2013 – 09:44	1.374
			D2_C20_bbc	EN	17/06/2013 – 13:00	
			D3_C20_welt	DE	17/06/2013 – 00:00	
C21	Mundo	Queda de ponte na Itália	D1_C21_g1	PT	14/08/2018 – 07:29	1.761
			D2_C21_npr	EN	14/08/2018 – 07:38	
			D3_C21_bild	DE	14/08/2018 – 21:30	
Total de palavras						31.555

Fonte: Elaborado pelo autor com base em Tosta (2014).

A seguir, descreve-se a anotação léxico-conceitual dos novos textos do CM3News.

3.4 A anotação léxico-conceitual do CM3News

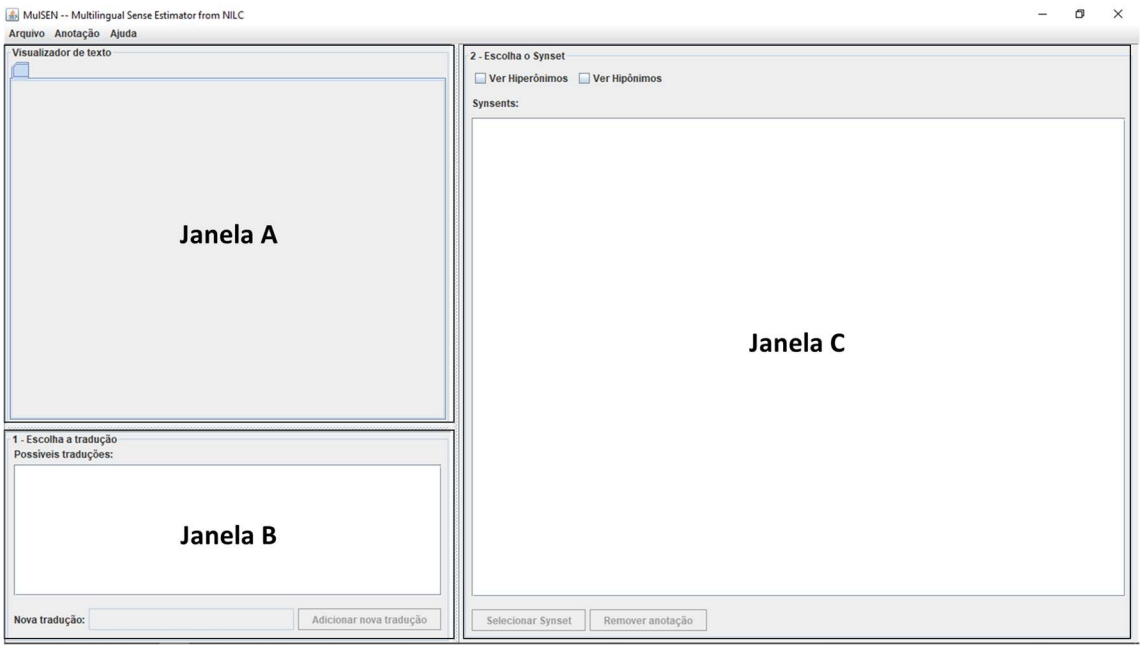
A anotação léxico-conceitual do CM3News concentrou-se quase que exclusivamente nos 20 textos em língua alemã, com exceção da coleção C21, cujos textos em português e inglês também foram anotados. Para tal anotação, utilizaram-se dois recursos básicos:

(i) o *Multilingual Sense Estimator from NILC*[13] (MulSEN[14]), editor que fora desenvolvido para a anotação dos textos em português e em inglês do CM2News, e (ii) o conjunto de regras de anotação de Tosta (2014), que foram anteriormente aplicadas ao CM2News. Após a descrição geral de tais recursos, apresenta-se, na Seção 3.4.3, como se deu a anotação dos textos-fonte em alemão em função desses recursos.

3.4.1 O editor MulSEN e suas funcionalidades gerais

Para a anotação do *corpus* CM2News em nível léxico-conceitual, desenvolveu-se o MulSEN[15] (TOSTA, 2014), que é uma ferramenta com interface gráfica de auxílio à anotação manual dos conceitos expressos pelos nomes por meio dos *synsets* da WN.Pr. Na Figura 3, exibe-se a tela principal do editor MulSEN.

Figura 3 – Interface do editor MulSEN



A Janela A é um visualizador de texto que exibe o texto a ser anotado, enquanto as Janelas B e C são diretamente responsáveis pela anotação do conteúdo.

Quando o usuário abre um arquivo para anotação, o MulSEN realiza um pré-processamento do texto antes de exibi-lo na Janela A. Um etiquetador morfossintático

[13] O Núcleo Interinstitucional de Linguística Computacional (NILC) é um dos grupos de pesquisa em PLN mais antigos do Brasil. Sediado no ICMC/USP/São Carlos, ele é formado por linguistas e cientistas da computação de diferentes instituições. Este trabalho, aliás, vem sendo conduzido no âmbito do NILC.
[14] Disponível em http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/files/MulSEN.zip
[15] O referido editor foi desenvolvido em um trabalho colaborativo entre Fabrício Élder da Silva Tosta e Fernando A. A. Nóbrega, sob supervisão do Prof. Dr. Thiago A. S. Pardo, do ICMC/USP.

ou *tagger* identifica automaticamente os nomes comuns, marcando-os com uma borda vermelha. O MulSEN aplica o etiquetador MXPOST (RATNAPARKHI, 1986) para os textos-fonte em português e o *TreeTagger* (SHIMID, 1994) para os textos em inglês.

Os dados resultantes do processo de *tagging* são utilizados por um módulo de desambiguação lexical de sentido (em inglês, *word sense disambiguation* ou DLS) para determinar o conceito subjacente a um nome mediante o contexto (sentença, texto, documento etc.) e um repositório de conceitos (AGIRRE; EDMONDS, 2006). Desse modo, a premissa da ferramenta é a de que a identificação dos nomes e a determinação do *synset* mais adequado para cada nome sejam automáticas, cabendo ao usuário confirmar as decisões tomadas pelo sistema.

Após essas duas tarefas de pré-processamento, o MulSEN exibe o texto como ilustrado pela Figura 4. Nessa figura, tem-se parte da tela do MulSEN, com destaque para o visualizador de texto. Note-se que a imagem apresenta algumas palavras marcadas com fundo em amarelo. Esses são os nomes comuns que já passaram pelas etapas de etiquetação morfossintática e desambiguação lexical de sentido, o que significa que a ferramenta tem uma sugestão de *synset* a ser assinalado. Além disso, há palavras marcadas apenas com uma borda vermelha, que são os termos identificados pelo *tagger* como nomes comuns (em ambos os casos da imagem, houve um erro de etiquetação automática), mas que não retornaram quaisquer *synsets*.

Figura 4 – Visualizador de texto no MulSEN

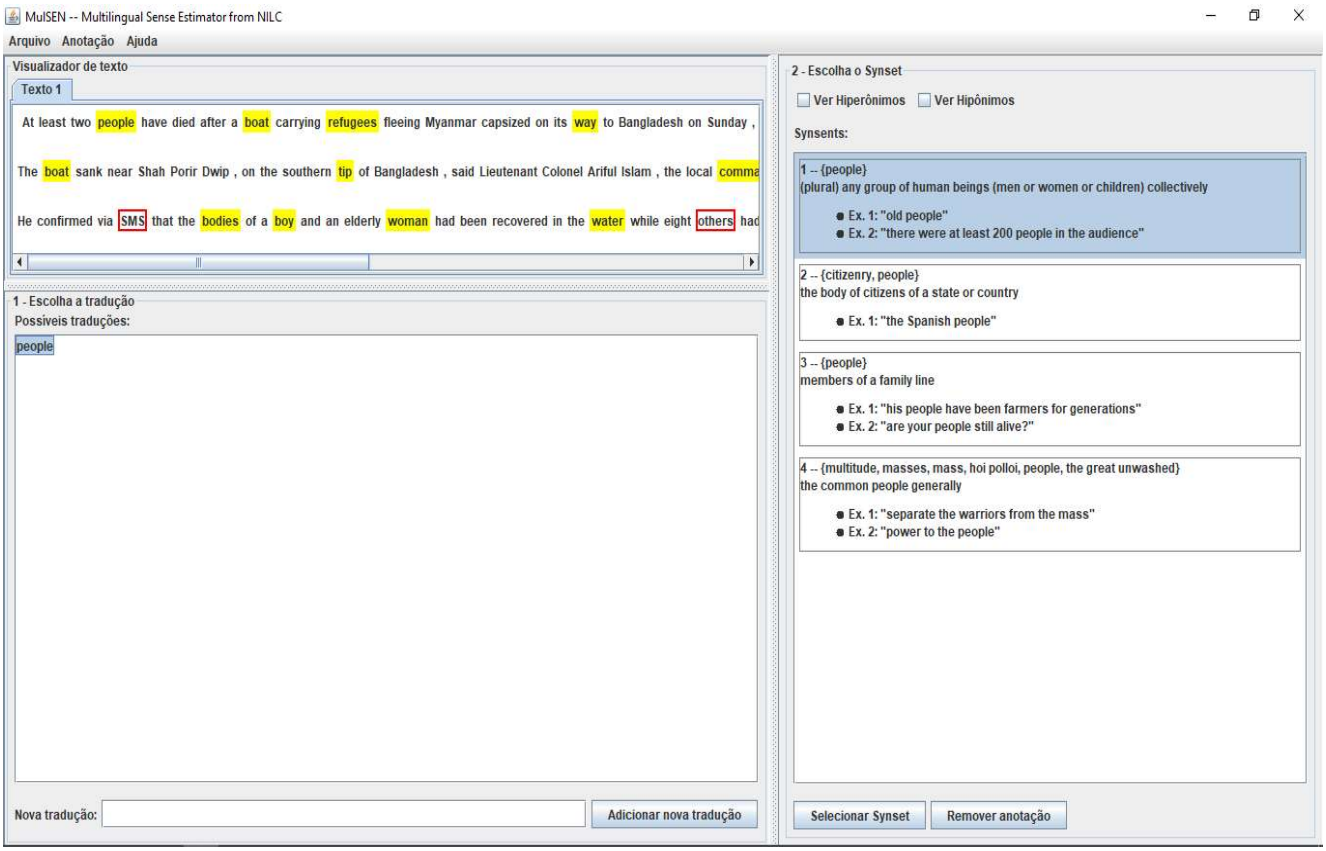
Ao clicar sobre um dos nomes, o usuário da ferramenta deverá ver, na Janela B, uma lista de traduções em língua inglesa da palavra, caso o texto esteja em outro idioma[16]. Se o texto em anotação estiver em inglês, o próprio vocábulo selecionado aparecerá

[16] O MulSEN foi desenvolvido para a anotação de textos em inglês, português e espanhol. As possíveis traduções listadas pela ferramenta vêm do dicionário *on-line* WordReference (Disponível em: http://www.wordreference.com/).

nessa janela. Além disso, o usuário também tem a possibilidade de inserir manualmente a tradução de qualquer palavra, caso não considere as sugestões adequadas ou caso o MulSEN não encontre traduções para a palavra.

Para exemplificar o funcionamento do MulSEN, considera-se o primeiro nome comum do texto da Figura 5 (*people*).

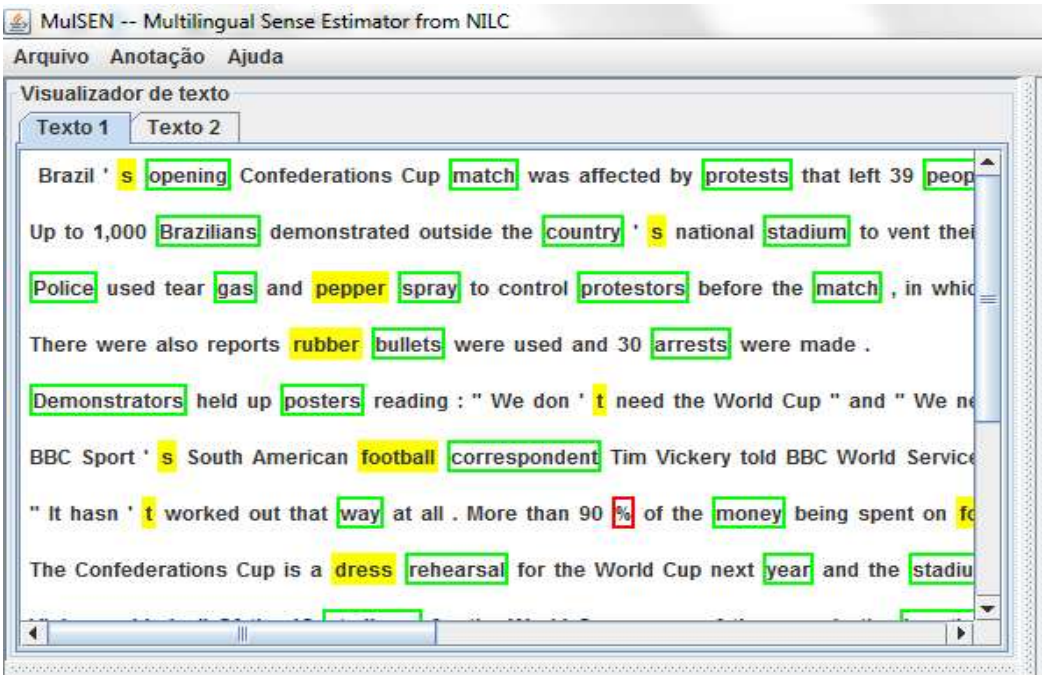
Figura 5 – Seleção da palavra, tradução e recuperação do *synsets*.



Por ser um texto em inglês, observa-se que o próprio item do texto foi utilizado como "tradução" em inglês (Janela B) para que o editor recuperasse todos os *synsets* da WN.Pr constituídos por *people,* os quais são exibidos na Janela C. No caso, o módulo de DLS do editor vinculou automaticamente o nome *people* ao primeiro *synset* da lista, que está destacada em azul. Se o anotador estiver de acordo com a anotação prévia do MulSEN, deverá apenas clicar no botão "Selecionar *Synset*" e confirmar. Caso contrário, deverá selecionar o *synset* que considerar adequado antes de clicar nesse botão. Ainda, se o usuário entender que a palavra foi incorretamente etiquetada como nome e que, portanto, não deve passar por anotação, basta clicar em "Remover anotação".

Essa janela ainda conta com um recurso que permite a visualização de hiperônimos e hipônimos dos *synsets*, o que pode ser relevante quando, por exemplo, a WN.Pr não contiver um *synset* que codifica o conceito específico expresso pelo nome no texto. Nesse caso, a seleção de um hiperônimo permite que a anotação não perca por total o conceito que ocorreu no texto. Com a confirmação do *synset* que representa o conceito subjacente ao nome, encerra-se a anotação deste, que passará a ser marcado com uma borda verde, e o usuário poderá prosseguir à anotação do próximo nome do texto. Na Figura 6, exibe-se um texto em inglês após sua anotação no editor MulSEN.

Figura 6 – Exibição do texto-fonte em inglês após anotação léxico-conceitual



Observa-se que todos os nomes devidamente anotados estão com a borda verde e que outros nomes (como *pepper* e *rubber*) permanecem em amarelo, indicando que suas anotações não foram confirmadas ou descartadas. Os motivos pelos quais tais palavras não foram incluídas na anotação serão discutidos na próxima Seção. Ao final, o MulSEN gera um arquivo no formato XML (do inglês, *Extensible Markup Language*) (Quadro 1), no qual é possível visualizar os nomes anotados e os respectivos equivalentes de tradução que permitiram recuperar os *synsets* da WN.Pr, além dos próprios *synsets* selecionados e seus códigos identificadores.

Quadro 1 – Trecho de arquivo XML gerado pelo MulSEN

O vulcão<volcano,Noun@9470550[vent, volcano]> **Grimsvotn, o mais ativo da Islândia, registrou no sábado**<saturday,Noun@15164570[Saturday,Sabbatum,Sat]> **passado a erupção**<eruption,Noun@7436475[volcanic eruption,eruption]> **inicial mais violenta dos últimos cem anos**<year,Noun@15203791[year,twelvemonth,yr]>, **provocando uma imensa nuvem**<cloud,Noun@11439690[cloud]> **de cinzas**<ash,Noun@14769160[ash]>**.**

3.4.2 As regras de anotação de Tosta (2014)

O *corpus* CM2News, que deu origem ao CM3News, fora anotado em nível léxico-conceitual segundo um manual de regras gerais e específicas propostas por Tosta (2014). Para a anotação do CM3News, buscou-se empregar as mesmas diretrizes.

Segundo o autor, as regras gerais são: (i) ler cuidadosamente os textos-fonte de cada coleção, (ii) iniciar a anotação preferencialmente pelo texto em inglês da coleção, posto que esse texto pode fornecer os equivalentes de tradução para a anotação dos textos nas demais línguas, (iii) anotar todos os nomes comuns e siglas do *corpus,* pois se entende que estes carregam boa parte da carga semântica de um texto jornalístico, (iv) refinar a anotação morfossintática automática, visto que os etiquetadores não são completamente precisos, (v) ignorar palavras anotadas equivocadamente como nome e (vi) selecionar o mesmo *synset* para anotar diferentes expressões linguísticas do mesmo conceito na coleção.

A regra geral (vi), em especial, busca 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.

Tosta (2014) previu ainda um conjunto de cinco diretrizes específicas. A primeira estabelece que, quando da anotação de expressões multipalavras[17], apenas o núcleo nominal será anotado com um *synset* que codifica o conceito da expressão como um todo. Isso se deve ao fato de que os *taggers* não detectam expressões (mas apenas unigramas), não permitindo que o MulSEN associe uma sequência de unigramas a um único *synset*. Para exemplificar, o autor cita *gás de pimenta* e *pepper spray,* cujo conceito, expresso pela glosa "*a nonlethal aerosol spray made with the pepper derivative oleoresin capsicum; used to cause temporary blindness and incapacitate an*

[17] As expressões multipalavras são sequências de palavras que apresentam idiossincrasias lexicais, sintáticas, semânticas, pragmáticas ou estatísticas e incluem, por exemplo, expressões compostas, como "carro de polícia" e "bode expiatório" (VILLAVICENCIO *et al.*, 2010).

attacker"[18], é codificado pelo *synset* específico {*pepper spray*}. Nesse caso, apenas os núcleos *gás* e *spray* foram associados a {*pepper spray*}.

A segunda regra estabelece que todos os nomes constitutivos de um sintagma recorrente livre[19] (SRL) devem ser anotados com seus respectivos *synsets.* Para ilustrar, Tosta (2014) cita *foco da dengue*. No caso, o nome *foco* foi anotado com o *synset* {*beginning, origin, root, rootage, source*}*,* que codifica o conceito "*the place where something begins, where it springs into being"*[20], e o nome *dengue* foi anotado com o *synset* {*dengue, dengue fever, dandy fever, breakbone fever*}, que expressa em inglês o conceito "*an infectious disease of the tropics transmitted by mosquitoes and characterized by rash and aching head and joints*"[21]. Dessa forma, essa regra busca permitir que a anotação capture o conceito representado pelo sintagma completo.

As expressões multipalavras e os SRLs são comumente compostas por mais de um item lexical, podendo englobar, aliás, mais de um nome. Diferenciá-los nem sempre é uma tarefa simples, pois requer uma análise da cristalização da expressão na língua. Uma estratégia produtiva é a observação de sua lexicalização em dicionários, por exemplo. No caso das línguas focalizadas neste estudo, diversos dicionários registram, na forma de entrada ou subentrada, uma definição específica para *gás de pimenta* ou algum equivalente próximo, como *spray de pimenta* ou a forma mais genérica *gás lacrimogêneo*, o que não costuma ocorrer com as expressões identificadas como SRLs.

A terceira regra determina que os anotadores devem analisar todas as possíveis traduções fornecidas pelo MulSEN, bem como seus respectivos *synsets*, antes de concluir o processo. Isso é importante porque a tradução adequada pode não ser a primeira na lista de opções apresentada pelo editor.

A quarta regra se aplica a casos nos quais as traduções precisam ser inseridas manualmente no editor, pois ele (i) não encontrou uma tradução no WordReference ou (ii) não forneceu uma tradução apropriada na lista de sugestões. Essa regra estabelece que, para inserir uma tradução, o anotador deve testar todos os possíveis equivalentes encontrados em outros recursos antes de incluir o mais adequado no MulSEN.

---

[18] "Um *spray* de aerossol não-letal, feito com a oleorresina Capsicum derivada da pimenta; usado para causar cegueira temporária e incapacitar um agressor" (tradução nossa).
[19] Os sintagmas recorrentes livres são combinações de palavras que, embora frequentes, apresentam níveis baixos de estabilidade e fixação (BENTIVOGLI; PIANTA, 2003).
[20] "O local onde algo começa, onde passa a existir" (tradução nossa).
[21] "Doença tropical infecciosa que é transmitida por mosquitos e se caracteriza por erupções cutâneas e dores de cabeça e nas articulações" (tradução nossa).

A quinta regra determina que, se não houver um *synset* adequado para codificar o conceito subjacente a um nome, deve-se selecionar um mais genérico. Isso significa que, se nenhum dos *synsets* exibidos com base na tradução escolhida for adequado, os anotadores devem procurar um *synset* hiperônimo satisfatório.

Na sequência, discorre-se sobre como o MulSEN e as regras de anotação foram efetivamente aplicadas na anotação do CM3News.

3.4.3 A anotação do CM3News via MulSEN e diretrizes do CM2News

A anotação do CM3News foi realizada por um único linguista em sessões diárias de aproximadamente 40 minutos, no período de cinco semanas.

Quanto ao MulSEN, destaca-se que as funcionalidades do editor voltadas para a anotação de textos em português e inglês descritas na Seção 3.4.1 foram aplicadas à anotação dos textos nessas línguas que compõem a coleção C21. No entanto, parte das funcionalidades não pode ser aplicada à anotação dos textos em alemão. Embora o editor pudesse ser alterado para contemplar a língua alemã, não foi possível que as alterações necessárias fossem feitas pelo desenvolvedor do editor no período em que a anotação em questão estava em curso. Assim, alguns dos recursos básicos desse editor gráfico não puderam ser utilizados de forma automática, o que tornou a anotação mais demorada. Porém, mesmo com esses entraves, decidiu-se pelo uso do MulSEN para que se mantivesse a padronização das anotações já realizadas nas outras línguas.

Para a anotação das notícias jornalistas em alemão via MulSEN, foi necessário realizar um "pré-processamento" semiautomático dos textos-fonte, que consistiu na substituição das letras *ä*, *ö*, *ü* e *ß* por *ae*, *oe*, *ue* e *ss*, respectivamente, as quais seguem o padrão ortográfico do alemão quando redigido em dispositivos que não oferecem suporte às letras especiais. Essa substituição foi necessária porque o editor, ao não reconhecer as letras *ä*, *ö*, *ü* e *ß* pertencentes ao alfabeto alemão, as substituía por pontos de interrogação entre espaços (p.ex., *Ern ? hrung* ao invés de *Ernährung*), gerando muito ruído nos textos. Uma vez que essas substituições foram feitas nos arquivos *txt* dos textos-fonte, estes foram submetidos individualmente ao editor.

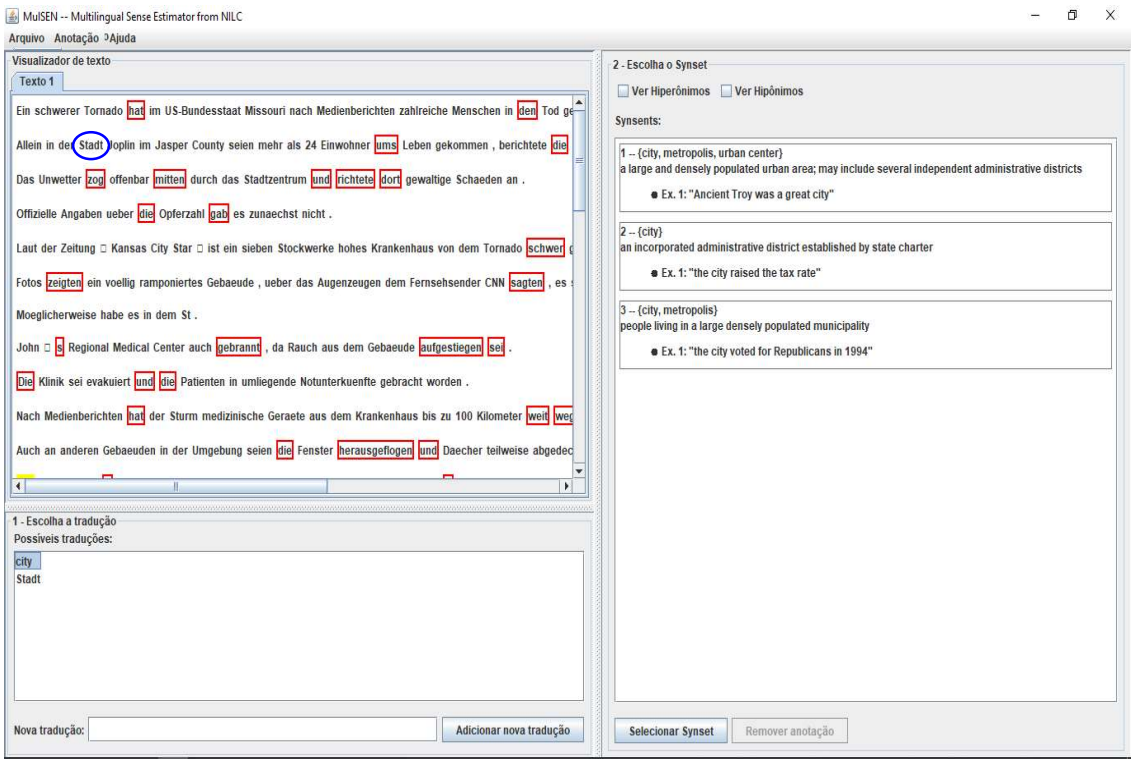
Como os *taggers* do MulSEN são dependentes das línguas que compõem o CM2News (PT e EN), a detecção automática dos nomes em alemão não foi realizada ou foi realizada de forma equivocada. Quando um nome não era identificado pelo editor, o anotador humano clicava sobre esse nome quando o texto estivesse sendo exibido no

painel “Visualizador de texto” e prosseguia com a sua anotação. A identificação errônea de muitas palavras como nome pelo editor ocorreu porque estas também integram o vocabulário do inglês. O artigo definido em alemão *die*, por exemplo, por corresponder graficamente a um nome[22] do inglês, fora assinalado como nome. Todos os elementos equivocadamente etiquetados como nome foram ignorados, não sendo, portanto, anotados em nível léxico-conceitual.

Além da anotação morfossintática, a TA para o inglês e a DLS também não ocorreram para os textos em alemão, sendo realizadas manualmente pelo anotador humano, ainda que por meio da interface do editor.

Com base na Figura 7, ilustra-se a anotação conceitual, via MulSEN e diretrizes gerais de Tosta (2014), de um nome (*Stadt*), que está circulado, do texto em alemão de C9.

Figura 7 – Ilustração da anotação conceitual nos nomes das notícias em alemão



Seguindo Tosta (2014), os textos-fonte em inglês e português (de C9) e suas respectivas anotações léxico-conceituais foram lidos antes da anotação do texto em alemão. Assim, para a anotação do nome *Stadt,* verificou-se que “*city*” e “cidade”, em inglês e

[22] Em português, o equivalente é *dado,* isto é, “cubo com números ou símbolos em suas faces”.

português, respectivamente, já haviam sido anotadas com o *synset* {*city, metropolis, urban center*}, que é especificado pela glosa "*a large and densely populated urban area; may include several independent administrative districts*"[23]. Quando da anotação de *Stadt,* o equivalente de tradução "*city*" (proveniente do texto em inglês) fora digitado no campo "Nova tradução" para que o editor recuperasse todos os *synsets* da WN.Pr que contivessem "*city*", inclusive {*city, metropolis, urban center*}, o qual seria selecionado para a anotação. Após confirmar a anotação*,* todas as ocorrências de *Stadt* no texto foram automaticamente associadas ao *synset* selecionado.

Nem todos os nomes em alemão, no entanto, têm uma composição morfológica simples como *Stadt*. No tocante à formação dos nomes, o alemão se caracteriza pela flexibilidade em recorrer tanto à derivação quanto à composição (ROMÃO, 2018). No primeiro caso, tem-se a junção de um termo autônomo (palavra-base) a um termo não-autônomo (normalmente um afixo), como é o caso de *Freiheit* ("liberdade"), formado pelo adjetivo *frei* ("livre") e pelo sufixo *-heit*, utilizado para transformar adjetivos em nomes. Na composição, duas palavras autônomas são justapostas ou, eventualmente, ligadas por um interfixo. Nesse processo, o primeiro termo é designado como palavra determinativa e o segundo como palavra-base.

De acordo com a classe gramatical da palavra determinativa, tem-se os seguintes padrões de composição para nomes:

- nome-nome: *Apfelbaum* ("macieira") → *Apfel* ("maçã") + *Baum* ("árvore");
- adjetivo-nome: *Nacktschnecke* ("lesma") → *nackt* ("nu") + *Schnecke* ("caramujo");
- verbo-nome: *Schreibtisch* ("escrivaninha") → *schreiben* ("escrever") + *Tisch* ("mesa");
- advérbio-nome: *Innenpolitik* ("política interna") → *innen* ("dentro, dentro de") + *Politik* ("política").

Diante disso, os nomes em alemão acabam por expressar conceitos bastante específicos que, quando não referidos por um único *synset,* poderiam teoricamente ser representados pela composição de *synsets.* Como o MulSEN permite associar apenas um *synset* a um unigrama ou um *token* (sequência de caracteres separada por espaços

[23] "Uma área urbana grande e densamente povoada; pode incluir vários distritos administrativos independentes" (tradução nossa).

em branco), priorizou-se, como Tosta (2014), a seleção dos *synsets* que expressam o sentido global (e específico) dos nomes compostos. Esse foi o caso, por exemplo, de *Stadtzentrum* ("centro da cidade"), anotado com o *synset* {*business district, downtown*}, cuja glosa é "*the central area or commercial center of a town or city*"[24].

No entanto, para vários nomes compostos, a WN.Pr não dispunha de *synsets* que representassem tais conceitos específicos. Esse é o caso, por exemplo, de *Opferzahl* ("número de vítimas"), cuja palavra-base *Zahl* ("número") é antecedida pela palavra determinativa *Opfer* ("vítima"). Diante da inexistência de um *synset* específico para esse conceito e a impossibilidade de selecionar dois *synsets*, o *token Opferzahl* foi anotado com um *synset* que representa um conceito mais genérico, no caso, o conceito subjacente à palavra-base *Zahl* ({*number, figure*}).

Pelo fato de o trabalho de pesquisa envolver a SAMM nas línguas portuguesa, inglesa e alemã, optou-se pela WN.Pr por ser a mais ampla. No entanto, a língua alemã tem um repositório lexical específico, desenvolvido na Eberhard Karls Universität, em Tübingen (Alemanha). A GermaNet[25] conta com mais de 136 mil *synsets* na versão lançada em 2019 e leva em consideração as expressões multipalavra e a natureza profundamente composicional do idioma. Quanto a essa primeira categoria, há *synsets* disponíveis para cobrir tais expressões quando elas apresentam um grau percebido de cristalização: quando são usadas com frequência em conjunto e quando atuam como unidades lexicais devido a uma forte relação entre suas partes. Os compostos nominais, por sua vez, são separados em suas partes constituintes de forma automática e, posteriormente, revisados e enriquecidos com outras informações relevantes.

No Quadro 2, exemplificam-se compostos lexicalizados, cujos conceitos específicos foram anotados com seus respectivos *synsets*, e compostos não-lexicalizados, os quais foram anotados com *synsets* que codificam o conceito subjacente às palavras-base (em negrito). Todos os exemplos pertencem à C9 do CM3News.

Além do fato de que nem sempre foi possível anotar o conceito específico, sendo estes anotados de forma mais genérica, alguns nomes não foram de fato anotados. Pode-se dividir tais ocorrências em dois grupos:

---

[24] "A área central ou o centro comercial de uma cidade" (tradução nossa).
[25] Disponível em: http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml.

- Nomes pertencentes a sintagmas não-nominais: esses são os casos, por exemplo, de *Leben* (“vida”) no sintagma verbal *ums Leben kommen* (“morrer”) e no sintagma adjetival *ums Leben gekommen* (“morto”), *Zuge* no sintagma preposicional *im Zuge* (“durante, no decorrer de”) e *Bezug* no sintagma preposicional *mit Bezug auf* (“acerca de, a respeito de”);
- Nomes específicos sem *synsets* correspondentes ou mesmo hiperônimos adequados: esse é o caso, por exemplo, de *Islamisierung* (“islamização”) na coleção C20, para o qual não se identificou um *synset* que minimamente refletisse esse conceito.

Quadro 2 – Exemplos de anotação dos nomes em alemão do CM3News

Compostos lexicalizados		
Nome	**Tradução (em português)**	***Synset***
Stadtzentrum	centro da cidade	*{business district, downtown}*
Fernsehsender	emissora de televisão	*{television station, TV station}*
Telefoninterview	entrevista por telefone	*{telephone interview}*
Hauptstraßen	ruas principais	*{highway, main road}*
Ortszeit	hora/horário local	*{civil time, standard time, local time}*
US-Präsident	presidente dos EUA	*{President of the United States, United States President, President, Chief Executive}*
Compostos não-lexicalizados		
Nome	**Tradução (em português)**	***Synset***
Medien**bericht**	relato/cobertura da mídia	*{report, news report, story, account, write up}*
Online**ausgabe**	edição on-line	*{edition}*
Opfer**zahl**	número de vítimas	*{number, figure}*
Not**unterkünfte**	abrigos de emergência	*{shelter}*
Lokal**zeitung**	jornal local	*{newspaper, paper}*
Strom- und Telefon**netz**	rede de energia e telefone	*{network}*
Notfall-**Zentrum**	centro de emergência	*{center, centre}*
Unwetter**system**	sistema de tempestade	*{system}*
Not**stand**	estado de emergência	*{state}*
Europa**reise**	viagem pela Europa	*{trip}*

Fonte: Elaborado pelo autor.

Nas Tabelas 10 e 11, apresentam-se os dados finais da anotação dos nomes dos textos em alemão para cada uma das 20 coleções do CM3News.

Na Tabela 10, tem-se, para cada coleção: (i) o total de nomes, (ii) o total de nomes simples, (iii) o total de nomes compostos e (iv) o total de nomes não anotados.

Na Tabela 11, por sua vez, detalham-se os nomes compostos anotados dos textos-fonte em alemão, destacando que estes são lexicalizados (isto é, se foram anotados com um *synset* específico que representa o conceito em questão) ou se não são lexicalizados (isto é, se foram anotados com um *synset* mais genérico que representa o conceito subjacente apenas à palavra-base).

Os dados da Tabela 11 evidenciam que 17,2% dos nomes presentes nos 20 textos em alemão do CM3News são compostos. Desse universo, retratado na Tabela 11, 25,5% foram anotados com *synsets* específicos e 74,5% dos nomes compostos (isto é, 12,8% do total de nomes em alemão do *corpus*) foram anotados com *synsets* mais genéricos.

A despeito dos entraves que se apresentaram, 85,6%[26] dos nomes do *corpus* em alemão foi anotado com seu sentido integral, evidenciando que a anotação conseguiu capturar conteúdo considerável veiculado pelos textos-fonte, o que é relevante para métodos de SAMM pautados em conhecimento léxico-conceitual.

Tabela 10 – Estatística da anotação conceitual dos nomes em alemão do *corpus*

Coleção	**Total de nomes**	**Nomes simples**		**Nomes compostos**		**Nomes não anotados**	
C1	211	160	75,8%	39	18,5%	12	5,7%
C2	114	96	84,2%	17	14,9%	1	0,9%
C3	160	128	80,0%	32	20,0%	0	0,0%
C4	40	30	75,0%	10	25,0%	0	0,0%
C5	211	178	84,4%	30	14,2%	3	1,4%
C6	180	159	88,3%	21	11,7%	0	0,0%
C7	58	50	86,2%	7	12,1%	1	1,7%
C8	91	71	78,0%	18	19,8%	2	2,2%
C9	113	95	84,1%	16	14,2%	2	1,8%
C10	185	138	74,6%	45	24,3%	2	1,1%
C11	123	101	82,1%	21	17,1%	1	0,8%
C12	110	88	80,0%	21	19,1%	1	0,9%

[26] Valor obtido pela soma de 81,2% (total de nomes simples) e 4,4% (compostos lexicalizados no conjunto de todos os nomes).

C13	108	85	78,7%	23	21,3%	0	0,0%
C15	124	112	90,3%	10	8,1%	2	1,6%
C16	107	91	85,0%	13	12,1%	3	2,8%
C17	155	132	85,2%	20	12,9%	3	1,9%
C18	67	58	86,6%	9	13,4%	0	0,0%
C19	78	56	71,8%	21	26,9%	1	1,3%
C20	90	70	77,8%	18	20,0%	2	2,2%
C21	163	122	74,8%	37	22,7%	4	2,5%
TOTAL	**2488**	**2020**	**81,2%**	**428**	**17,2%**	**40**	**1,6%**

Fonte: Elaborado pelo autor.

Tabela 11 – Estatística da anotação conceitual dos nomes compostos em alemão

Coleção	**Nomes compostos**	**Compostos lexicalizados**		**Compostos não lexicalizados**	
C1	39	10	25,6%	29	74,4%
C2	17	1	5,9%	16	94,1%
C3	32	0	0,0%	32	100,0%
C4	10	4	40,0%	6	60,0%
C5	30	8	26,7%	22	73,3%
C6	21	5	23,8%	16	76,2%
C7	7	2	28,6%	5	71,4%
C8	18	8	44,4%	10	55,6%
C9	16	6	37,5%	10	62,5%
C10	45	11	24,4%	34	75,6%
C11	21	4	19,0%	17	81,0%
C12	21	5	23,8%	16	76,2%
C13	23	7	30,4%	16	69,6%
C15	10	1	10,0%	9	90,0%
C16	13	6	46,2%	7	53,8%
C17	20	5	25,0%	15	75,0%
C18	9	4	44,4%	5	55,6%
C19	21	7	33,3%	14	66,7%
C20	18	6	33,3%	12	66,7%
C21	37	9	24,3%	28	75,7%
TOTAL	**428**	**109**	**25,5%**	**319**	**74,5%**

Fonte: Elaborado pelo autor.

A escolha deliberada pela anotação léxico-conceitual com o uso de ferramentas como a MulSEN e a WN.Pr se deve às intenções de investigar o paradigma linguístico na SAMM e verificar seu estado da arte. Outros métodos produtivos empregam diferentes

abordagens, a exemplo das *word embeddings*. Pesquisas como as realizadas por Mohd *et al.* (2020) e Tilahun *et al.* (2020) mostram as vantagens, em termos de qualidade do sumário, do uso da semântica como referencial teórico e da identificação de vetores tanto para produzir como para avaliar sumários automáticos.

Uma vez que o CM3News foi anotado, os métodos CF e CFUL foram aplicados às coleções trilíngues para a geração de extratos com diferentes taxas de compressão e os sumários de referência para algumas coleções do *corpus* foram gerados. Tais tarefas são descritas na próxima Seção.

CAPÍTULO 4 – Produção dos extratos automáticos e sumários de referência

4.1 Geração dos extratos pelos métodos CF e CFUL

De acordo com a arquitetura dos métodos CF e CFUL, a geração de um extrato multilíngue para cada uma das 20 coleções do CM3News foi feita em duas etapas: (i) pontuação e ranqueamento automáticos das sentenças em função da frequência de ocorrência de seus conceitos/*synsets* na respectiva coleção e (ii) seleção das sentenças de acordo com as taxas de compressão estipuladas e método em questão (CF e CFUL).

4.1.1 Pontuação e ranqueamento das sentenças

Cada arquivo gerado pelo MulSEN após a anotação conceitual dos textos-fonte de uma coleção foi automaticamente processado[27] para o cálculo da frequência dos conceitos (*synsets*) na coleção e posterior ranqueamento das sentenças em função da soma da frequência de seus conceitos constitutivos. O ranque de uma coleção Cx, resultante do processo descrito e exemplificado a seguir, foi utilizado por ambos os métodos, CF e CFUL.

Com o auxílio das bibliotecas *stringr*, *foreach*, *plyr* e *xlsx* em uma ferramenta baseada na linguagem R, a frequência dos conceitos em cada uma das 20 coleções foi calculada e o ranque das sentenças foi disposto em uma planilha XLSX. Inicialmente, utilizou-se a função *readLines* para que os arquivos do MulSEN pudessem ser lidos pela ferramenta e, na sequência, criou-se um vetor que unifica as anotações nos três idiomas. Vale ressaltar que esses procedimentos foram realizados individualmente para cada coleção trilíngue.

Usando o recurso *str_split*, segmentaram-se todas as sentenças dos textos, que agora compõem um vetor único. Então, utilizou-se o recurso *str_extract_all* para a extração dos códigos identificadores de cada sentença. Considerando que essas ferramentas se baseiam em padrões e expressões regulares, identificou-se um padrão para a detecção automática dos caracteres que importam para a produção dos *rankings*.

---

[27] O trabalho estatístico foi realizado pelo Núcleo de Estatística Aplicada (NEA) (http://cemeai.icmc.usp.br/NEA/) do ICMC/USP. Em um trabalho colaborativo, os graduandos do Bacharelado em Estatística (BEst) Débora Rissato e Vinicius Rozemwinkel, sob supervisão da Profa. Dra. Juliana Cobre, calcularam automaticamente a frequência dos conceitos e ranquearam as sentenças.

Depois disso, todos os códigos da coleção foram armazenados em um pequeno banco de dados para facilitar a soma das frequências conceituais, que ocorrerá em uma etapa posterior. A transformação dos códigos identificadores em *data frames* permite o uso da função *count*, que contabiliza as repetições de cada código no conjunto de dados. Com a frequência de cada código identificador já calculada, as sentenças segmentadas pela função *str_split* receberam as devidas pontuações, isto é, a soma da frequência de todos os conceitos contidos em cada sentença. As sentenças e suas respectivas pontuações foram, então, agrupadas em um arquivo XLSX.

De forma abstrata, o processo de pontuação e ranqueamento das sentenças-fonte de uma coleção pode ser exemplificado com base na sentença anotada em (5)[28], extraída do textos-fonte em português de C16.

Para a criação do ranque de C16[29], por exemplo, calculou-se de forma automática a frequência de cada *synset* nessa coleção. A frequência de ocorrência de um *synset* em uma coleção C equivale ao número de vezes que o *synset* em questão foi anotado na coleção. Em (5), observa-se, por exemplo, que a frequência do *synset* <@103140>, indexado a *lançamento*, é 1 em C16, o que significado que esse *synset* ocorreu 16 vezes na coleção. Observa-se também que a frequência do *synset* <@9818022>, indexado a "*astronautas"*, é 18 em C16, o que significa que esse *synset* ocorreu 18 vezes na coleção (isto é, no total de 2 textos-fonte).

A soma da frequência dos *synsets* de uma sentença S resulta na pontuação final dessa sentença, que representa a sua importância. No caso, a soma da frequência de todos os *synsets* da sentença em (5) resultou na pontuação 26.

(5) O **presidente**<10467179>(2) Xi Jinping supervisionou pessoalmente o **lançamento**<103140>(1) de **terça-feira**<15164105>(3), dirigindo-se aos **astronautas**<9818022>(18) para lhes desejar **sucesso**<7319103>(2) e dizendo-se "enormemente feliz" por estar presente.

Uma vez pontuadas, as sentenças são ranqueadas em ordem decrescente, ou seja, partindo da sentença de pontuação mais alta em direção à de pontuação mais baixa. Na

[28] Por uma questão de brevidade, os conceitos estão descritos apenas pelos seus números identificadores.
[29] Os textos-fonte da C16 estão no Apêndice A.

Tabela 12, apresenta-se o ranque obtido para todas as sentenças da C16. Nesse ranque, a sentença em (5) ocupa a 19ª posição dada a sua pontuação.

Tabela 12 – Ranque sentencial com base na frequência dos conceitos (C16)

Posição	Sentença	Pontuação
1ª	Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.	73
2ª	Der 48-Jährige ist 2005 bereits mit "Shenzhou 6" ins All geflogen und mit diesem Flug der älteste Astronaut Chinas im All.	55
3ª	Europe in particular has opened a dialogue that could eventually result in flight opportunities for its astronauts on the proposed Chinese space station.	53
4ª	Fast auf den Tag genau 50 Jahre nach dem ersten Flug einer Frau ins All ist mit Wang Yaping zum zweiten Mal eine chinesische Astronautin an Bord.	50
5ª	Em junho de 2012, a China realizou sua primeira manobra bem sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.	45
6ª	Wang is China's second female astronaut and she will beam the country's first lesson from space to students on Earth.	42
7ª	A quinta viagem tripulada da China ao espaço desde 2003 foi acompanhada pelas habituais manifestações de orgulho nacional e propaganda do Partido Comunista, incluindo crianças vestidas em trajes de minorias étnicas, acenando para os três astronautas no centro espacial.	42
8ª	Als Chinas erste Lehrerin im All wird die Astronautin Wang Yaping Themen wie Schwerelosigkeit, Oberflächenspannung von Flüssigkeiten sowie Gewicht und Masse erläutern.	41
9ª	15 Tage sollen die Taikonauten im All bleiben - solange wie noch kein chinesischer Raumfahrer zuvor.	38
10ª	China wäre dann die einzige Nation, die einen ständigen Aussenposten im All hätte, da die Internationale Raumstation Iss ausläuft.	36
11ª	The commander, Nie Haisheng, and his crew, Zhang Xiaoguang and Wang Yaping, plan to spend just under two weeks at the orbiting Tiangong space lab.	32
12ª	The Shenzhou-9 crew - which included China's first female astronaut, Liu Yang - hooked up with the module for nearly 10 days in June 2012.	32
13ª	"We are looking at possibilities to use this space station," the European Space Agency's human spaceflight director Thomas Reiter told the BBC last month.	31
14ª	Auf dem fünften bemannten Raumflug Chinas planen die Astronauten ein manuelles und ein automatisches Andockmanöver mit dem Raummodul "Tiangong 1", das seit September 2011 die Erde umkreist.	31
15ª	Die Experimente und Übungen gelten als wichtige Voraussetzung für den langen Marsch	28

	der jungen Raumfahrernation zum Bau einer Raumstation bis 2020.	
16ª	Three astronauts blasted away from the Jiuquan base in Inner Mongolia on a Long March 2 F rocket at 17:38 Beijing time (09:38 GMT).	27
17ª	Kommandeur des Fluges ist der erfahrene Astronaut Nie Haisheng.	26
18ª	Essa será a mais longa missão já feita por astronautas chineses.	26
19ª	O presidente Xi Jinping supervisionou pessoalmente o lançamento de terça-feira, dirigindo-se aos astronautas para lhes desejar sucesso e dizendo-se "enormemente feliz" por estar presente.	26
20ª	"Die Astronauten werden etwa zwölf Tage in dem "Himmelspalast" wohnen.	25
21ª	Die Reise der drei Taikonauten zum Raumlabor "Tiangong 1" (Himmelspalast), das die Erde in rund 335 Kilometern Höhe umkreist, dauert 40 Stunden.	25
21ª	Auch nahm sie im selben Jahr an Impfaktionen aus der Luft zum Abregnen von Regenwolken während der Olympischen Spiele in Peking teil.	25
23ª	China has launched its latest Shenzhou manned space mission.	23
24ª	Bei einem Treffen mit den Astronauten zuvor sagte der Präsident : "Sie machen das chinesische Volk stolz.	22
25ª	It is the latest step in China's plan to eventually put a permanently manned station above the Earth.	22
26ª	Als erste Frau war die heute 76-jährige Russin Valentina Tereschkowa am 16. Juni 1963 in den Weltraum gestartet.	21
27ª	Nie's team aims to stay a few days longer, and like the crew of Shenzhou-9 will practise both manual and automatic dockings during the mission.	21
28ª	Auch gebe es neue Nahrung für die Astronauten.	19
29ª	Seither sind schon mehr als 50 Frauen im All gewesen.	18
30ª	Sie sollen "neue Technologien zum Bau der Raumstation" sowie lebenserhaltende Systeme testen.	18
31ª	China setzt seinen langen Marsch zu einer eigenen Raumstation fort.	16
32ª	Wie der 46-jährige Zhang Xiaoguang ist Wang Yaping ein Neuling im All.	16
33ª	Beijing hopes to launch its fully-fledged station at the turn of the decade.	15
34ª	Auf einer Rakete vom Typ "Langer Marsch 2 F" hob die Mission "Shenzhou 10" am Dienstag um 17.38 Uhr Ortszeit (11.38 Uhr MESZ) vom Kosmodrom Jiuquan in der Inneren Mongolei ab.	15
35ª	Há, no entanto, quem critique tamanho gasto na exploração espacial por parte de um país ainda em desenvolvimento, confrontado por questões mais prementes - da segurança alimentar à poluição e aos incêndios em fábricas.	15
36ª	O programa espacial chinês avançou muito desde que Mao Tsé-tung, fundador do regime comunista em 1949, lamentou o fato de seu país não ser capaz nem mesmo de colocar uma batata em órbita.	14
37ª	"Vocês são o orgulho do povo chinês, e esta missão é ao mesmo tempo gloriosa e sagrada", disse Xi, segundo a imprensa estatal.	13
38ª	Earlier in the day, Chinese TV carried pictures of President Xi Jinping wishing the crew luck.	12
39ª	"It should take just over 40 hours to raise the craft's orbit to the operating altitude of	12

	Tiangong some 335 km (210 miles) above the planet's surface.	
40ª	Mission controllers clapped enthusiastically once the ship's solar panels had been deployed.	11
41ª	The crew's capsule was ejected from the upper-stage of the rocket about nine minutes after lift-off.	11
42ª	Vor dem Start sagte die 33-Jährige, der Flug sei die Erfüllung des "chinesische Traums" von einem starken und wohlhabenden China.	11
43ª	Bei dem Flug von "Shenzhou 10" sollen erstmals auch chinesische Mittel- und Grundschüler über Video unterrichtet werden.	10
44ª	This mission, the fifth manned venture by China and scheduled to be the longest, is designated Shenzhou-10.	9
45ª	A Shenzhou 10 foi lançada em uma base remota no deserto de Gobi, no extremo oeste chinês, às 17h38 (6h38 em Brasília), numa tarde quente e de céu claro, conforme imagens transmitidas pela TV estatal.	9
46ª	Auch baut das Land gegenwärtig ein Satellitennetz für ein unabhängiges, weltumspannendes Navigationssystem.	8
47ª	China's human spaceflight programme is conducted largely in isolation to the ISS partners.	8
48ª	"You have trained and prepared yourselves carefully and thoroughly, so I am confident in your completing the mission successfully.	8
49ª	It is expected to have a mass of about 60 tonnes and comprise a number of interlocking modules.	7
50ª	"Por que não gastam esse dinheiro resolvendo os verdadeiros problemas da China em vez de desperdiçá-lo desse jeito?", escreveu um usuário no Sina Weibo, espécie de Twitter chinês.	7
51ª	In diesem Jahr will China noch eine Sonde auf dem Mond landen.	6
52ª	No entanto, o avanço chinês nesse campo gera temores sobre uma corrida armamentista espacial.	5
53ª	"The way ahead is that we will likely see first an exchange of experiments.	5
54ª	Die Majorin ist eine erfahrene Pilotin und flog Einsätze nach dem Erdbeben 2008 mit 87'000 Toten in Sichuan.	5
55ª	But this could change in the next few years.	4
56ª	Chinas Staats- und Parteichef Xi Jinping verfolgte den erfolgreichen Start am Raumfahrtbahnhof.	4
57ª	"You have made Chinese people feel proud of ourselves," Xi told Nie and his colleagues.	4
57ª	It was launched in 2011 to provide a target to test rendezvous and docking technologies.	4
59ª	And there are now also a few colleagues at the European Astronaut Centre who have started Chinese language training.	4
60ª	"I wish you success and look forward to your triumphant return.	3
61ª	Like the International Space Station (ISS), it will have long-duration residents and be supplied by robotic freighters.	2
62ª	Tiangong-1 is the demonstrator.	1
63ª	Die Abfallverarbeitung sei verbessert worden.	1

64ª	A China ainda está distante de se equiparar a EUA e Rússia, superpotências espaciais estabelecidas.	1

Fonte: Elaborado pelo autor.

4.2 Seleção de conteúdo e construção dos extratos

Com o ranque elaborado automaticamente para cada coleção, teve início a etapa manual de produção dos extratos com base em cada um dos métodos de Tosta (2014) e em função da variação da taxa de compressão. Assim, para cada coleção trilíngue, produziram-se quatro extratos nos seguintes cenários: (i) método CF com 70% de compressão, (ii) método CF com 30% de compressão, (iii) método CFUL com 70% de compressão e (iv) método CFUL com 30% de compressão.

Para cada uma das coleções do CM3News, calcularam-se as referidas taxas de compressão, as quais foram utilizadas por ambos os métodos. Embora a taxa de compressão tenha sido calculada, em Tosta (2014), com base no maior texto-fonte da coleção seguindo as diretrizes da literatura, optou-se, neste trabalho, por utilizar a média do número (ou quantidade) de palavras da coleção.

Essa opção foi adotada devido à diferença de extensão entre os textos-fonte em algumas coleções (cf. Tabela 8). Se a quantidade de palavras do texto mais extenso fosse a base para o cálculo da taxa de compressão, haveria a possibilidade de que os extratos ficassem mais extensos que algum dos textos-fonte. Ao basear a compressão na média da quantidade de palavras da coleção, mitiga-se esse problema de alguma forma.

Para exemplificação, considera-se a coleção C16 do *corpus* CM3News, cujos textos-fonte têm, em média, 410 palavras. Tendo em vista o critério aqui adotado, 70% de taxa de compressão significa que os métodos CF e CFUL devem gerar extratos com 30% da média de palavras dos textos-fonte. No caso de C16, isso representa um extrato com 123 palavras. Para 30% de compressão, os métodos devem gerar extratos com 70% da média da coleção, isto é, 287 palavras.

Uma vez que as duas taxas de compressão distintas foram especificadas para cada coleção, procedeu-se à geração efetiva dos extratos por cada um dos métodos profundos de SAMM.

4.2.1 Geração dos extratos pelo método CF

Como mencionado, o método CF seleciona conteúdo com base exclusivamente no ranque das sentenças em função da frequência dos conceitos. O Quadro 3 apresenta o algoritmo do método CF para auxiliar na compressão da geração dos extratos.

Quadro 3 – Algoritmo do método CF.

Método CF	
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 3. Pontuar as sentenças em função da frequência de ocorrência dos conceitos/*synsets* na coleção 4. Ranquear as sentenças em função da pontuação dos conceitos 5. Selecionar a 1ª sentença do ranque e traduzir para o português (se necessário)[30] 6. Caso a taxa de compressão não tenha sido atingida: 6.a. Selecionar a próxima sentença do ranque 6.b.Traduzir a sentença selecionada para o português, caso necessário 6.c. Verificar a redundância da sentença em questão com a já selecionada 6.d. Eleger a sentença somente se não for redundante 7. Repetir o passo 6 até que a taxa de compressão seja atingida
Síntese	8. Justapor as sentenças na ordem em que foram selecionadas 9. Ordenar as sentenças pela posição de ocorrência nos textos-fonte[31].

Fonte: Tosta (2014).

As traduções das sentenças não são revisadas, a fim de garantir que o resultado espelhe de forma mais precisa o desempenho das ferramentas automáticas.

De acordo com o algoritmo, a geração do extrato com 70% de compressão, no caso, começa com a seleção da 1ª sentença do ranque (Tabela 12), que está em português e contém 43 palavras. Como a extensão de 123 palavras não foi atingida, selecionou-se a 2ª sentença do ranque, em alemão. Após a TA desta para o português, verificou-se que a sentença tem 24 palavras, totalizando 67 palavras.

Para evitar a redundância entre as sentenças, aplicou-se um fator de redundância pautado na sobreposição de conceitos (ou *synsets*) (*concept overlap*)[32]. Caso a

[30] Em caso de empate, a sentença com menor número de palavras aparece antes no ranque.

[31] Em caso de empate, segue-se a ordem das línguas: português > inglês > alemão. Tal ordem foi definida com base na observação de que, no momento, as traduções automáticas inglês-português apresentam menos erros que as traduções alemão-português.

sobreposição de conceitos entre uma sentença selecionada e outra candidata a compor o sumário fosse superior a um limiar determinado empiricamente (do inglês, *threshold*), a sentença candidata não era selecionada. Com base em uma análise manual nas coleções do *corpus*, definiu-se o limiar de 0.4 (ou 40%) para redundância. Para isso, realizou-se um teste simples e curto no qual sentenças do CM3News que fossem equivalentes e estivessem em diferentes idiomas foram comparadas no *script*. Desse modo, obteve-se uma média próxima a 40% em relação a essas sentenças com conteúdo similar. Na Tabela 13[33], tem-se, por exemplo, a sobreposição de conceitos entre as 13 primeiras sentenças do ranque de C16.

Tabela 13 – Sobreposição de *synsets* entre 13 sentenças de C16

Sentença /Posição	**1ª**	**2ª**	**3ª**	**4ª**	**5ª**	**6ª**	**7ª**	**8ª**	**9ª**	**10ª**	**11ª**	**12ª**	**13ª**
1ª	-	7%	14%	5%	4%	5%	6%	5%	16%	6%	5%	15%	6%
2ª		-	33%	42%	7%	20%	33%	20%	50%	14%	11%	16%	28%
3ª			-	20%	6%	7%	11%	7%	14%	11%	0%	12%	22%
4ª				-	5%	14%	20%	14%	25%	9%	7%	10%	18%
5ª					-	5%	6%	5%	7%	14%	12%	15%	0%
6ª						-	16%	12%	20%	7%	6%	8%	7%
7ª							-	16%	33%	25%	9%	12%	10%
8ª								-	20%	7%	6%	8%	7%
9ª									-	14%	11%	40%	12%
10ª										-	9%	0%	0%
11ª											-	10%	0%
12ª												-	11%
13ª													-

Fonte: Elaborado pelo autor.

[32] O cálculo da redundância foi feito automaticamente em colaboração com Raphael Rocha da Silva, mestrando do ICMC/USP e pesquisador do NILC. Para tanto, desenvolveu-se um *script* em *Python*.

[33] Nessa tabela, o sinal "-" indica que a sobreposição de conceitos/*synsets* é sempre calculada entre uma sentença e outra, diferente desta (isto é, que não seja ela mesma).

Com base na Tabela 13, a sobreposição entre 1ª sentença e a 2ª (traduzida para o português) do ranque é inferior a 40% (7%), o que permite incluir a 2ª sentença no extrato. Posto que o extrato em construção tem apenas 67 palavras, não atingindo ainda o tamanho desejado de 123 palavras, selecionou-se a 3ª sentença do ranque, a qual está em inglês e contém, pós-tradução, 22 palavras. Como a redundância entre essa sentença candidata e as duas já previamente selecionadas (1ª e 2ª) é inferior ao limiar estipulado (cf. Tabela 13), a 3ª sentença também é incluída no extrato, totalizando 89 palavras.

Na sequência, selecionou-se a 4ª sentença mais bem pontuada, que é proveniente do texto em alemão e que, traduzida para o português, contém 29 palavras. A aplicação do fator de redundância evidenciou que há sobreposição superior a 40% (no caso, 42%) (cf. Tabela 13) entre essa sentença candidata e a 2ª sentença já selecionada, o que a impediu de compor o extrato.

Assim, selecionou-se a 5ª sentença mais bem pontuada do ranque que, advinda do texto em português, contém 49 palavras. Como a 5ª sentença não apresenta redundância frente às demais selecionadas, essa sentença pode compor o extrato, o qual passa a ter 138 palavras, que é um valor superior ao tamanho desejado de 128 palavras. Nesses casos, aplicou-se um critério de parada para a seleção das sentenças, que está pautado na extensão parcial do extrato que mais se aproxima da taxa de compressão. Tendo em vista que a quantidade de 138 palavras está mais próxima do tamanho desejado de 128 palavras do que 89 (que é o tamanho do extrato sem a inserção de 5ª sentença), a 5ª sentença é incluída no extrato e a seleção de conteúdo termina.

Assim, baseado na taxa de compressão de 70%, o método CF selecionou as sentenças do Quadro 4 para compor o extrato, as quais estão dispostas na ordem em que foram selecionadas do ranque.

Quadro 4 – Seleção de conteúdo: CF com 70% de compressão (C16).

Sentenças selecionadas para o extrato	**Posição**
Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.	1ª
O piloto de 48 anos voou para o espaço em 2005 com "Shenzhou 6" e foi o astronauta mais velho no espaço neste vôo.	2ª
A Europa, em particular, abriu um diálogo que poderá resultar em oportunidades de voo para os astronautas da proposta estação espacial chinesa.	3ª

Em junho de 2012, a China realizou sua primeira manobra bem-sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.	5ª

Fonte: Elaborado pelo autor.

Por fim, gerou-se o extrato pela justaposição das sentenças selecionadas considerando a posição em que ocorrem nos textos-fonte. No caso, tem-se a seguinte ordem: 1ª sentença (S1_pt) > 5ª sentença (S3_pt) > 3ª sentença (S23_en) > 2ª sentença (S27_de). No Quadro 5, apresenta-se o extrato final produzido pelo método CF para a C16 com base em 70% de compressão. Nesse quadro, a posição e o texto de origem das sentenças estão descritos entre colchetes por questão didática.

Quadro 5 – Extrato da C16: método CF com 70% de compressão

[S1_pt] Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.
[S3_pt] Em junho de 2012, a China realizou sua primeira manobra bem sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.
[S23_en] A Europa, em particular, abriu um diálogo que poderá resultar em oportunidades de voo para os astronautas da proposta estação espacial chinesa.
[S27_de] O piloto de 48 anos voou para o espaço em 2005 com "Shenzhou 6" e foi o astronauta mais velho no espaço neste vôo.

Fonte: Elaborado pelo autor.

O mesmo procedimento foi aplicado para a geração do extrato com 30% de taxa de compressão, o que representa 287 palavras. No caso, o CF gerou, para C16, o extrato do Quadro 6, que tem 300 palavras e é composto pela seguinte ordem das: S1_pt > S3_pt > S3_en > S4_en > S7_pt > S16_en > S18_de > S22_de > S23_en > S27_de.

Quadro 6 – Extrato da C16: método CF com 30% de compressão

[S1_pt] Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.
[S3_pt] Em junho de 2012, a China realizou sua primeira manobra bem sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.
[S3_en] O comandante, Nie Haisheng, e sua equipe, Zhang Xiaoguang e Wang Yaping, planejam passar pouco

menos de duas semanas no laboratório espacial em órbita de Tiangong.
[S4_en] Wang é a segunda astronauta do sexo feminino da China e irá transmitir a primeira lição do país do espaço para os estudantes da Terra.
[S7_pt] A quinta viagem tripulada da China ao espaço desde 2003 foi acompanhada pelas habituais manifestações de orgulho nacional e propaganda do Partido Comunista, incluindo crianças vestidas em trajes de minorias étnicas, acenando para os três astronautas no centro espacial.
[S16_en] A tripulação da Shenzhou-9 - que incluiu o primeiro astronauta da China, Liu Yang - conectou-se ao módulo por quase 10 dias em junho de 2012.
[S18_de] A China seria então a única nação que teria um posto permanente no espaço quando a Estação Espacial Internacional Iss expirasse.
[S22_de] Como o primeiro instrutor espacial da China, o astronauta Wang Yaping irá explorar tópicos como ausência de peso, tensão superficial do fluido, peso e massa.
[S23_en] A Europa, em particular, abriu um diálogo que poderá resultar em oportunidades de voo para os astronautas da proposta estação espacial chinesa.
[S27_de] O piloto de 48 anos voou para o espaço em 2005 com "Shenzhou 6" e foi o astronauta mais velho no espaço neste vôo.

Fonte: Elaborado pelo autor.

4.2.2 Geração dos extratos pelo método CFUL

A partir do mesmo ranque utilizado pelo CF, o método CFUL realiza a seleção de conteúdo com base na língua do usuário. Especificamente, o CFUL apenas seleciona as sentenças em português mais bem pontuadas para compor o extrato até que a taxa de compressão seja atingida.

Quadro 7 – Algoritmo do método CFUL

Método CFUL	
Análise	1. Analisar cada um dos textos da coleção em nível léxico-conceitual, ou seja, anotar os nomes comuns com os conceitos/synsets da WordNet de Princeton.
Transformação	2. Calcular a taxa de compressão 3. Pontuar as sentenças em função da frequência de ocorrência dos *synsets*/conceitos na coleção 4. Ranquear as sentenças em função da pontuação dos conceitos 5. Selecionar a 1ª sentença do ranque que seja proveniente do texto em português[34] 6. Caso a taxa de compressão não tenha sido atingida: 6.a. Selecionar a próxima sentença em português do ranque 6.b. Verificar a redundância da sentença em questão com a já selecionada 6.c. Eleger a sentença somente se não for redundante 7. Repetir o passo 6 até que a taxa de compressão seja atingida

[34] Em caso de empate, a sentença com menor número de palavras aparece antes no ranque.

Síntese	8. Justapor as sentenças na ordem em que foram selecionadas 9. Ordenar as sentenças pela posição de ocorrência nos textos-fonte.

Fonte: Tosta (2014).

De acordo com o algoritmo, a geração do extrato com 70% de compressão teve início com a seleção da 1ª sentença do ranque descrito no Quadro 2, que está em português e contém 43 palavras. Como a extensão de 123 palavras não foi atingida, selecionou-se a próxima sentença do ranque em português. No caso, trata-se da 5ª sentença, que tem 49 palavras. Verificando a não redundância entre elas (cf. Tabela 13), esta foi incluída no extrato, o qual tem parcialmente 92 palavras. Como a taxa de compressão não foi atingida, selecionou-se a 7ª sentença, pois esta é a próxima mais bem pontuada em português. A 7ª sentença tem 39 palavras e, como a redundância entre ela e as já selecionadas (1ª e 5ª) está abaixo do *threshold,* a sentença em questão pode compor o extrato, o qual passa a ter 131 palavras. Conferindo o critério de parada ou truncamento, verificou-se que a extensão de 131 palavras é mais próxima do tamanho de desejado (123 palavras) do que 92 palavras. Assim, a 7ª sentença é efetivada a compor o extrato.

Ao final, o conjunto final de sentenças (provenientes exclusivamente do texto em português) selecionadas do ranque de C16 pelo método CFUL com base em 70% de taxa de compressão está descrito no Quadro 8.

Quadro 8 – Sentenças selecionadas da C16: método CFUL com 70% de compressão

Sentenças selecionadas para o extrato	**Posição**
Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.	1ª
Em junho de 2012, a China realizou sua primeira manobra bem-sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.	5ª
A quinta viagem tripulada da China ao espaço desde 2003 foi acompanhada pelas habituais manifestações de orgulho nacional e propaganda do Partido Comunista, incluindo crianças vestidas em trajes de minorias étnicas, acenando para os três astronautas no centro espacial.	7ª

Fonte: Elaborado pelo autor.

Com base na posição de ocorrência das sentenças no texto-fonte em português, tem-se a seguinte ordenação para o extrato (Quadro 9): 1ª sentença (S1_pt) > 5ª sentença (S3_pt) > 7ª sentença (S7_pt).

Quadro 9 – Extrato da C16: método CFUL com 70% de compressão

[S1_pt] Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.
[S3_pt] Em junho de 2012, a China realizou sua primeira manobra bem-sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.
[S7_pt] A quinta viagem tripulada da China ao espaço desde 2003 foi acompanhada pelas habituais manifestações de orgulho nacional e propaganda do Partido Comunista, incluindo crianças vestidas em trajes de minorias étnicas, acenando para os três astronautas no centro espacial.

Fonte: Elaborado pelo autor.

O mesmo procedimento foi aplicado para a geração do extrato com 30% de taxa de compressão, o que representa 287 palavras. No caso, o CFUL gerou, para C16, o extrato do Quadro 10, que tem 290 palavras e é composto pela seguinte ordem das: 1ª sentença (S1_pt) > 45ª sentença (S2_pt) > 5ª sentença (S3_pt) > 19ª sentença (S4_pt) > 37ª sentença (S5_pt) > 18ª sentença (S6_pt) > 7ª sentença (S7_pt) > 35ª sentença (S8_pt) > 36ª sentença (S10_pt).

Quadro 10 – Extrato da C16: método CFUL com 30% de compressão

[S1_pt] Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.
[S2_pt] A Shenzhou 10 foi lançada em uma base remota no deserto de Gobi, no extremo oeste chinês, às 17h38 (6h38 em Brasília), numa tarde quente e de céu claro, conforme imagens transmitidas pela TV estatal.
[S3_pt] Em junho de 2012, a China realizou sua primeira manobra bem-sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.
[S4_pt] O presidente Xi Jinping supervisionou pessoalmente o lançamento de terça-

feira, dirigindo-se aos astronautas para lhes desejar sucesso e dizendo-se “enormemente feliz” por estar presente.
[S5_pt] “Vocês são o orgulho do povo chinês, e esta missão é ao mesmo tempo gloriosa e sagrada”, disse Xi, segundo a imprensa estatal.
[S6_pt] Essa será a mais longa missão já feita por astronautas chineses.
[S7_pt] A quinta viagem tripulada da China ao espaço desde 2003 foi acompanhada pelas habituais manifestações de orgulho nacional e propaganda do Partido Comunista, incluindo crianças vestidas em trajes de minorias étnicas, acenando para os três astronautas no centro espacial.
[S8_pt] Há, no entanto, quem critique tamanho gasto na exploração espacial por parte de um país ainda em desenvolvimento, confrontado por questões mais prementes - da segurança alimentar à poluição e aos incêndios em fábricas.
[S10_pt] O programa espacial chinês avançou muito desde que Mao Tsé-tung, fundador do regime comunista em 1949, lamentou o fato de seu país não ser capaz nem mesmo de colocar uma batata em órbita.

Fonte: Elaborado pelo autor.

4.3 Produção dos sumários de referência

Para atingir os objetivos descritos neste trabalho, produziram-se sumários de referência para as vinte coleções do *corpus* CM3News.

Tendo em vista um *corpus* trilíngue como o CM3News, a confecção desses sumários requeria que os sumarizadores humanos idealmente fossem capazes de ler os 3 textos-fonte (português, inglês e alemão) de uma coleção e produzir um sumário abstrativo em português, já que esses sumários seriam utilizados, no caso, para avaliar os extratos em português produzidos pelos métodos CF e CFUL. Contudo, devido ao objetivo de investigar a influência da língua materna dos redatores na confecção dos sumários de referência, optou-se por permitir que cada um escrevesse os sumários em seu próprio idioma materno. Essa decisão também se pautou no fato de que, mesmo com um nível considerável de proficiência em português, um falante não-nativo dificilmente redigirá um sumário com a mesma naturalidade que teria ao escrever em seu próprio idioma. Para a posterior análise da informatividade via ROUGE, os sumários de referência em alemão foram automaticamente traduzidos via *Google Translator*.

Neste trabalho, os sumários de referência foram escritos por 5 sumarizadores humanos, sendo 3 falantes nativos do português e 2 do alemão. No âmbito desta

pesquisa, não foi possível contar com falantes nativos do inglês com proficiência de leitura/escrita em português e alemão.

Todos os participantes produziram os sumários remotamente e com base em um protocolo contendo os seguintes passos: (i) ler os 3 textos-fonte da coleção trilíngue, (ii) redigir um sumário abstrativo em sua língua materna com tamanho equivalente a 70% da média das palavras da coleção (isto é, 30% de taxa de compressão), e (iii) reduzir o sumário produzido em (ii) de forma a produzir outro com tamanho equivalente a 30% da média das palavras da coleção (isto é, 70% de compressão).

Para as tarefas descritas em (ii) e (iii), os sumarizadores humanos foram informados sobre as médias de palavras de suas respectivas coleções, assim como os valores referentes às taxas de 30% e 70% de compressão. Para os responsáveis pela coleção C16, por exemplo, forneceram-se as seguintes informações: (i) a média de palavras (dos textos-fonte) da coleção C16 é 410; (ii) 70% de taxa de compressão para C16 equivale a um sumário de 123 palavras (com desvio possível entre 120 e 130), e (iii) 30% de taxa de compressão para C16 equivale a um sumário com 287 palavras (com desvio possível entre 280 e 290).

No total, os 5 sumarizadores produziram 24 sumários de referência referentes a nove coleções do *corpus* CM3News, sendo metade composta por textos-fonte mais extensos e metade por textos menores. Na Tabela 14, tem-se a distribuição dos sumários de referência das nove coleções por taxa de compressão e língua materna do sumarizador.

Tabela 14 – Língua e taxa de compressão dos sumários de referência do CM3News

Coleção	**Língua/taxa de compressão do sumário de referência**			
	PT/30%	**PT/70%**	**DE/30%**	**DE/70%**
C2	1	1		
C4	1	1	1	1
C8			1	1
C9	1	1		
C15	1	1		
C16	1	1	1	1
C17	1	1	1	1
C18	1	1		
C19	1	1		

Fonte: Elaborado pelo autor.

Uma vez produzidos, os sumários de referência foram utilizados para:

i) investigar se a língua materna dos produtores dos sumários multilíngues de referência influencia a produção desses textos, a ponto de os mesmos conterem mais conteúdo proveniente do texto-fonte do respectivo idioma materno, sob a hipótese de que o melhor desempenho do CFUL se deve ao fato de que os sumários manuais do *corpus* CM2News foram produzidos exclusivamente por falantes nativos do português, contendo preferencialmente informações advindas dos textos-fonte nessa língua;
ii) avaliar o desempenho dos métodos quando da geração de sumários com taxas de compressão diferentes, sob a hipótese de que extratos menores gerados pelo método CF, por exemplo, podem apresentar poucos problemas de qualidade linguística, sobretudo aqueles que resultam da TA dos textos-fonte.

Na sequência, descreve-se o modo como os itens (i) e (ii) foram investigados. Ademais, discutem-se os resultados obtidos em ambas as investigações.

CAPÍTULO 5 – Explorando a avaliação em SAMM

5.1 A influência da taxa de compressão no desempenho dos métodos

Quanto à influência da variação da taxa de compressão no desempenho dos métodos de SAMM, realizou-se a avaliação da qualidade linguística e informatividade dos extratos (em português) com 30% e 70% de compressão gerados pelos métodos CF e CFUL.

5.1.1 A avaliação da qualidade linguística

Especificamente, a qualidade linguística dos 80 extratos automáticos (4 para cada uma das 20 coleções, sendo 2 do CF e 2 do CFUL) foi avaliada manualmente conforme os 5 critérios na DUC'05 (gramaticalidade, não-redundância, clareza referencial, foco e estrutura/coerência).

Ao longo de dois meses, treze participantes (onze dos quais com formação em Letras/Linguística) atribuíram notas de 1 a 5 aos extratos em função de cada um dos critérios da DUC'05. Inicialmente, os avaliadores receberam um manual que continha principalmente a descrição da tarefa e de cada um dos cinco critérios de análise da DUC'05. Ademais, para cada um dos critérios, o manual fornecia (i) um caso problemático do próprio CM3News que penalizaria o extrato caso este o contivesse e (ii) comentário ou explicação sobre o problema-exemplo, como ilustrado no Quadro 11.

Quadro 11 – Exemplos de problemas do CM3News que afetam a qualidade linguística

Critério	Sentenças-fonte com problema	Comentário
Gramaticalidade	Uma loja Hugo Boss e um mesa de câmbio em Sloane Square foram atacados entre segunda à noite e a madrugada de terça-feira, antes dos saqueadores alvo lojas em Pimlico Road, disse a polícia. (C1, S28_en)	Problemas de gramaticalidade se caracterizam, por exemplo, pelos desvios de concordância nominal (p.ex.: "um mesa") e outros como "saqueadores alvo lojas".
Não-redundância	Os números incluem dois rapazes de 17 anos e um homem de 18 anos preso por um incêndio criminoso que destruiu um depósito da Sony. (C1, S2_en) Dois dos três adolescentes presos em conexão com o fogo - um de 17 anos e um homem de 18 anos - permanecem sob custódia da polícia. (C1, S26_en)	As duas sentenças apresentam informações que se sobrepõem, como as idades e o crime cometido. Caso ocorram em um único texto, caracterizam problema de redundância.

Clareza referencial	Na semana passada, o mesmo Garotinho afirmou que a bancada evangélica, composta por 74 deputados, não votaria nada. (C2, S6_pt)	Caso a sentença-exemplo seja a 1ª do extrato, esta apresenta uma violação do tipo "1ª menção sem explicação", já que a entidade "Garotinho", embora popular para a maioria dos leitores, é inserida no discurso sem uma descrição adequada, o que é agravado pelo uso de "mesmo", como se esse nome já tivesse sido mencionado.
Foco temático	Eles expressaram sua solidariedade aos manifestantes do movimento Passe Livre, que luta desde a semana passada em São Paulo e no Rio de Janeiro contra o aumento nos preços de ingresso de ônibus, trem e metrô em 20 centavos por viagem única. (C17, S2_de) Mais de 90% do dinheiro gasto em estádios de futebol é dinheiro público. (C17, S8_en)	Embora as sentenças possam ter relação em um contexto mais ampliado (por exemplo, ao se ler as notícias na íntegra), a justaposição dessas sentenças representa uma mudança de foco muito brusca.
Estrutura/ Coerência	Entre outros pontos, o Código define a isenção da reserva legal para as propriedades de quatro módulos (20 a 400 hectares, dependendo do Estado), ponto que o governo é contra. (C5, S12_pt)	Caso a sentença em questão seja a 1ª do extrato, a estrutura ou coerência do texto fica prejudicada, pois o texto se inicia com a expressão "Entre outros pontos", que pressupõe a ocorrência de informações prévias.

Fonte: Elaborado pelo autor.

Os avaliadores receberam arquivos do Microsoft Word ou um *link* para formulários do *Google Forms* com os extratos, nos quais puderam dar as notas a cada critério. Cada avaliador recebeu quatro coleções distintas do *corpus*, totalizando 16 extratos, o que significa que cada extrato foi avaliado por duas ou três pessoas, conforme indica o Quadro 12.

Quadro 12 – Distribuição das coleções do CM3News pelos avaliadores

Coleções	**Avaliadores**
C1, C2, C3, C4	A1, A6, A9
C5, C6, C7, C8	A2, A10
C9, C10, C11, C12	A3, A7, A11
C13, C15, C16, C17	A4, A8, A12
C18, C19, C20, C21	A5, A13

Fonte: Elaborado pelo autor.

Antes de se investigar a influência da taxa de compressão no desempenho dos métodos, verificou-se o desempenho geral do CF e CFUL no CM3News, independentemente da compressão. Corroborando os resultados de Tosta (2014) quando da aplicação dos métodos ao CM2News (cf. Tabela 2), o CFUL foi superior ao CF (Tabela 15). Isso indica que um extrato formado exclusivamente por sentenças originais de um único texto-fonte (no caso, em português) tem qualidade linguística superior a de um extrato composto por sentenças originais (em português) e sentenças traduzidas (do alemão ou inglês para o português). Especificamente, a média obtida pelo CFUL foi superior à do CF em todos os critérios, sendo que essa superioridade variou de 7% a 11% (0,29 a 0,43 pontos). A maior diferença entre os eles reside no critério gramaticalidade, pois os extratos do CF podem ter sido prejudicados por problemas advindos da TA das sentenças em língua estrangeira (para o português) via *Google Tradutor*. Além disso, o sumário também é penalizado em tal critério quando há erros de formatação, os quais podem ou não ter sido gerados durante o processamento.

Tabela 15 – Avaliação da qualidade linguística dos extratos do CF e CFUL no CM3News

Critério	**Método**	
	CF	**CFUL**
Gramaticalidade	3,96	4,39
Não-redundância	3,95	4,32
Clareza referencial	3,74	4,14
Foco temático	4,09	4,38
Estrutura e coerência	3,85	4,20

Fonte: Elaborado pelo autor.

Na Tabela 16, tem-se as médias obtidas por cada um dos métodos em função das diferentes taxas de compressão dos extratos. Como esperado, o CFUL foi superior ao CF em ambos os cenários. Além disso, os resultados indicam que, independentemente do método, os extratos mais curtos (70% de compressão) obtiveram notas mais altas que os de maior extensão (30% de compressão) em todos os critérios. Aliás, os extratos do CFUL com 30% de compressão apresentam qualidade linguística igual ou inferior aos extratos do CF com 70% de compressão em todos os critérios. Isso pode ser um indicativo de que a taxa de compressão seja, de fato, um elemento que afeta o desempenho de um método de SAMM e, consequentemente, a qualidade linguística do

extrato multilíngue, pois um extrato mais extenso apresenta maior probabilidade de conter problemas linguísticos.

Tabela 16 – Resultado da avaliação da qualidade linguística em função da compressão

Critério	**Taxa de compressão/Método**			
	30% de compressão		**70% de compressão**	
	CF	**CFUL**	**CF**	**CFUL**
Gramaticalidade	3,72	4,20	4,20	4,59
Não-redundância	3,54	4,09	4,35	4,54
Clareza referencial	3,46	3,91	4,02	4,37
Foco temático	3,83	4,22	4,35	4,54
Estrutura e coerência	3,52	4,04	4,17	4,35

Fonte: Elaborado pelo autor.

Avaliou-se também a influência da taxa de compressão na informatividade dos extratos. Ressalta-se que essa investigação foi feita a partir das nove coleções da Tabela 12, o que englobou todos os 24 sumários de referência, inclusive os oito traduzidos do alemão para o português.

5.1.2 A avaliação da informatividade

A informatividade dos extratos dessas coleções foi avaliada via ROUGE. Por um lado, tais métricas podem parecer pouco produtivas em termos de conteúdo. Porém, a ROUGE se mostra útil no quesito forma, considerando-se todo o contexto já apresentado. Decidiu-se utilizá-la, neste momento, porque ela já foi empregada em vários estudos da área, incluindo os trabalhos de Tosta (2014) e Di-Felippo, Tosta e Pardo (2016), com o qual se compartilha parte do *corpus* utilizado, e porque ela pode dar indícios sobre as escolhas lexicais na sumarização humana. Se a comparação entre sumário de referência e extrato automático gerar resultados consideráveis, poder-se-á entender que os textos-fonte basearam fortemente o sumário abstrativo, sem que houvesse um alto grau de escrita espontânea, ou que, de fato, o extrato em questão sumariza bem o conteúdo dos textos-fonte.

Na Tabela 17, apresentam-se as médias obtidas via ROUGE-1 e 2 em um contexto de 48 comparações entre sumário de referência e extrato automático (12 de cada método/taxa de sumarização).

Tabela 17 – Resultado da avaliação automática da informatividade

	ROUGE-1			ROUGE-2		
	Cobertura	**Precisão**	**Medida-f**	**Cobertura**	**Precisão**	**Medida-f**
CF 30%	0,40431	0,39525	0,39919	0,11782	0,11325	0,11537
CFUL 30%	0,41747	0,41417	**0,41547**	0,14280	0,14060	**0,14162**
CF 70%	0,31888	0,34195	0,32878	0,09897	0,10293	0,10053
CFUL 70%	0,32925	0,33457	0,33007	0,11210	0,11176	0,11132

Fonte: Elaborado pelo autor.

Os resultados indicam que, considerando a mesma taxa de compressão (30% ou 70%), o método CFUL gera extratos mais informativos que o CF tanto na sobreposição de unigramas como na de bigramas. Por outro lado, independentemente do método empregado, os extratos maiores (com 30% de compressão) são mais informativos que os extratos menores (70%), uma vez que obtiveram medidas-f mais altas.

5.2 A influência da língua materna nos sumários de referência

A fim de investigar a influência da língua materna dos sumarizadores humanos no processo de seleção de conteúdo a compor um sumário de referência multilíngue, decidiu-se analisar a origem das informações contidas nos sumários de referência em português por meio de alinhamentos. Especificamente, realizou-se o alinhamento manual entre os sumários de referência e seus respectivos textos-fonte, de tal forma que esses alinhamentos evidenciassem a origem da informação constante dos sumários.

Mani (2001) menciona que, no contexto multidocumento, os sumarizadores humanos tendem a escolher um texto-fonte como a base para a produção de *abstracts*. Tal processo pode ser influenciado por elementos como a sequência cronológica dos textos, os autores dos materiais originais, o prestígio do veículo de comunicação e aspectos de textualidade. Nesta pesquisa, investigou-se se o fator “língua materna” também pode ser relevante no contexto multidocumento multilíngue.

5.2.1 O alinhamento dos sumários de referência e textos-fonte

Nos moldes de Camargo (2013), realizou-se o alinhamento entre todos os sumários de referência disponíveis nesta pesquisa e seus respectivos textos-fonte (nas diferentes línguas). Assim, as 239 sentenças que compõem os 24 sumários de referência foram alinhadas a seu(s) respectivo(s) texto(s)-fonte. Especificamente, a tarefa de alinhamento foi realizada por um único linguística computacional de forma manual. Seguindo Camargo (2013), os alinhamentos foram feitos em nível sentencial a partir da sobreposição de conteúdo entre sumários e textos-fonte. Essa sobreposição pode ser referente à informação principal (6) ou secundária (7) das sentenças, ignorando-se inconsistências numéricas ou de grau de generalização e de assertividade. Ademais, seguiu-se a diretriz de que as sentenças de um sumário de referência (SS) deviam ser alinhadas a todas as sentenças dos textos-fonte (ST) com as quais compartilhavam conteúdo, fossem elas pertencentes a um único texto ou a textos distintos (8).

(6) SS: Garotinho, vice-líder da Frente Parlamentar Evangélica, chegou mesmo a pedir a demissão do ministro da educação.
ST: Ontem, no plenário, o deputado Anthony Garotinho (PR-RJ) chegou a pedir a demissão do ministro da Educação, Fernando Haddad. (C2, S5_pt)

(7) SS: Laut Satellitenbildern der NASA sind ungefähr 30 Millionen Tonnen Eis des Gletschers in den gleichnamigen See abgerutscht und habe dabei Wellen von bis zu dreieinhalb Metern verursacht, berichten Augenzeugen. → De acordo com imagens de satélite da NASA, cerca de 30 milhões de toneladas de gelo da geleira se desprenderam no mar de mesmo nome e teriam causado ondas de até 3,5 metros, relatam testemunhas.
ST: Eine 1200 Meter lange und 75 Meter breite Eiszunge sei abgerutscht und in den See geplumpst, dreieinhalb Meter hohe Wellen wogten an die Ufer, berichten Augenzeugen. (C8, S3_de) → Uma língua de gelo com 1.200 metros de comprimento e 75 metros de largura teria se desprendido e caído no mar, causando ondas de 3,5 metros de altura na orla, relatam testemunhas.

(8) SS: Angelina Jolie revelou ter se submetido à retirada das mamas.
ST_pt: Jolie, 37, revelou ontem, em artigo publicado no jornal "New York Times", ter se submetido ao procedimento, que, de acordo com os médicos, tem sido cada vez mais procurado na rede privada. (C15, S2_pt)
ST_en: Hollywood actress Angelina Jolie has undergone a double mastectomy to reduce her chances of getting breast cancer. (C15, S1_en) → A atriz de Hollywood Angelina Jolie passou por uma mastectomia dupla para reduzir suas chances de contrair câncer de mama.
ST_de: Jetzt hat Jolie einen mutigen Schritt gewagt: in der „New York Times" beschreibt sie, wie sie sich aus Angst vor Brustkrebs beide Brüste hat amputieren lassen. (C15, S4_de) → Agora, Jolie deu um passo corajoso: no "New York Times", ela descreve como amputou ambas as mamas por medo do câncer de mama.

Para ilustrar, tem-se, nos Quadros 13 e 14, os alinhamentos identificados somente entre os sumários de referência (originais) em português de C4 (com 30% e 70% de compressão, respectivamente) e os respectivos textos nas diferentes línguas-fonte. A título de esclarecimento, destaca-se que, no Quadro 10, a SS1, por exemplo, foi alinhada às (i) sentenças ST1_pt, ST2_pt e ST9_pt do texto-fonte em português e (ii) sentenças ST1_de, ST4_de, ST5_de e ST6_de do texto-fonte em alemão. A SS1, no entanto, não possui nenhum alinhamento com o texto em inglês, indicando que o seu conteúdo não advém desse texto.

Quadro 13 – Alinhamento em C4: sumário (30% de compressão) e textos-fonte

Sentença/ Sumário	**Sentença/Texto-fonte**		
	Português	**Inglês**	**Alemão**
SS1	S1, S2, S9		S1, S4, S5, S6
SS2	S7	S19	
SS3	S9, S10, S11	S3	S10, S11
SS4	S9, S10, S11	S3	S10, S11
SS5		S25	
SS6	S15	S2, S12	
SS7		S9, S10, S11	
SS8		S1, S2, S15, S17, S21	

Fonte: Elaborado pelo autor.

Quadro 14 – Alinhamento em C4: sumário (70% de compressão) e textos-fonte

Sentença/ Sumário	**Sentença/Texto-fonte**		
	Português	**Inglês**	**Alemão**
SS1	S1, S2, S5, S9, S10, S11	S3, S13	S1, S4, S5, S6, S10, S11
SS2	S15	S2, S12	
SS3		S1, S2, S15, S17, S21	

Fonte: Elaborado pelo autor.

5.2.2 Um estudo de caso a partir dos alinhamentos da coleção C4

Para investigar, com base nos alinhamentos, a influência da língua materna dos redatores na seleção de conteúdo a compor os sumários de referência, realizou-se um estudo de caso a partir do alinhamento dos sumários (um com 30% e outro com 70% de compressão, redigidos por um falante nativo do português) e textos-fonte da coleção C4[35] (Quadros 10 e 11).

O estudo de caso consistiu em dois cálculos distintos sobre os alinhamentos, buscando identificar aquele que mais contribuía para a análise da influência do texto-fonte escrito originalmente em sua materna para a produção do sumário de referência.

O primeiro cálculo consistiu em determinar a quantidade de sentenças distintas de cada texto-fonte alinhadas aos sumários, buscando verificar se o texto-fonte escrito na língua materna do sumarizador humano teve predominância nos sumários.

Para exemplificar, destaca-se que o texto-fonte em português de C4 tem 15 sentenças distintas, sendo que, segundo o Quadro 13, apenas sete sentenças diferentes foram alinhadas ao sumário com 30% de compressão (a saber: ST1_pt, ST2_pt, ST9_pt, ST7_pt, ST10_pt, ST11_pt e ST15_pt). Isso significa dizer que aproximadamente 46% das sentenças do texto-fonte foram alinhadas ao sumário. Já o texto em inglês possui 28 sentenças distintas, sendo que doze delas foram alinhadas ao sumário com 30% de compressão (a saber: ST19_en, ST3_en, S2T5_en, ST2_en, ST12_en, ST9_en, ST10_en, ST11_en, ST1_en, ST15_en, ST17_en e ST21_en) (Quadro 13). Isso significa dizer que aproximadamente 43% das sentenças do texto-fonte foram alinhadas ao sumário.

Na Tabela 18, tem-se os valores (absolutos e porcentagens) resultantes do cálculo descrito anteriormente para os 2 sumários de referência de C4.

[35] Os textos-fonte e os sumários de referência da C4 estão no Apêndice B.

Tabela 18 – Quantidade de sentença dos textos-fonte alinhadas aos sumários (C4)

Texto-fonte	**Sumário de referência (original em português)**	
	30% de compressão	**70% de compressão**
Português	46% (7/15)	46% (7/15)
Inglês	43% (12/28)	29% (8/28)
Alemão	54% (6/11)	54% (6/11)

Fonte: Elaborado pelo autor.

Com base na Tabela 18, ambos os sumários têm menor quantidade de sentenças alinhadas ao texto-fonte em inglês (43% e 29%, respectivamente). No entanto, nos Quadros 13 e 14, observa-se que certas SSs têm conteúdo advindo exclusivamente dos textos-fonte em inglês. No sumário com 30% de compressão, por exemplo, SS5, SS7 e SS8 foram alinhadas somente ao texto em inglês. O mesmo ocorreu com SS3 do sumário com 70% de compressão.

O segundo cálculo do estudo de casos consistiu em dividir o número de sentenças de um sumário alinhadas a alguma sentença-fonte pelo número total de sentenças que compõem o sumário, buscando evidenciar o quanto de um sumário de referência é composto por conteúdo proveniente de cada texto-fonte. Os resultados da Tabela 19 indicam que o produtor dos sumários de referência, cuja língua materna no caso é o português, redigiu sentenças com conteúdo predominantemente advindo do texto-fonte em inglês, já que, (i) das 8 sentenças do sumário com 30% de compressão, 7 foram alinhadas a pelo menos uma sentença do texto em inglês (7/8=87%), e (ii) todas as 3 sentenças do sumário com 70% de compressão foram alinhadas ao texto em inglês (3/3=100%).

Tabela 19 – Quantidade de sentença dos sumários alinhadas a cada texto-fonte (C4)

Texto-fonte	**Sumário de referência (original em português)**	
	30% de compressão	**70% de compressão**
Português	62% (5/8)	66% (2/3)
Inglês	87% (7/8)	100% (3/3)
Alemão	37% (3/8)	33% (1/3)

Fonte: Elaborado pelo autor.

Diante do estudo de caso, optou-se por utilizar apenas o segundo cálculo, pois este parece ser mais útil para o foco do estudo em questão. Para que se possa verificar de forma mais direta se a língua materna do produtor do sumário de referência tem impacto sobre o texto redigido, realizou-se o cálculo da quantidade de sentenças (i) alinhadas ao texto-fonte na língua materna e (ii) alinhadas aos textos-fonte nas duas línguas estrangeiras, sem que houve distinção entre elas. Assim, para o sumário 1 da Tabela 20, por exemplo, fizeram-se seis alinhamentos entre o sumário de referência e o texto-fonte no idioma materno (português, nesse caso), além de onze alinhamentos entre o sumário de referência e os textos-fonte nas línguas estrangeiras (cinco com um idioma e seis com o outro).

Na Tabela 20, tem-se o resultado do referido cálculo para os 24 sumários de referência do CM3News.

Tabela 20 – Alinhamentos no *corpus* CM3News

Sumário de referência	**Total de sentenças do sumário de referência**	**Qt de alinhamentos (textos-fonte)**		
		Língua materna	**Língua estrangeira 1**	**Língua estrangeira 2**
1	10	60% (6/10)	50% (5/10)	60% (6/10)
2	7	57% (4/7)	43% (3/7)	57% (4/7)
3	11	18% (2/11)	27% (3/11)	91% (10/11)
4	5	20% (1/5)	20% (1/5)	100% (5/5)
5	8	63% (5/8)	63% (5/8)	38% (3/8)
6	3	67% (2/3)	100% (3/3)	33% (1/3)
7	14	29% (4/14)	64% (9/14)	57% (8/14)
8	8	25% (2/8)	63% (5/8)	75% (6/8)
9	15	53% (8/15)	73% (11/15)	60% (9/15)
10	8	50% (4/8)	88% (7/8)	75% (6/8)
11	21	67% (14/21)	43% (9/21)	19% (4/21)
12	12	83% (10/12)	33% (4/12)	8% (1/12)
13	14	64% (9/14)	64% (9/14)	64% (9/14)
14	8	75% (6/8)	75% (6/8)	63% (5/8)
15	9	44% (4/9)	67% (6/9)	78% (7/9)
16	5	60% (3/5)	60% (3/5)	100% (5/5)
17	15	93% (14/15)	33% (5/15)	27% (4/15)
18	7	100% (7/7)	29% (2/7)	29% (2/7)
19	15	73% (11/15)	47% (7/15)	47% (7/15)
20	8	88% (7/8)	50% (4/8)	38% (3/8)

21	11	91% (10/11)	18% (2/11)	18% (2/11)
22	5	100% (5/5)	40% (2/5)	20% (1/5)
23	14	50% (7/14)	57% (8/14)	64% (9/14)
24	6	67% (4/6)	67% (4/6)	67% (4/6)
TOTAL	**239**	**62,3% (149/239)**	**51,5% (123/239)**	**50,6% (121/239)**

Fonte: Elaborado pelo autor.

5.2.3 A origem das informações dos sumários de referência multilíngues

Com base na Tabela 20, observa-se que, do total de 239 sentenças que compõem os 24 sumários de referência do CM3News, 149 sentenças distintas foram alinhadas ao texto-fonte na língua materna do redator (português ou alemão, dependendo do sumário). Vale ressaltar, porém, que tais alinhamentos nem sempre são do tipo 1:1, pois uma mesma SS foi alinhada a uma ou mais sentenças de textos-fonte distintos. Ademais, houve 244 alinhamentos com os textos-fonte nas línguas estrangeiras (123+121), totalizando 393 alinhamentos. Preliminarmente, observa-se, portanto, que a língua materna esteve envolvida em 37,9% dos alinhamentos realizados (149/393), valor superior ao obtido pelas duas línguas estrangeiras avaliadas individualmente: 31,3% (123/393) para língua estrangeira 1 e 30,8% (121/393) para Língua estrangeira 2.

Dos 24 sumários de referência, 16 foram escritos por falantes nativos do português e 8 por nativos do alemão. Na Tabela 21, exibem-se os resultados dos alinhamentos com base na língua materna específica.

Tabela 21 – Alinhamentos no *corpus* CM3News por língua dos sumários de referência

Sumário de referência	**Total de sentenças**	**Alinhamentos por texto-fonte**		
		Português	**Inglês**	**Alemão**
Português	157	**104** (66,2%)	83 (52,9%)	72 (45,9%)
Alemão	82	40 (48,8%)	49 (59,8%)	**45** (54,9%)

Fonte: Elaborado pelo autor.

Com base nesses dados, pode-se dizer que os falantes nativos de português redigiram sumários de referência predominantemente com base no texto-fonte nessa língua (66,2%). Por outro lado, os nativos do alemão produziram sumários mais fortemente baseados no texto-fonte em inglês (59,8%), estando seu idioma materno em segundo lugar (54,9%).

Os resultados obtidos nessa etapa de avaliação podem ser vistos como indícios sobre a sumarização humana e, por conseguinte, sobre a influência da língua materna na produção de sumários. Por isso, não é possível tirar conclusões assertivas sobre a matéria, uma vez que a investigação se baseou em um *corpus* relativamente pequeno e os sumários de referência foram traduzidos para a comparação com os extratos automáticos. Além disso, deve-se levar em conta que a inexistência de *gold standards* para a SA, assim como ocorre em outras áreas da Linguística Computacional, dificulta e torna subjetiva a interpretação do que é tido como ideal.

CAPÍTULO 6 – Considerações finais

Neste trabalho, explorou-se o processo de avaliação de extratos produzidos pelos métodos de SAMM denominados CF e CFUL (TOSTA, 2014), sobretudo no que diz respeito à taxa de compressão e à língua materna dos produtores dos sumários de referência, que são utilizados na avaliação automática da informatividade dos extratos.

6.1 Contribuições

Acredita-se que este trabalho tenha produzido algumas contribuições para a área da SAMM.

A primeira delas diz respeito à ampliação do *corpus* CM2News, que passou a se chamar CM3News, sobretudo pela inclusão do alemão como segunda língua estrangeira a compor o referido *corpus* multidocumento multilíngue. Assim, contendo atualmente vinte coleções trilíngues (português, inglês e alemão) de notícias jornalísticas, pode-se dizer que CM3News é o principal *corpus* para subsidiar pesquisas em SAMM que envolvam a língua portuguesa. Em breve, esse recurso estará disponível para toda a comunidade linguística e do PLN pelo *website* do projeto Sustento[36].

A segunda contribuição feita por este trabalho é a anotação léxico-conceitual dos novos textos-fonte do CM3News por meio do MulSEN. Embora o referido editor tenha facilitado a tarefa de anotação dos textos/notícias em alemão, a impossibilidade de adaptá-lo por completo, no período de desenvolvimento desta pesquisa, para lidar com a língua alemã e as próprias características da referida língua fizeram da anotação léxico-conceitual uma tarefa desafiadora.

Ademais, como produto deste trabalho, destacam-se a geração automática de extratos multilíngues em português e a produção de sumários de referência para as coleções do CM3News. Quanto aos extratos, cada um dos métodos profundos de SAMM (CF e CFUL) gerou dois extratos multilíngues em português, sendo um com 30% de taxa de compressão (em relação à média da coleção) e um com 70% de taxa de compressão. Por conseguinte, o CM3News possui ao todo 80 extratos automáticos multilíngues. Sobre os sumários de referência, produziram-se, ao todo, 16 sumários

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

humanos do tipo *abstract* em português e 8 em alemão (posteriormente traduzidos para o português), os quais também compõem as respectivas coleções do CM3News.

Por fim, este trabalho revelou alguns indícios sobre o impacto da taxa de compressão e da língua materna do sumarizador humano na SAMM.

Sobre a taxa de compressão, as pesquisas realizadas ao longo dos 24 meses deste Mestrado mostraram que extratos automáticos mais curtos, com 70% de compressão, têm índices de qualidade linguística mais altos (os quais variam de 4,02 a 4,59 em uma escala de 0 a 5) que os extratos mais longos, com 30% de compressão (cujos valores variam entre 3,46 e 4,22). Os dados ajudam a comprovar a hipótese de que extratos mais curtos apresentam menos deficiências gramaticais e textuais. Tendo menos palavras, entende-se que há uma chance menor de erros de gramática (critério "gramaticalidade") e repetição de conteúdo (critério "não-redundância"). Além disso, a presença de menos sentenças também valoriza o critério "foco temático", pois reduz a probabilidade de haver segmentos que não se relacionam entre si.

Os critérios "clareza referencial" e "estrutura/coerência" obtiveram, no geral, as pontuações mais baixas nos sumários com ambas as taxas de compressão, o que mostra, mais uma vez, que os estudos linguísticos sobre a SAMM podem se concentrar, no futuro, na mitigação dos problemas resultantes da perda de referências e do emprego equivocado de elementos de coesão nos extratos, entre outros.

Por outro lado, em termos de informatividade, os extratos mais reduzidos, no geral, obtiveram medida-f mais baixa que os extratos com apenas 30% de compressão. A interpretação desses dados tem relação direta com os critérios de menor pontuação na avaliação da qualidade linguística. Extratos curtos dispõem de menos espaço para veicular o conteúdo e, por consequência, detalhamentos sobre entidades apresentadas nos textos (significados de siglas, explicações sobre pessoas etc.) acabam sendo sacrificados pela compressão, reduzindo seu grau de informatividade. Ainda, a seleção de sentenças de um ou mais textos para a geração de um extrato tende a prejudicar a estrutura desse material, já que parte considerável do(s) texto(s)-fonte não passa a compor o sumário.

Por fim, quanto à origem das informações contidas nos sumários de referência, verificou-se que 62,3% das sentenças desses *abstracts* possuem conteúdo presente nos textos-fonte escritos na língua materna dos produtores dos sumários de referência. Os dados também indicaram que, no contexto dessa avaliação envolvendo as três línguas-

fonte, 37,9% dos alinhamentos ocorreram com a língua materna, taxa que supera ligeiramente o valor de 1/3, caso os alinhamentos fossem uniformemente distribuídos entre os três idiomas.

Considerando sua inserção em uma esfera pioneira para a SAMM de base léxico-conceitual envolvendo a língua portuguesa, espera-se que este trabalho de exploração tenha apontado caminhos sobre o que pode ser melhorado e sobre o que já demonstra bons resultados. Futuramente, por meio de colaborações entre pesquisadores da Linguística, da Ciência da Computação e de áreas afins, vislumbra-se a possibilidade de que os métodos aqui investigados possam resultar em ferramentas computacionais para a geração de sumários.

Sabendo que a qualidade linguística e a informatividade estão diretamente relacionadas à extensão dos extratos, será possível tomar decisões mais consistentes na produção de métodos e/ou ferramentas de SA. Contudo, isso não significa que esses mesmos elementos não devam ser analisados sob outras perspectivas. Entre outras, seria relevante analisar o papel da TA nos métodos que a empregam (neste caso, apenas o método CF).

Partindo do princípio segundo o qual a tradução prévia de todos os textos-fonte é desnecessária e ocasiona falhas na gramaticalidade dos extratos, verifica-se que a abordagem aqui empregada gerou bons resultados. Aliás, o critério de gramaticalidade obteve a melhor pontuação na avaliação manual da qualidade linguística e foi aquele que apresentou a maior variação na comparação dos métodos CF e CFUL. Contudo, a mera observação do funcionamento dos tradutores automáticos mais conhecidos hoje em dia (sobretudo aqueles *on-line* e gratuitos) pode levar a questionamentos do tipo "Será que ainda vale a pena usar a *late translation*?". Nesse caso, com os métodos de sumarização aplicados a coleções multilíngues traduzidas automaticamente, poder-se-ia dizer que não houve mais sumarização multidocumento multilíngue, e sim monolíngue.

6.2 Dificuldades e limitações da pesquisa

Embora os resultados gerados por esta pesquisa possam contribuir para os estudos em SAMM, algumas ressalvas são importantes.

A primeira delas diz respeito ao tamanho do CM3News, que, contendo vinte coleções, é considerado pequeno, mesmo com a inclusão de um novo texto-fonte em

uma segunda língua estrangeira (alemão) a cada coleção. Aliás, ressalta-se que, como as coleções às quais os textos em alemão foram inseridos possuíam notícias sobre eventos ocorridos entre 2011 e 2013, a compilação atual de notícias em alemão sobre os mesmos eventos foi bastante difícil, mas bem-sucedida na medida do possível.

A observação sobre a limitação da extensão do *corpus* também se refere à quantidade de sumários de referência. Embora a investigação de 24 sumários humanos tenha permitido levantar alguns indícios sobre a relevância da taxa de compressão no desempenho dos métodos CF e CFUL e sobre a influência da língua materna na confecção dos sumários de referência, reconhece-se a necessidade de uma investigação mais ampla para a validação dos resultados.

Outra ressalva diz respeito às ferramentas e recursos disponíveis para a realização deste trabalho. Dado que a anotação léxico-conceitual do *corpus* foi feita por meio de uma ferramenta sem suporte para a língua alemã, alguns recursos automáticos não puderam ser empregados (no caso, a etiquetação morfossintática e a tradução), o que tornou a tarefa mais demorada do que o previsto e mais suscetível à subjetividade do anotador humano. Além disso, destaca-se que a WN.Pr, embora ainda seja um dos recursos léxico-conceituais mais respeitáveis no PLN, não continha muitos dos conceitos expressos nos textos em alemão devido a vários fatores: (i) a WN.Pr é um repositório conceitual fortemente baseado no léxico do inglês, (ii) a expressividade lexical da língua alemã pode ser bastante específica devido a seus processos morfológicos característicos e (iii) a limitação de atualização da WN.Pr frente ao surgimento de novos conceitos. Diante disso, foi necessário empregar várias estratégias para contornar tais problemas, sendo que se reconhece que tais estratégias geraram perda de conteúdo na anotação léxico-conceitual.

6.3 Trabalhos futuros

Tendo em vista os resultados (e contribuições) e as limitações encontradas no decorrer desta pesquisa, propõem-se os trabalhos futuros listados a seguir, todos eles restritos ao paradigma de PLN baseado em conhecimento linguístico.

- Investigar outros mecanismos de avaliação automática da informatividade, de modo a complementar para fins de comparação ou substituir as métricas ROUGE; uma

possibilidade, como inicialmente investigado por Ng e Abrecht (2015), é a integração de *word embeddings* para contornar o fato de que a ROUGE se pauta exclusivamente na forma dos n-gramas, em detrimento das possíveis correlações semânticas entre palavras.

- Ampliar o *corpus* CM3News, sobretudo pela inclusão de mais sumários (multilíngues) de referência, para validar os resultados indicativos que foram produzidos neste trabalho;
- Refinar o editor MulSEN para que todas as funcionalidades automáticas estejam efetivamente disponíveis para a anotação léxico-conceitual da língua alemã e/ou de outra(s) língua(s);
- Explorar outros métodos de análise da influência, na confecção dos sumários de referência, do texto-fonte escrito na língua materna do redator; tal análise pode ser feita, por exemplo, por meio de alinhamentos em nível léxico-conceitual entre os sumários de referência e os textos-fonte;
- Analisar o desempenho dos métodos CF e CFUL ou de outros que incluam a língua portuguesa na geração de extratos automáticos nas outras línguas do *corpus*, dado que os mesmos pares de línguas em sentidos opostos (p.ex., português → inglês vs. inglês → português) podem apresentar resultados diferentes após a TA.

REFERÊNCIAS BIBLIOGRÁFICAS

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

BENTIVOGLI, L.; PIANTA, E. Beyond Lexical Units: Enriching Wordnets with Phrasets. In: **EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS**, No. 3, 2003, Budapeste, Hungria. Proceedings... Budapeste, 2003, p. 67-70.

CAMARGO, R.T. **Investigação de estratégias de sumarização humana multidocumento**. 2013. 132 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, 2013.

CAMARGO, Y.V. **Sumarização Automática Multilíngue Multidocumento: seleção de conteúdo e tratamento da redundância com base em conhecimento léxico-conceitual**. 2019. 73f. Qualificação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, 2019.

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), 2019. Salvador/BA. **Proceedings...** Salvador, 2019, pp. 239-243.

CHAUD, M.R. **Investigação de estratégias de seleção de conteúdo baseadas na UNL (Universal Networking Language)**. 2015. 171 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, 2015.

CREMMINS, E.T. **The art of abstracting**. Arlington, Virginia: Information Resources Press, 1996.

CROSSLEY, S.A.; KYLE, K.; MCNAMARA, D.S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. **Behavior Research Methods**, V. 48, 2016, p. 1227-1237.

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

DI-FELIPPO, A.; TOSTA, F.E.S.; PARDO, T.A.S. Applying Lexical-Conceptual Knowledge for Multilingual Multi-Document Summarization. In: PROPOR, 12, 2016, Tomar. **Proceedings...Lecture Notes in Computer Science**, Vol 9727, Springer, Tomar, 2016, p. 38-49.

EBERHARD, D.M.; SIMONS, G.F.; FENNIG, C.D. (eds.) **Ethnologue**: Languages of the World. 22. ed. Dallas: SIL Internacional, 2019. Disponível em: <https://www.ethnologue.com>. Acesso em: 12 jul. 2019.

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

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

FELLBAUM, C. (ed.) **Wordnet**: an electronic lexical database (Language, speech and communication). Cambridge, MA: The MIT Press, 1998.

JURAFSKY, D; MARTIN, J. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2007. 1024p.

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: The 5th Annual International Conference on Systems Documentation*, 1986*, New York, NY, USA. **Proceedings…** New York, NY, 1986, p. 24–26.

LIN, C-Y.; HOVY, E.H. Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In: LANGUAGE TECHNOLOGY CONFERENCE, 2003, Edmonton/Canada. **Proceedings…** Edmonton, 2003.

LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, Cambridge, MA, Vol. 39, No. 2, 2013, p. 267-300.

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

MANI, I.; MAYBURY, M.T. (eds.) **Advances in automatic text summarization**. Cambridge, MA; London: The MIT Press, 2001.

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

MENZEL W. Robust processing of natural language. In: WACHSMUTH I., ROLLINGER CR., BRAUER W. (Eds) KI-95: Advances in Artificial Intelligence. KI 1995. **Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)**, v. 981. Springer, Berlin, Heidelberg, 1995.

MOHD, M.; JAN, R.; SHAH, M. Text document summarization using word embedding. In: **Expert Systems with Applications**, v. 143, 2020. Disponível em: <https://doi.org/10.1016/j.eswa.2019.112958>.

NENKOVA, A.; PASSONNEAU, R. Evaluating Content Selection in Summarization: The Pyramid Method. In: **Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics** (HLT-NAACL). Boston: 2004.

NG, J.P.; ABRECHT, V. Better Summarization Evaluation with Word Embeddings for ROUGE. **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. Association on Computational Linguistics, Lisboa, 2015, p. 1925-1930.

NÓBREGA, F. A. A. **Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento.** 2013. 126 p. Dissertação (Mestrado, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo, São Carlos, SP, 2013.

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

OUYANG, J.; SONG, B.; MCKEOWN, K. A Robust Abstractive System for Cross-Lingual Summarization. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:** Human Language Technologies, V. 1, Minneapolis, 2019, p. 2025-2031.

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

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

ROARK, B.; FISHER, S. OGI/OHSU baseline multilingual multi-document summarization system. In: MULTILINGUAL SUMMARIZATION EVALUATION (MSE) (Association for Computational Linguistics Workshop), 2005, Michigan, United States of America. *Proceedings...* Michigan, USA, 2005.

ROGERS, P.; PURYEAR, R.; ROOT, J. **Infobesity**: The enemy of good decisions. Bain & Company, jun. 2013. Disponível em: <https://www.bain.com/insights/infobesity-the-enemy-of-good-decisions>. Acesso em: 5 nov. 2018.

ROMÃO, T.L.C. Composição nominal em alemão: algumas peculiaridades do modelo ”adjetivo + substantivo”. **Trama**, Vol. 14, No. 31, 2018, p. 152-161.

SAGGION, H; RADEV, D.; TEUFEL, S.; WAI LAM, STRASSEL, S. M. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2002), 3, 2002, Las Palmas. **Proceedings...** Las Palmas: ELRA, 2002. p. 747-754.

SCHLUTER, N. The limits of automatic summarisation according to ROUGE. **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics**: Volume 2, Short Papers, Valencia, 2017, p. 41-45.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: INTERNATIONAL CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING, Manchester, UK. **Proceedings**… Manchester, 1994. p. 44-49.

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

SPARCK JONES, K. **Automatic summarising**: a review and discussion of the state of the art. Cambridge: University of Cambridge, 2007. (Technical Report UCAM-CL-TR-679).

SPARCK-JONES, K.; GALLIERS, J.R. **Evaluating natural language processing systems**: an analysis and review. Spring-Verlag HeidelBerlag, 1996.

TOSTA, F.E.S. **Aplicação de conhecimento léxico-conceitual na Sumarização Multidocumento Multilíngue**. 2015. 116 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, 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 IC EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 4, 2013. **Proceedings**... Fortaleza, 2013, p. 34-36.

VAN-HALTEREN, H.; TEUFEL, S. Examining the consensus between human summaries: initial experiments with factoid analysis. In: HLT-NAACL DUC WORKSHOP, 2003, Edmonton. **Proceedings**… Edmonton, 2003. p. 57-64.

VILLAVICENCIO, A.; RAMISCH, C.; MACHADO, A.; CASELI, H.M.; FINATTO, M.J. Identificação de Expressões Multipalavra em Domínios Específicos. **Linguamática**, Vol. 2, No. 1, abr. 2010, p. 15-34.

WAN, X.; LI, H.; XIAO, J. Cross-language document summarization based on machine translation quality prediction. In: **Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**. Uppsala, 2010, p. 917-926

W3TECHS. **Historical trends in the usage of content languages for websites**. Disponível em: <https://w3techs.com/technologies/history_overview/content_language>. Acesso em: 5 jul. 2019.

Apêndice A – Textos-fonte da coleção C16 do CM3News

1. Texto-fonte em português (Folha de São Paulo, 11/06/2013)

Uma nave da China decolou nesta terça-feira com três taikonautas - como são chamados os astronautas chineses - a bordo para uma missão de 15 dias em um laboratório espacial experimental, em mais um passo rumo ao desenvolvimento de uma estação espacial própria.

A Shenzhou 10 foi lançada em uma base remota no deserto de Gobi, no extremo oeste chinês, às 17h38 (6h38 em Brasília), numa tarde quente e de céu claro, conforme imagens transmitidas pela TV estatal.

Em junho de 2012, a China realizou sua primeira manobra bem sucedida de acoplagem no espaço, ligando-se ao pequeno módulo Tiangong 1, o que foi um marco na aquisição das capacidades tecnológicas e logísticas necessárias à operação de uma estação espacial completa, capaz de abrigar tripulantes por longos períodos.

O presidente Xi Jinping supervisionou pessoalmente o lançamento de terça-feira, dirigindo-se aos astronautas para lhes desejar sucesso e dizendo-se "enormemente feliz" por estar presente.

"Vocês são o orgulho do povo chinês, e esta missão é ao mesmo tempo gloriosa e sagrada", disse Xi, segundo a imprensa estatal.

Essa será a mais longa missão já feita por astronautas chineses.

A quinta viagem tripulada da China ao espaço desde 2003 foi acompanhada pelas habituais manifestações de orgulho nacional e propaganda do Partido Comunista, incluindo crianças vestidas em trajes de minorias étnicas, acenando para os três astronautas no centro espacial.

Há, no entanto, quem critique tamanho gasto na exploração espacial por parte de um país ainda em desenvolvimento, confrontado por questões mais prementes - da segurança alimentar à poluição e aos incêndios em fábricas.

"Por que não gastam esse dinheiro resolvendo os verdadeiros problemas da China em vez de desperdiçá-lo desse jeito?", escreveu um usuário no Sina Weibo, espécie de Twitter chinês.

O programa espacial chinês avançou muito desde que Mao Tsé-tung, fundador do regime comunista em 1949, lamentou o fato de seu país não ser capaz nem mesmo de colocar uma batata em órbita.

A China ainda está distante de se equiparar a EUA e Rússia, superpotências espaciais estabelecidas. No entanto, o avanço chinês nesse campo gera temores sobre uma corrida armamentista espacial.

2. Texto-fonte em inglês (BBC, 11/06/2013)

China has launched its latest Shenzhou manned space mission.
Three astronauts blasted away from the Jiuquan base in Inner Mongolia on a Long March 2F rocket at 17:38 Beijing time (09:38 GMT).
The commander, Nie Haisheng, and his crew, Zhang Xiaoguang and Wang Yaping, plan to spend just under two weeks at the orbiting Tiangong space lab.
Wang is China's second female astronaut and she will beam the country's first lesson from space to students on Earth.
The crew's capsule was ejected from the upper-stage of the rocket about nine minutes after lift-off. Mission controllers clapped enthusiastically once the ship's solar panels had been deployed.
Earlier in the day, Chinese TV carried pictures of President Xi Jinping wishing the crew luck.
"You have made Chinese people feel proud of ourselves," Xi told Nie and his colleagues. "You have trained and prepared yourselves carefully and thoroughly, so I am confident in your completing the mission successfully. "I wish you success and look forward to your triumphant return."
It should take just over 40 hours to raise the craft's orbit to the operating altitude of Tiangong some 335km (210 miles) above the planet's surface.
This mission, the fifth manned venture by China and scheduled to be the longest, is designated Shenzhou-10.
It is the latest step in China's plan to eventually put a permanently manned station above the Earth.
Tiangong-1 is the demonstrator. It was launched in 2011 to provide a target to test rendezvous and docking technologies.
The Shenzhou-9 crew - which included China's first female astronaut, Liu Yang - hooked up with the module for nearly 10 days in June 2012.
Nie's team aims to stay a few days longer, and like the crew of Shenzhou-9 will practise both manual and automatic dockings during the mission.
Beijing hopes to launch its fully-fledged station at the turn of the decade.
It is expected to have a mass of about 60 tonnes and comprise a number of interlocking modules.
Like the International Space Station (ISS), it will have long-duration residents and be supplied by robotic freighters.
China's human spaceflight programme is conducted largely in isolation to the ISS partners. But this could change in the next few years.
Europe in particular has opened a dialogue that could eventually result in flight opportunities for its astronauts on the proposed Chinese space station.
"We are looking at possibilities to use this space station," the European Space Agency's human spaceflight director Thomas Reiter told the BBC last month.
"The way ahead is that we will likely see first an exchange of experiments. And there are now also a few colleagues at the European Astronaut Centre who have started Chinese language training."

3. Texto-fonte em alemão (20 Minuten, 11/06/2013)

China setzt seinen langen Marsch zu einer eigenen Raumstation fort. Auf einer Rakete vom Typ «Langer Marsch 2F» hob die Mission «Shenzhou 10» am Dienstag um 17.38 Uhr Ortszeit (11.38 Uhr MESZ) vom Kosmodrom Jiuquan in der Inneren Mongolei ab. Die Reise der drei Taikonauten zum Raumlabor «Tiangong 1» (Himmelspalast), das die Erde in rund 335 Kilometern Höhe umkreist, dauert 40 Stunden. 15 Tage sollen die Taikonauten im All bleiben - solange wie noch kein chinesischer Raumfahrer zuvor.

Fast auf den Tag genau 50 Jahre nach dem ersten Flug einer Frau ins All ist mit Wang Yaping zum zweiten Mal eine chinesische Astronautin an Bord. Vor dem Start sagte die 33-Jährige, der Flug sei die Erfüllung des «chinesische Traums» von einem starken und wohlhabenden China. Als erste Frau war die heute 76-jährige Russin Valentina Tereschkowa am 16. Juni 1963 in den Weltraum gestartet. Seither sind schon mehr als 50 Frauen im All gewesen.

Auf dem fünften bemannten Raumflug Chinas planen die Astronauten ein manuelles und ein automatisches Andockmanöver mit dem Raummodul «Tiangong 1», das seit September 2011 die Erde umkreist. Chinas Staats- und Parteichef Xi Jinping verfolgte den erfolgreichen Start am Raumfahrtbahnhof. Bei einem Treffen mit den Astronauten zuvor sagte der Präsident: «Sie machen das chinesische Volk stolz.»

Die Astronauten werden etwa zwölf Tage in dem «Himmelspalast» wohnen. Sie sollen «neue Technologien zum Bau der Raumstation» sowie lebenserhaltende Systeme testen. Die Abfallverarbeitung sei verbessert worden. Auch gebe es neue Nahrung für die Astronauten.

Die Experimente und Übungen gelten als wichtige Voraussetzung für den langen Marsch der jungen Raumfahrernation zum Bau einer Raumstation bis 2020. China wäre dann die einzige Nation, die einen ständigen Aussenposten im All hätte, da die Internationale Raumstation ISS ausläuft.

In diesem Jahr will China noch eine Sonde auf dem Mond landen. Auch baut das Land gegenwärtig ein Satellitennetz für ein unabhängiges, weltumspannendes Navigationssystem.

Bei dem Flug von «Shenzhou 10» sollen erstmals auch chinesische Mittel- und Grundschüler über Video unterrichtet werden. Als Chinas erste Lehrerin im All wird die Astronautin Wang Yaping Themen wie Schwerelosigkeit, Oberflächenspannung von Flüssigkeiten sowie Gewicht und Masse erläutern.

Die Majorin ist eine erfahrene Pilotin und flog Einsätze nach dem Erdbeben 2008 mit 87'000 Toten in Sichuan. Auch nahm sie im selben Jahr an Impfaktionen aus der Luft zum Abregnen von Regenwolken während der Olympischen Spiele in Peking teil.

Wie der 46-jährige Zhang Xiaoguang ist Wang Yaping ein Neuling im All. Kommandeur des Fluges ist der erfahrene Astronaut Nie Haisheng. Der 48-Jährige ist 2005 bereits mit «Shenzhou 6» ins All geflogen und mit diesem Flug der älteste Astronaut Chinas im All.

Apêndice B – Textos-fonte e sumários de referência da coleção C4 do CM3News

1 – TEXTOS-FONTE

1.1 Texto-fonte em português

O terrorista de direita norueguês Anders Behring Breivik, autor confesso do massacre de 22 de julho passado na Noruega, usou produtos ilegais quando realizou seus ataques, informou a polícia nesta segunda-feira.
"Posso confirmar que ele usou entorpecentes ilegais.
Não desejo comentar que tipo de produto, mas ele os consumiu", declarou à AFP o procurador da polícia, Paal-Fredrik Hjort Kraby, referindo-se ao resultado dos exames de sangue.
No manifesto que difundiu na internet, logo depois de explodir um carro-bomba no centro de Oslo e abrir fogo contra uma reunião de jovens na ilha de Utoeya, Behring Breivik explica a forma com que utilizou esteroides anabolizantes.
Em 26 de julho, seu advogado, Geir Lippestad, também se referiu à utilização de substâncias dopantes para que ele se sentisse "forte, eficaz, alerta" no momento dos ataques, que deixaram 77 mortos.
"Os esteroides, claro, mas também usou outros medicamentos sobre os quais não quero dar detalhes", assinalou Hjort Kraby nesta segunda.
Por outra parte, o procurador da polícia afirmou que os psiquiatras designados para examinar o estado mental de Behring Breivik com a finalidade de determinar se é ou não responsável por seus atos iniciaram seus trabalhos.
Os dois especialistas devem entregar seu informe no mais tardar em 1º de novembro.
No dia 22 de julho, Breivik cometeu dois atentados em Oslo, na capital da Noruega, que deixaram 77 mortos.
Na primeira ação, um carro-bomba explodiu próximo à sede do governo, no centro de Oslo.
No segundo ataque, Breivik atirou contra os participantes de uma colônia de férias da juventude do Partido Trabalhista (no poder) na ilha de Utoya, 40 km a oeste da capital.
Os dois ataques foram cometidos com apenas duas horas de diferença.
A hipótese mais sólida era de que o suspeito tinha ativado o carro-bomba que explodiu na capital para depois seguir em direção à ilha, situada a cerca de 40 quilômetros da capital.
Um documento de 1.500 páginas redigido aparentemente pelo norueguês revela que o ataque já era preparado desde o outono (boreal) de 2009.
O documento, publicado na internet diariamente, inclui um manual sobre como montar bombas e um discurso contra o Islã e o marxismo.

1.2 Texto-fonte em inglês

The man who has confessed to killing 77 people in Norway has made a list of "unrealistic" demands, his lawyer says.
Anders Behring Breivik wanted the government to resign and Japanese specialists to assess his mental health, Geir Lippestad told reporters.
The far-right extremist admits killing eight people with a bomb in Oslo and shooting dead 69 on Utoeya island.
Meanwhile, the leader of the right-wing Progress Party has warned that Norway still faces a serious Islamist threat.
"All the debates that we had prior to 22 July will come back.
All the challenges that Norway was facing and the challenges that the world was facing are still there.
Al-Qaeda is still there," Siv Jensen told the AFP news agency.
"The new thing is that we have been in a horrible way reminded of the fact that terrorism can come in many different forms, with different rhetoric behind it, with different crazy ideas behind it."
Ms Jensen also said in another interview that the anti-Muslim views of Mr Breivik, who was a member of the Progress Party between 1999 and 2006, were "perversely unique" and that it was not aware of his plans.
"It was impossible for us to foresee at the time.
He obviously changed in recent years without anyone knowing," she told the Associated Press.
Mr Breivik blames the governing Labour Party for increased immigration in Norway.
Its youth wing was on Utoeya for a summer camp when the attack took place, while the bomb was set-off near government buildings.
Mr Lippestad said Mr Breivik's list of demands was "far from the real world" and "completely impossible to fulfil" and showed "he doesn't know how society works".
"His demands here includes the complete overthrowing of both the Norwegian and European societies," he told the Associated Press.
"But it shows that he doesn't understand the situation he's in."
The 32-year-old had linked his demands to his willingness to share information about other alleged terrorist cells, Mr Lippestad said.
Norwegian police have previously cast doubt on Mr Breivik's claims that he was part of a broader network but said they would investigate them.
A court has appointed two psychiatrists to try to examine Mr Breivik's actions, with a mandate to report back by 1 November.
Mr Lippestad said Mr Breivik had asked that he also be examined by Japanese mental health specialists as he believes "the Japanese understand the idea and values of honour" and would understand him better than Europeans.
The lawyer has previously said his client is probably insane.
Mr Lippestad added that a second list from his client requested items like cigarettes and civilian clothes.

Mr Breivik has been charged under the criminal law for acts of terrorism.
The charges include the destabilisation of vital functions of society, including government, and causing serious fear in the population.
At a court appearance on 25 July, Mr Breivik admitted carrying out the attacks but did not plead guilty to the charges.
He was remanded in custody for eight weeks, with the first four to be in solitary confinement.
The attacks on 22 July traumatised Norway, one of the most politically stable and tolerant countries in Europe.
The government plans to set up an independent "July 22 Commission" to examine the attacks, including investigating whether police reacted too slowly to the shootings at Utoeya.

1.3 Texto-fonte em alemão

Die Polizei bestätigt, dass im Blut von Anders Behring Breivik Rauschmittel nachgewiesen wurden.
Nähere Angaben macht sie nicht.
Breiviks Anwalt spricht von einem "Medikamente-Cocktail".
Der norwegische Doppelattentäter Anders Behring Breivik stand während seiner Taten am 22. Juli unter dem Einfluss von Drogen.
Ein Polizeisprecher bestätigte am Wochenende gegenüber dem Norwegischen Rundfunk NRK, dass in der bei dem 32-Jährigen nach seiner Festnahme genommenen Blutproben Rauschmittel nachgewiesen werden konnten.
Breivik sei "unter Einfluss" gestanden, sagte Polizei-Staatsanwalt Christian Hatlo gegenüber NRK.
Dazu, um welche Mittel oder Stoffe es sich dabei gehandelt habe, wolle er sich im Detail nicht äußern.
Laut der Online-Ausgabe von "Verdens Gang" handelte es sich unter anderem um anabole Stereoide sowie eine hohe Konzentration von Koffein.
Der Verteidiger Breiviks, Geir Lippestad, sagte, sein Mandant habe erklärt, vor dem Doppel-Attentat einen "Medikamente-Cocktail" geschluckt zu haben.
Bei dem Bombenanschlag auf das Regierungsviertel in Oslo und dem Massaker auf der rund 40 Kilometer entferneten Insel Utöya wurden insgesamt 77 Menschen getötet und fast ebensoviele unterschiedlichen Grades verletzt.
Auf Utöya hatte das traditionelle Feriencamp der Sozialdemokratischen Parteijugend stattgefunden.

2 – SUMÁRIOS DE REFERÊNCIA EM PORTUGUÊS

2.1 Sumário com 30% de compressão

Segundo relato da polícia norueguesa, os resultados do exame de sangue de Anders Behring Breivik apontam que o terrorista estava sob influência de drogas durante os dois atentados que cometeu no dia 22 de julho de 2018.
Perante a isso, será investigado se o norueguês de 32 anos pode ser responsabilizado pelos seus atos no fatídico dia.
Os atentados, dos quais Breivik é acusado, ocorreram em frente ao prédio do governo e em um acampamento de jovens afiliados ao Partido dos Trabalhadores – atualmente no poder.
No primeiro, houve a explosão de um carro-bomba e, no segundo, o autor abriu fogo contra os presentes no local, culminando na morte de 77 pessoas e aproximadamente essa mesma quantidade de feridos.
Breivik admite ter executado tais ações, porém, não se considera culpado.
Com base em um texto redigido pelo extremista de direita, há indícios de que a sua motivação advém do medo de um suposto processo de islamização do ocidente e de uma grave insatisfação com a atual gestão do governo.
Indagada a esse respeito, a líder do partido de direita Progress Party, do qual Breivik já pertenceu, afirma não ter estado ciente dos planos do terrorista e que ele mudou radicalmente nos últimos anos.
Geir Lippestad, advogado de defesa de Breivik, defende a hipótese de que seu cliente está insano, dado que, além das razões do crime, ele tem feito exigências para colaborar com as investigações, dentre as quais: ser analisado por psiquiatras japoneses e que o governo renuncie.

2.2 Sumário com 70% de compressão

Os resultados do exame de sangue de Anders Behring Breivik apontam que o terrorista estava sob influência de drogas durante os dois atentados que cometeu: um em frente ao prédio do governo e o outro em um acampamento de jovens do Partido dos Trabalhadores (atualmente no poder), culminando na morte de 77 pessoas.
Há indícios de que a sua motivação advém do medo de uma suposta islamização do ocidente e de uma grave insatisfação com o atual governo.
Geir Lippestad, advogado de defesa de Breivik, defende a hipótese de que seu cliente está insano, pois, além das razões do crime, ele tem feito exigências para colaborar com as investigações, dentre elas a renúncia do governo.