

Validating Semantic Content in Sentence Simplification using OpenIE and Relational Graphs

Samuel Rios da Silva¹, Larrissa Dantas¹, Daniela Barreiro Claro¹,
Aline Paes², Maria José Finatto³, Rerisson Cavalcante⁴, Silvana Ribeiro⁴

¹FORMAS - Centro de Pesquisa em Dados e Linguagem Natural
Instituto de Computação, Universidade Federal da Bahia (UFBA)
Salvador, BA, Brasil

{samuelrs, larrissasilva, dclaro}@ufba.br

²Instituto de Computação, Universidade Federal Fluminense (UFF)
Niterói, RJ, Brasil
alinepaes@ic.uff.br

³Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brasil
mariafinatto@gmail.com

⁴Instituto de Letras, Universidade Federal da Bahia
Salvador, BA, Brasil
{rerisson.cavalcante, silvanar}@ufba.br

Abstract. *Sentence simplification often alters sentence structure while aiming to preserve meaning, making semantic fidelity difficult to evaluate. In this paper, we investigate whether graph-based representations derived from Open Information Extraction (Open IE) can capture semantic preservation across different simplification levels. Using the PorSimplesSent corpus, we extract relational triples from original and simplified sentences to build semantic graphs for comparison. Our results provide exploratory evidence that graph-based analysis can capture meaningful semantic differences and provides complementary signals beyond surface-level lexical overlap for evaluating simplification quality.*

Resumo. *A simplificação de sentenças frequentemente altera a estrutura textual enquanto preserva o significado, tornando a fidelidade semântica difícil de avaliar. Neste trabalho, investigamos se representações baseadas em grafos derivadas de Extração Aberta de Informação (Open IE) podem capturar a preservação semântica em diferentes níveis de simplificação. Utilizando o corpus PorSimplesSent, extraímos triplas relacionais de sentenças originais e simplificadas para construir grafos semânticos comparáveis. Os resultados fornecem indícios exploratórios de que a análise baseada em grafos pode capturar diferenças semânticas relevantes e fornece sinais complementares além da sobreposição lexical superficial para avaliar a qualidade da simplificação.*

1. Introdução

A simplificação textual tem ganhado relevância crescente no contexto brasileiro, especialmente após a promulgação da Lei nº 15.263/2025, que instituiu o uso de linguagem simples em documentos e comunicações públicas¹. Essa iniciativa buscou ampliar a acessibilidade às informações prestadas por órgãos públicos e garantir que seus conteúdos

¹<https://www2.camara.leg.br/legin/fed/lei/2025/lei-15263-14-novembro-2025-798293-publicacaooriginal-177011-pl.html>

fossem compreendidos por uma parcela mais ampla da população, incluindo pessoas com diferentes níveis de letramento e proficiência linguística, e pessoas com necessidades especiais [Finatto 2024].

Nesse cenário, a área de Processamento de Linguagem Natural (PLN) tem desempenhado um papel central no desenvolvimento de métodos automáticos voltados à simplificação textual [Leal et al. 2024, Leal et al. 2026]. Paralelamente, diversos estudos têm se dedicado à construção de recursos e insumos voltados à avaliação automática da complexidade textual e dos níveis de legibilidade [de Menezes et al. 2023].

As técnicas contemporâneas de simplificação automática de textos combinam abordagens linguísticas, métodos de aprendizado de máquina e, mais recentemente, modelos de linguagem de grande porte (LLMs) [Kim 2022, Liu et al. 2024, Alves et al. 2023, Scalercio et al. 2024, Shardlow et al. 2024, Baez and Saggion 2023, Scalercio et al. 2025]. Tradicionalmente, essas abordagens utilizam conjuntos de características linguísticas e corpora paralelos, compostos por textos complexos e suas respectivas versões simplificadas, para orientar modelos supervisionados ou estratégias heurísticas de simplificação [Chandrasekar et al. 1996b]. As transformações podem ocorrer em diferentes níveis linguísticos, desde o nível lexical, com a substituição de termos complexos por alternativas mais frequentes ou mais acessíveis, até o nível sintático, envolvendo a divisão de sentenças, a redução de constituintes e a reorganização de estruturas para formas mais diretas e canônicas [Saggion and Hirst 2017].

Um *corpus* paralelo pioneiro que oferece acervos de simplificações de frases para o português do Brasil é o PorSimplesSent [Leal et al. 2018]. O PorSimplesSent utiliza técnicas de simplificação de sentenças baseadas em aspectos lexicais e sintáticos e detém uma estrutura composta por uma sentença original complexa (Original), acompanhadas de duas versões simplificadas. A primeira versão sofreu apenas simplificação lexical e é denominada Natural. A segunda versão (chamada *Strong*) emprega técnicas sintático-lexicais e inclui elementos psicolinguísticos, com a incorporação e/ou o resgate de elementos mencionados em sentenças anteriores.

Nesse processo, preservar o significado da sentença original em suas versões simplificadas permanece um dos principais desafios da simplificação automática de textos [Agrawal and Carpuat 2024]. Embora as transformações realizadas possam melhorar aspectos como legibilidade e acessibilidade, elas também podem introduzir perdas de informação, omissões relevantes ou alterações semânticas sutis, comprometendo a fidelidade do conteúdo original. Nesse contexto, diferentes trabalhos, sobretudo na língua inglesa, têm investigado métodos para avaliar a preservação semântica na simplificação textual [Vásquez-Rodríguez et al. 2021].

A título ilustrativo, segue um exemplo de simplificação com omissão de informação (omissão que a ópera *estava moribunda*), acréscimo e/ou retomada de elementos (para o item *dele*, referido como *Pavarotti* em sentenças anteriores) e alteração da ordem de elementos (substantivo-adjetivo - ópera viva)².

- **Original:** A importância dele é ter mantido viva a ópera, que estava moribunda – diz Chaves.

²Sentença do extrato aleatório do *corpus* PorSimplesSent.

- **Simplificada (neutra):** Chaves diz que a importância de Pavarotti é ter mantido viva a ópera.
- **Simplificada (strong):** Chaves diz que a importância de Pavarotti é ter mantido a ópera viva.

A avaliação automática em simplificação textual ainda representa um desafio, especialmente no que se refere à preservação semântica, uma vez que métricas tradicionais frequentemente não capturam adequadamente omissões de informação, alterações de significado ou reorganizações estruturais mais profundas [Alva-Manchego et al. 2021]. Por outro lado, embora avaliações humanas sejam capazes de identificar tais fenômenos com maior precisão, sua realização em larga escala torna-se impraticável devido ao alto custo de anotação, à necessidade de especialistas e ao tempo requerido para análises consistentes. Dessa forma, torna-se necessária a investigação de novas formas de avaliar a simplificação textual, principalmente no que cerne a preservação da semântica.

Com isso em mente, o presente trabalho realiza uma análise qualitativa das sentenças originais em relação às respectivas simplificações que integram o *corpus* PorSimplesSent, por meio de grafos gerados a partir de informações extraídas das sentenças. Partimos da hipótese de que o significado central de um texto pode ser representado por relações semânticas explícitas entre entidades e predicados. Assim, ao converter sentenças originais e simplificadas em grafos relacionais, torna-se possível comparar não apenas a similaridade lexical superficial, mas também a preservação da estrutura semântica subjacente. Divergências estruturais entre os grafos podem indicar perda de informação, omissão de relações relevantes ou alterações semânticas introduzidas pela simplificação, enquanto similaridades sugerem maior fidelidade ao conteúdo original.

Para investigar essa hipótese de forma exploratória, triplas de sentenças foram extraídas do *corpus* PorSimplesSent. O principal objetivo foi explorar como a estrutura relacional dos grafos gerados reflete a preservação semântica em diferentes níveis de simplificação. Com auxílio desses grafos e da apreciação de avaliadores humanos, descrevemos em que medida esse método pode apoiar a verificação da preservação semântica em diferentes processos de simplificação.

Especificamente, pretende-se responder às seguintes questões de pesquisa:

- Como os diferentes níveis e estratégias de simplificação presentes no *corpus* PorSimplesSent se refletem na estrutura dos grafos gerados por Open IE?
- Em casos de preservação semântica, de que forma essa conservação se manifesta na estrutura relacional dos grafos?
- Em casos de distorção, omissão ou alteração semântica, como essas modificações impactam a estrutura dos grafos gerados?

Os experimentos foram divididos em quatro etapas principais. A primeira etapa analisou a omissão de informação e a mudança de sentido com a participação de avaliadores linguistas. A segunda buscou observar se os grafos gerados a partir das sentenças simplificadas detinham graus menores de complexidade que os grafos de sentenças originais. A terceira etapa buscou descrever em que medida tanto omissões de informação quanto mudanças de sentido repercutem na geração de novos nós e arestas nos grafos das sentenças simplificadas. Na quarta etapa, foram comparadas as anotações realizadas pelos linguistas com os resultados obtidos a partir dos grafos gerados.

O presente artigo está organizado como segue. A Seção 2 apresenta os trabalhos relacionados. A seção 3 descreve a tarefa de simplificação. A Seção 4 descreve a proposta deste trabalho. A Seção 5 descreve os experimentos e resultados, e casos particulares são discutidos na Seção 6. Por fim, a Seção 7 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Diversos trabalhos investigam os desafios do processo de avaliação da simplificação textual, seja por meio de métricas automáticas, de análise morfossintática ou ainda de avaliação humana.

Métricas Automáticas e Morfossintáticas: As métricas automáticas baseadas em referências, como SARI [Xu et al. 2016] e BERTScore [Zhang et al. 2020], são utilizadas na literatura para medir simplicidade e preservação de significado [Scalercio et al. 2025]. Para avaliações sem referência, índices clássicos de legibilidade como o Flesch Readability Index e o Flesch-Kincaid Grade Level (FKGL) [Kincaid et al. 1975] ainda são adotados para verificar a legibilidade da sentença resultante [Scalercio et al. 2026, Alva-Manchego et al. 2020]. Métricas morfossintáticas baseadas em hipóteses psicolinguísticas [Heineman et al. 2023] têm sido utilizadas para validar o processo de simplificação, como a razão Lemma/Token para diversidade lexical, a proporção de voz passiva para ativa, proporção de orações adverbiais que precedem a oração principal, capturando tendências na estrutura da frase, e razão entre orações relativas totalmente desenvolvidas e reduzidas, refletindo simplificações sintáticas [Scalercio et al. 2025, Scalercio et al. 2026].

Avaliação Humana e Análise de Erros: Diante da insuficiência das métricas automáticas na simplificação de sentenças, a avaliação humana permanece como o padrão-ouro. Abordagens tradicionais analisam o texto simplificado sob uma escala Likert de três eixos fundamentais: adequação (preservação do significado), simplicidade (as sentenças geradas são mais simples?) e fluência (as frases geradas são gramaticais ou bem formadas?) [Alva-Manchego et al. 2020]. [Wu and Arase 2026] propõe uma abordagem focada em falhas e voltada ao resultado, mapeando sete tipos de erros (como Altered Meaning e Lack of Simplicity, subdivididos em eixos lexicais e estruturais).

Embora alguns métodos propostos [Ivanov and Solnyshkina 2020, Silva et al. 2025, Queiroz et al. 2023] utilizem grafos derivados de representações semânticas, nenhum dos trabalhos mapeados investiga a preservação do conteúdo semântico por meio da extração de informação aberta e da análise de divergências estruturais entre grafos gerados a partir de diferentes níveis de simplificação.

3. Simplificação de Sentenças

A tarefa de simplificação de sentenças consiste em transformar uma sentença de entrada, tipicamente complexa, em uma ou mais sentenças de saída linguisticamente mais simples, com o objetivo de facilitar sua leitura e compreensão, preservando ao máximo seu conteúdo semântico original [Chandrasekar et al. 1996a, Al-Thanyyan and Azmi 2021].

A simplificação pode envolver diferentes tipos de operações, incluindo substituições lexicais por termos mais frequentes ou acessíveis, reestruturação sintática por meio da decomposição de sentenças longas e complexas em unidades mais curtas, e reorganização discursiva para melhorar a fluidez e a clareza textual.

Tipicamente, um *corpus* de simplificação, como o PorSimpleSent [Leal et al. 2018], inclui pares alinhados entre uma sentença original e suas versões simplificadas em diferentes níveis de intensidade, permitindo investigar sistematicamente os efeitos de simplificação na preservação semântica.

Dentre o processo de simplificação das sentenças no PorSimpleSent, há sentenças que são simplificadas por uma operação de *split*. Consequentemente, a simplificação de uma sentença original pode resultar em duas ou mais sentenças simplificadas. Como a simplificação por *split* gera triplas e grafos independentes para cada nova sentença, a comparação estrutural entre versões simplificadas e originais torna-se mais difícil e menos interpretável. Por essa razão, foram selecionadas apenas sentenças cuja simplificação não envolvesse processos de *split*.

4. Análise semântica da simplificação

Com o intuito de validar o conteúdo semântico na simplificação de sentenças por meio de extração de informação [Oliveira et al. 2022] e seus respectivos grafos, a metodologia foi organizada em quatro etapas principais: (1) a preparação do corpus e a representação linguística das sentenças; (2) a análise semântica das sentenças simplificadas sob a perspectiva da omissão de informação e preservação da semântica; (3) análise da simplificação via extração de informação e seus grafos, e, por fim, (4) uma análise comparativa quanto aos grafos gerados, considerando limitações em sentenças complexas e simplificações com mudança semântica significativa. A Figura 1 descreve o pipeline de validação da preservação do conteúdo semântico das sentenças simplificadas.

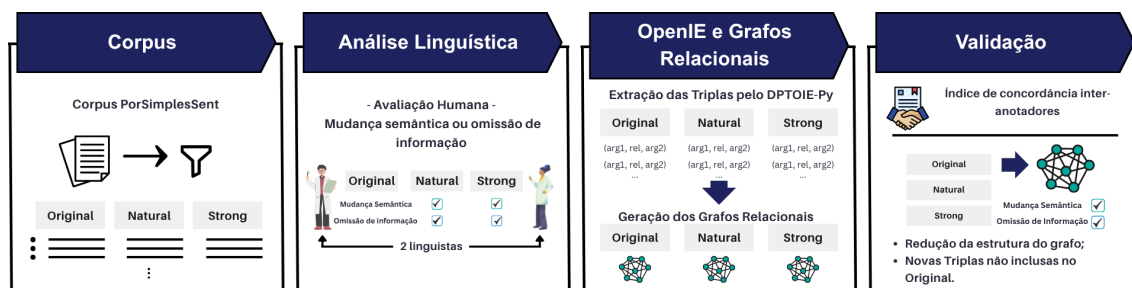


Figura 1. Pipeline

A **primeira etapa** consistiu na preparação do corpus a ser utilizado na validação da simplificação. Considerando os dois níveis de simplificação: $O \rightarrow N$ (Original para Natural) e $O \rightarrow S$ (Original para Strong), foi aplicado um filtro no *corpus* PorSimpleSent [Leal et al. 2018] para selecionar os pares de sentenças que não sofreram o processo de *split* e foram modificadas no processo de simplificação (i.e. cuja sentença original se difere léxico-sintaticamente da sentença simplificada), resultando em um total de 264 sentenças.

Desse conjunto, foram selecionadas 20 sentenças originais alinhadas com suas versões simplificadas, objetivando obter um conjunto diverso de representações linguísticas do PorSimpleSent para análise e validação por meio deste trabalho. O processo de aleatoriedade foi feito selecionando 20 valores no intervalo de [1, 264].

A **segunda etapa** foi responsável pela análise semântica das sentenças simplificadas sob a perspectiva da omissão da informação e da mudança de sentido. As 20 sentenças

selecionadas foram avaliadas manualmente por dois linguistas, responsáveis por analisar, independentemente, cada simplificação, observando dois critérios definidos:

- *Omissão de informação*: a omissão de informação resulta na remoção de conteúdos relevantes das sentenças simplificadas que estavam presentes na sentença original.
- *Mudança de sentido*: a mudança de sentido corresponde às alterações semânticas que alteram o significado da sentença original.

A **terceira etapa** consistiu na definição do método de extração de triplas a ser utilizado e na geração dos grafos. Cada sentença s , em sua versão Original ou Simplificada, foi processada através do método de extração de informação denominado *DPToie-Py*³ e gerou um conjunto de triplas $T(s) = \{(arg_1, rel, arg_2), \dots\}$, onde $T(s)$ representa a função que mapeia uma sentença para um conjunto de triplas relacionais [Oliveira et al. 2022].

Associamos à sentença s o grafo $G(s) = (V_s, E_s)$, onde o conjunto de vértices V_s é gerado através dos argumentos das triplas ($arg_1, arg_2 \in V_s$) e o conjunto de arestas E_s é gerado através da relação rel de seus argumentos. $G(s)$, portanto, é a função de construção do grafo da sentença s a partir de triplas extraídas em $T(s)$ [Oliveira et al. 2022].

Com o intuito de dar um maior subsídio à análise dos grafos, decorrente do processo de simplificação textual, a árvore de dependência de cada sentença $UD(s) = PortParser(s)$ também foi gerada através do *Portparser.v2* [Lopes and Pardo 2024]. Assim, a análise do processo de simplificação é representada através de cada sentença s e seus elementos associados $\beta_s = \{T(s), G(s), UD(s)\}$ nas versões **Original**, **Natural** e **Strong**.

A hipótese central é a de que simplificações de sentenças que preservam a semântica tendem a manter a mesma estrutura informacional do grafo original, ainda que apresentem redução estrutural. A validação dos grafos considerou dois aspectos principais:

- *Redução estrutural do grafo*: análise da simplificação topológica decorrente da redução textual, observando diminuição no número de nós, arestas ou relações;
- *Introdução de novas informações*: identificação de triplas inexistentes na sentença original, indicando possíveis distorções semânticas, inferências indevidas ou mudanças de sentido decorrentes da simplificação.

Por fim, na **quarta etapa**, foram comparadas as anotações dos linguistas aos resultados obtidos a partir dos grafos gerados. Essa comparação teve como objetivo verificar o grau de correspondência entre a avaliação humana de preservação semântica na simplificação de sentenças e os indícios estruturais observados nos grafos de conhecimento gerados a partir destas sentenças.

5. Experimentos e Resultados

Esta seção descreve os experimentos conduzidos neste trabalho e discute os resultados obtidos.

³Versão melhorada do *DPToie*[Oliveira et al. 2022] em Python, com melhorias na extração de relações semânticas

5.1. Experimento A - Análise Linguística

Considerando a efetiva redução de complexidade observada nas sentenças simplificadas de acordo com o PorSimplesSent, foram adotados dois critérios de avaliação com o objetivo de analisar o comportamento das simplificações em termos de *omissão de informação* e *preservação da semântica*. Esses critérios são descritos a seguir:

- *Omissão de informação*: a omissão de informação é responsável pela remoção de conteúdo relevante nas sentenças simplificadas, que estavam presentes na sentença original.
- *Mudança de sentido*: a mudança de sentido identifica as alterações semânticas que modifiquem o significado da sentença original.

Cada conjunto de sentenças **Original**, **Natural** e **Strong** foi analisado sob esses dois critérios por dois linguistas seniores, que realizaram as suas avaliações de maneira independente.

5.1.1. Resultados

Os resultados foram divididos em duas etapas de avaliação da simplificação: $O \rightarrow N$ e $O \rightarrow S$. Em relação à *omissão de informação*, das 20 sentenças simplificadas, os anotadores apresentaram concordância de 80% para a classe negativa, ou seja, não há *omissão de informação* nas simplificações, tanto para as $O \rightarrow N$ quanto $O \rightarrow S$. Somente em um caso, ambos concordaram com a *omissão da informação* nas sentenças, como segue na Tabela 1.

Versão	Sentença
Original	A importância dele é ter mantido viva a ópera, que estava moribunda – diz Chaves.
Natural	Chaves diz que a importância de Pavarotti é ter mantido viva a ópera.

Tabela 1. Exemplo de simplificação textual com *omissão de informação* entre as versões $O \rightarrow N$ com concordância inter-anotadores.

Em relação à *mudança de sentido*, observa-se que, na simplificação $O \rightarrow N$, a concordância observada inter-anotadores é 70%, sendo quatro casos de concordância que houve *mudança de sentido* na frase simplificada em relação à original e 10 casos de concordância para a classe negativa. Além disso, há seis casos nos quais o Anotador 2 considerou que houve *mudança de sentido* em divergência do Anotador 1 (Tabela 2).

Versão	Sentença
Original	Segundo eles, a pele humana ainda é bem mais complexa do que as imitações desenvolvidas até agora.
Natural	Eles dizem que a pele humana ainda é mais muito mais complexa do que as cópias desenvolvidas até agora.

Tabela 2. Exemplo de simplificação textual com *mudança de sentido* entre as versões $O \rightarrow N$ com concordância inter-anotadores.

Ainda em relação à *mudança de sentido*, observa-se que em relação às sentenças simplificadas das versões **Strong**, a concordância observada inter-anotadores decresceu

para 65%, sendo cinco casos de concordância que houve mudança de sentido na frase simplificada em relação à original. Além disso, há 7 casos nos quais o Anotador 2 considerou que houve *mudança de sentido*, mas esse comportamento não foi observado pelo Anotador 1.

Observa-se que, na avaliação humana, não houve divergência quanto à *omissão de informação* nas sentenças simplificadas. Por outro lado, há divergência entre anotadores quanto à preservação da semântica considerando a análise da *mudança de sentido*, definindo assim a tarefa como difícil. Nestes casos, evidencia-se a necessidade de analisar se esses comportamentos se refletem nos grafos gerados por meio das triplas.

5.2. Experimento B - Análise dos Grafos

Esse experimento analisa a simplificação sob a perspectiva dos grafos relacionais. Para cada sentença s dos pares $O \rightarrow N$ e $O \rightarrow S$ foi extraído um conjunto $T(s)$ de triplas (ver Tabelas 3 e 4) e, consequentemente, gerado o grafo $G(s)$ por meio das triplas extraídas (ver Figuras 2 e 3).

A avaliação da simplificação das sentenças por meio dos grafos teve como principal objetivo analisar as relações entre as omissões de informações e a estrutura do grafo, assim como as relações entre as mudanças de sentido nas sentenças simplificadas e as modificações nas arestas e nós dos grafos das sentenças simplificadas.

A análise de redução de um grafo foi realizada independentemente da omissão de informação, com o intuito de verificar se, mesmo na ausência de omissão, ocorre redução do grafo na simplificação de sentenças. Consequentemente, a observação de novos nós e arestas também foi analisada independentemente com o intuito de isolar os casos de preservação de semântica, mesmo em caso de não omissão de informação. As árvores de dependência $UD(s_i)$ foram utilizadas para ampliar a compreensão das extrações e grafos gerados.

5.2.1. Resultados

Considerando a análise de redução do grafo nas sentenças simplificadas $O \rightarrow N$, observou-se que apenas quatro grafos de sentenças simplificadas reduziram sua estrutura topológica, o que corresponde a uma redução de 20%. Em relação aos grafos simplificados $O \rightarrow S$, esse número aumentou para sete, perfazendo um total de 35% das simplificações. Neste caso, embora haja uma simplificação das sentenças, elas não são refletidas na estrutura topológica dos grafos.

Avaliando a correspondência com a *omissão de informação*, no único caso em que ambos os anotadores concordaram (Tabela 3), também não houve redução na estrutura do grafo (Figura 2).

Considerando a *preservação da semântica*, ou seja, que no processo de simplificação não houve *mudança de sentido* das sentenças simplificadas $O \rightarrow N$ e $O \rightarrow S$, os grafos gerados para cada uma delas foram avaliados quanto à existência de uma alteração nos nós e nas arestas, caracterizando novos fatos incorporados que não existiam na versão original da sentença.

Observou-se que, em 30% das simplificações $O \rightarrow N$, houve alterações nas arestas e nós dos grafos simplificados. Considerando as quatro concordâncias entre anotadores sobre a *mudança de sentido* de $O \rightarrow N$, em 50% também houve uma correspondência na modificação das arestas e dos nós. Os dois casos específicos nos quais os anotadores concordaram que houve uma mudança de sentido, mas não houve novo nó ou aresta, observou-se que o tema foi *política*, e a concordância de ambos se referiu ao significado do termo “união” que diverge de “coligação”. Esse resultado não tem impacto direto no grafo gerado, conforme a Figura 3.

#	Original			Natural		
	arg_1	rel	arg_2	arg_1	rel	arg_2
1	A importância de ele	é	mantido	A importância de Pavarotti	é	mantido
2	A importância de ele	é	mantido viva a ópera , estava moribunda	A importância de Pavarotti	é	mantido viva a ópera
3	A importância de ele	ter mantido	viva a ópera , estava moribunda	A importância de Pavarotti	é	viva a ópera
4				A importância de Pavarotti	ter mantido viva	a ópera
5				Chaves	ter mantido	viva a ópera, estava moribunda

Tabela 3. Conjuntos de triplas da sentença Original 12 e de sua versão Simplificada ($O \rightarrow N$).

Considerando as simplificações $O \rightarrow S$, em 40% dos casos houve *mudança de nós e arestas* nos grafos gerados. Nestes casos, dos cinco casos de concordância entre anotadores com *mudança de sentido* $O \rightarrow S$, em 80% dos casos houve alteração nas arestas e nós. Neste caso, há um forte indício de que o grafo reflete a mudança de sentido, quando há uma concordância entre os anotadores.

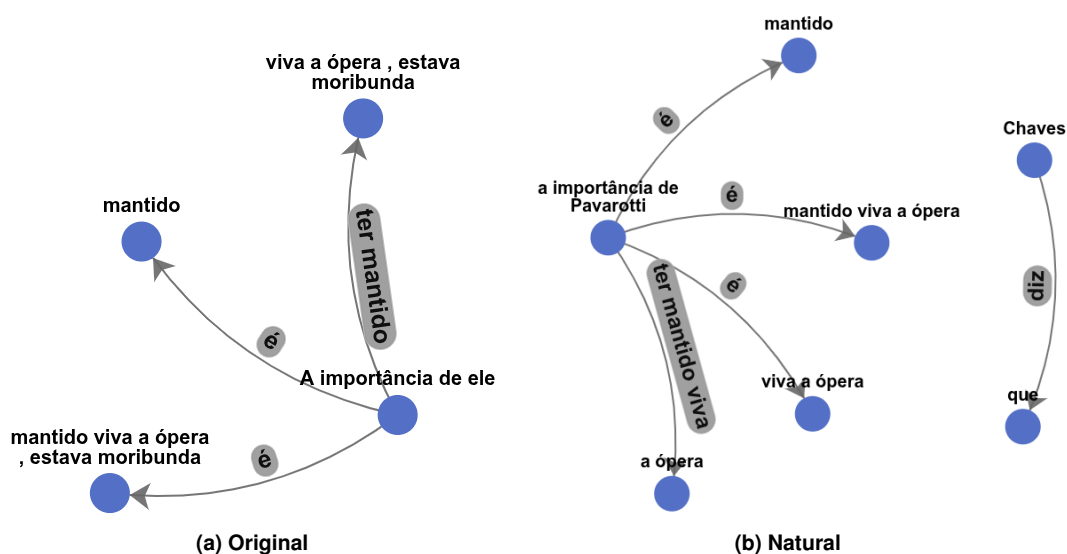


Figura 2. Representação em grafo da simplificação da sentença Original na versão Natural (Sentença 12).

Observou-se que a omissão de informação não foi uma das causas da mudança de sentido neste *corpus*, visto que poucas foram as ocorrências de omissões (1 de $O \rightarrow N$ e 3 de $O \rightarrow S$).

Em sua maioria, a mudança de sentido ocorreu em um único nó, com a substituição do termo por outro termo que não carregava mais o mesmo significado semântico, tais como “união” e “coligação”, “devota” e “apegada”. Nestes casos, não há mudança no grafo e nem essas nuances no significado são absorvidas para gerar novas arestas e novos nós.

#	Original			Natural		
	arg_1	rel	arg_2	arg_1	rel	arg_2
1	A regra	obriga fechar	apenas coligações não colidam com as nacionais	A regra	obriga fechar	apenas as uniões não se choquem com as nacionais
2	A regra	obriga fechar	em os Estados apenas coligações não colidam com as nacionais	A regra	obriga fechar	em os Estados apenas coligações as uniões não se choquem com as nacionais
3	A regra	obriga fechar	em os Estados	A regra	obriga fechar	em os Estados

Tabela 4. Conjuntos de triplas da sentença Original 14 e de sua versão Simplificada ($O \rightarrow N$).

6. Discussões

As discussões refletem algumas particularidades durante o processo de avaliação da geração de grafos a partir da Open IE para a validação da preservação semântica na simplificação de sentenças.

Nos casos em que houve a redução estrutural, não houve modificações nos grafos, ou seja, arestas novas ou nós novos não foram criados.

Os resultados indicam que a análise de preservação semântica por meio de grafos gerados por OpenIE fornece bons indícios de alterações semânticas, principalmente quando há uma concordância entre os anotadores. Essa concordância reflete que a mudança de sentido é observável. Além disso, quando a mudança ocorre no nível da sintaxe, envolvendo principalmente o deslocamento do “root” na Árvore de Dependência, os novos nós ou arestas nos grafos gerados por esse método de OpenIE tende a ser mais evidente, devido à heurística intrínseca ao método.

As mudanças no significado de um ou dois termos em uma sentença, tais como as diferenças políticas entre “união” e “coligação”, ou ainda casos como “devoto” e “apegado” não são refletidas em novas arestas ou nós no grafo, visto que estão inseridas no arg_1 ou arg_2 (Figura 3). Nestes casos, o grafo não reflete essa mudança de significado. Por outro lado, em se tratando de apostro, ou seja, uma omissão de informação da sentença original, o grafo reflete essa omissão tanto topologicamente, quando na mudança de sentido, caso o apostro seja permutado como agente da sentença.

7. Conclusões

Este trabalho investigou o uso de representações baseadas em grafos derivados de técnicas de Open IE, para apoiar a análise da preservação semântica em tarefas de simplificação

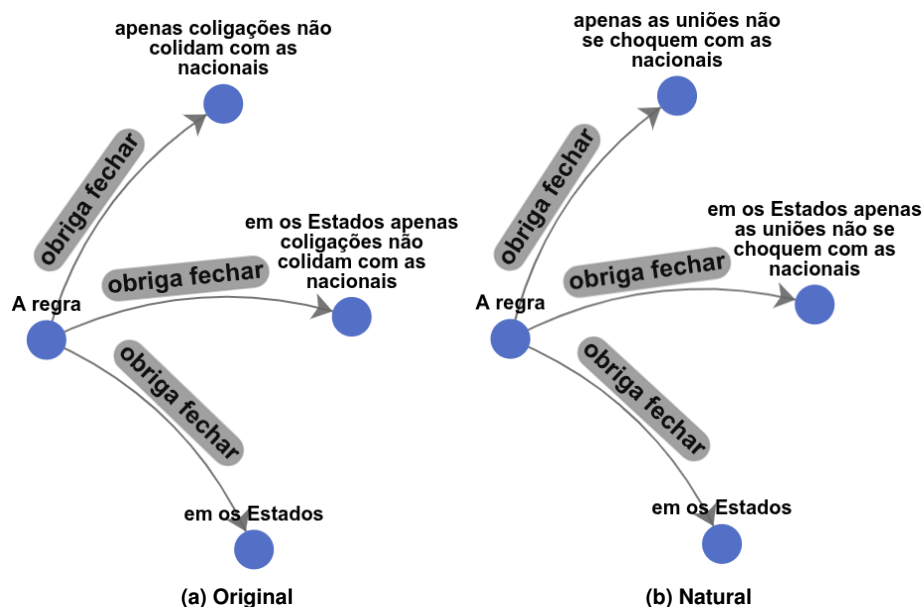


Figura 3. Representação em grafo da simplificação da sentença $O \rightarrow S$ (Sentença 14) com mudança de sentido mas sem mudança no grafo.

de sentenças em português. A partir do corpus PorSimpleSent, foram comparadas as versões original e simplificadas, com o objetivo de verificar em que medida as alterações de complexidade textual se refletem em mudanças estruturais nos grafos gerados.

Os resultados indicam que os grafos derivados de Open IE fornecem indícios úteis sobre a preservação ou a alteração de significado na simplificação textual. Em particular, divergências estruturais entre os grafos da sentença original em relação às simplificadas mostraram-se correlacionadas aos casos de mudança de sentido, conforme identificado na análise linguística. As omissões de informação não foram observadas sendo refletidas nas estruturas topológicas dos grafos.

Como trabalho futuro, pretendemos ampliar a avaliação para outros corpora, diferentes métodos de extração de informação e novos domínios textuais, investigando se os achados deste estudo exploratório permanecem consistentes em cenários mais amplos.

Limitações

Nesta seção descrevemos as limitações evidenciadas em relação a este trabalho. Algumas extrações, de acordo com as heurísticas definidas pelo DPTToie-Py, não foram realizadas, e, consequentemente, os grafos não foram gerados, resultando em grafos vazios (Sentenças *Strong 17* e *Strong 9*). Essa limitação impediu a análise nestes casos. Outra limitação deste trabalho é em relação ao subconjunto do *corpus* PorSimpleSent, definido como aleatório, não contemplando uma amplitude de casos linguísticos por tipos de simplificação, condicionando os experimentos a uma baixa cobertura para análise dos grafos. A expansão do *corpus* e a definição de um guia de anotação são trabalhos futuros para mitigar essas limitações.

Declaração Ética

Este trabalho investiga a avaliação da preservação semântica na simplificação textual, utilizando representações relacionais derivadas de *Open Information Extraction* (Open IE). Como a pesquisa foi conduzida sobre um corpus público de simplificação textual, não houve coleta de dados pessoais sensíveis nem participação direta de seres humanos no processo experimental. Entretanto, reconhece-se que tanto os sistemas de simplificação automática quanto os métodos de extração de informação podem introduzir vieses linguísticos, omissões semânticas ou interpretações incorretas das relações presentes no texto original. Tais limitações podem impactar especialmente grupos com diferentes níveis de letramento, bem como aplicações em domínios sensíveis, como educação, saúde e comunicação governamental. Além disso, as representações relacionais produzidas por Open IE dependem da qualidade das extrações automáticas, podendo refletir erros do sistema de extração e não apenas alterações semânticas introduzidas pela simplificação. Assim, os resultados deste trabalho devem ser interpretados como sinais auxiliares de preservação semântica, e não como uma medida definitiva de equivalência de significado.

Uso de IA generativa

Ferramentas de IA generativa foram utilizadas para o refinamento textual e a correção gramatical.

Agradecimentos

Os autores agradecem o financiamento parcial da bolsa CAPES do PIBIC 2025/2026, do apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 (processo nº 88887.970806/2024-00), do CNPq (processo 307088/2023-5) e da FAPERJ (processos SEI-260003/002930/2024 e SEI-260003/000614/2023). Este trabalho também conta com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (processo 408490/2024-1). E por fim, os autores agradecem o apoio do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), por meio da bolsa de Produtividade em Pesquisa (PQ, processo 307088/2023-5), do PROBIC FAPERGS-UFRGS e da FAPERGS, por meio do Edital 06/2025 – Programa Pesquisador Gaúcho (PqG) (processo 25/2551-0002702).

Referências

- Agrawal, S. and Carpuat, M. (2024). Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Al-Thanyyan, S. S. and Azmi, A. M. (2021). Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- Alva-Manchego, F., Scarton, C., and Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.
- Alva-Manchego, F., Scarton, C., and Specia, L. (2021). The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

- Alves, A., Miranda, P., Mello, R., and Nascimento, A. (2023). Automatic simplification of legal texts in portuguese using machine learning. In *Legal Knowledge and Information Systems*, pages 281–286. IOS Press.
- Baez, A. and Saggion, H. (2023). LSLlama: Fine-tuned LLaMA for lexical simplification. In Štajner, S., Saggion, H., Shardlow, M., and Alva-Manchego, F., editors, *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Chandrasekar, R., Doran, C., and Bangalore, S. (1996a). Motivations and methods for text simplification. In *COLING 1996 volume 2: The 16th international conference on computational linguistics*.
- Chandrasekar, R., Doran, C., and Srinivas, B. (1996b). Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.
- de Menezes, L. C., Paes, A., and Finatto, M. J. B. (2023). Abordagem baseada em aumento de dados para avaliação automática de leituraabilidade. *Domínios de Linguagem*, 17:e1721.
- Finatto, M. J. B. (2024). Da linguagem simples à acessibilidade textual e terminológica: um percurso com e na linguística. In SILVA, A. H. P.; LAGARES, X. C. M. M. o., editor, *Linguagem simples: para quem? A comunicação cidadã em debate*, pages 86–110. ABRALIN, Campinas.
- Heineman, D., Dou, Y., Maddela, M., and Xu, W. (2023). Dancing between success and failure: Edit-level simplification evaluation using SALSA. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.
- Ivanov, V. and Solnyshkina, M. (2020). A method for assessment of text complexity based on knowledge graphs. In *Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence*, volume 2852 of *CEUR Workshop Proceedings*, Moscow, Russia. CEUR-WS.org.
- Kim, T. (2022). Revisiting the practical effectiveness of constituency parse extraction from pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5398–5408, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for army enlisted personnel. Technical Report Research Branch Report 8-75, Chief of Naval Technical Training, Millington, TN, USA.
- Leal, S., Duran, M., and Scarton, C. (2024). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Lang Resources & Evaluation*, 58.
- Leal, S. E., Duran, M. S., and Aluísio, S. M. (2018). A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese. In Bender, E. M., Derczynski,

- L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Leal, S. E., Serras, F. R., Finger, M., and Aluísio, S. M. (2026). Complexidade textual e suas tarefas relacionadas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3, book chapter 6. BPLN, 4 edition.
- Liu, Y., Fabbri, A., Chen, J., Zhao, Y., Han, S., Joty, S., Liu, P., Radev, D., Wu, C.-S., and Cohan, A. (2024). Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.
- Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. I*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Oliveira, L., Claro, D. B., and Souza, M. (2022). Dptoie: a portuguese open information extraction based on dependency analysis. *Artif. Intell. Rev.*, 56(7):7015–7046.
- Queiroz, B., Cavalcante, R., and Claro, D. (2023). Desafios da tarefa de extração de informação aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 388–392, Porto Alegre, RS, Brasil. SBC.
- Saggion, H. and Hirst, G. (2017). *Automatic text simplification*, volume 32. Springer.
- Scalercio, A., Bertotto, E., Jesus, S., Finatto, M. J., and Paes, A. (2026). Annotation guidelines and challenges for automatic simplification of Portuguese drug leaflets. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 121–127, Salvador, Brazil. Association for Computational Linguistics.
- Scalercio, A., Finatto, M., and Paes, A. (2024). Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 15076–15091, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Scalercio, A. M. R. D. A., Souza, E. A. D., Finatto, M. J. B., and Paes, A. (2025). Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477, Vienna, Austria. Association for Computational Linguistics.

- Shardlow, M., Alva-Manchego, F., Batista-Navarro, R., Bott, S., Calderon Ramirez, S., Cardon, R., François, T., Hayakawa, A., Horbach, A., Hülsing, A., Ide, Y., Imperial, J. M., Nohejl, A., North, K., Occhipinti, L., Rojas, N. P., Raihan, N., Ranasinghe, T., Salazar, M. S., Štajner, S., Zampieri, M., and Saggion, H. (2024). The BEA 2024 shared task on the multilingual lexical simplification pipeline. In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z., editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- Silva, S., Athaydes, A., Mane, B., Claro, D., Souza, M., Neto, F. M., Dantas, L., and Cavalcante, R. (2025). Desafios dos grafos de conhecimento: uma proposta de avaliação de sistemas openie. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 718–722, Porto Alegre, RS, Brasil. SBC.
- Vásquez-Rodríguez, L., Shardlow, M., Przybyła, P., and Ananiadou, S. (2021). Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.
- Wu, X. and Arase, Y. (2026). An in-depth evaluation of large language models in sentence simplification with error-based human assessment. *ACM Trans. Intell. Syst. Technol.*, 17(4).
- Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.