**Núcleo Interinstitucional de Linguística Computacional – NILC**
Universidade de São Paulo - USP
Universidade Federal de São Carlos – UFSCar

# Alinhamento Manual dos Sumários Humanos e dos Textos-Fonte do *Corpus* Multidocumento CSTNews

## Relatório Técnico
## NILC-TR-01-12

**Verônica Agostini**
**Renata Tironi de Camargo**
**Ariani Di Felippo**
**Thiago Alexandre Salgueiro Pardo**

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

## Resumo

Neste relatório, descreve-se a tarefa manual de alinhamento dos sumários humanos (*abstracts*) multidocumento a seus respectivos textos-fonte, pois a tarefa mencionada se mostra relevante para o desenvolvimento, em especial, de pesquisas linguísticas e computacionais que podem subsidiar a sumarização automática multidocumento. Os sumários e textos alinhados compõem o CSTNews (Cardoso *et al.*, 2011), um *corpus* jornalístico multidocumento em português do Brasil composto por 50 coleções de textos provenientes de diferentes fontes e que versam sobre diferentes assuntos ou temas. O alinhamento em questão busca, especificamente, relacionar o conteúdo das sentenças que constituem os sumários às sentenças de origem dos textos-fonte. Para tanto, elaborou-se um manual de anotação que especifica um conjunto de regras gerais e específicas de alinhamento sumário-texto. O alinhamento ora descrito encontra-se codificado em esquemas XML de modo que o mesmo possa ser utilizado em futuras pesquisas linguístico-computacionais.



O trabalho relatado contou com o financiamento da FAPESP e CAPES.

# Índice

# 1. Introdução (Contexto e Justificativas)

O alinhamento consiste em relacionar segmentos textuais (palavras, sentenças ou parágrafos) de diferentes documentos ou até mesmo documentos inteiros. Essa tarefa é utilizada em várias aplicações desenvolvidas no Processamento Automático das Línguas Naturais (PLN), como tradução automática (p.ex.: Gale e Church, 1991, 1993; Yamada e Knight, 2001; Caseli, 2003), perguntas e respostas (p.ex.: Soricut e Brill, 2004), simplificação textual (p.ex.: Specia, 2010), sumarização automática (p.ex.: Marcu, 1999; Hirao *et al.*, 2004), entre outras. Em todas as aplicações, o processo de alinhamento (ou indexação) possibilita de uma forma geral a aquisição, manual ou automática, de conhecimento sobre a tarefa a ser automatizada.

Na tradução automática, por exemplo, busca-se relacionar especialmente um texto original em uma língua-fonte a sua versão em uma ou mais línguas-alvo distintas, ou seja, os chamados *textos paralelos*. Os *corpora* resultantes desse processo de alinhamento são recursos linguísticos extremamente importantes, pois permitem a aquisição automática de equivalências lexicais e padrões ou regras de tradução (p.ex.: Gale e Church, 1991, 1993; Yamada e Knight, 2001; Caseli, 2003). No Quadro 1, apresenta-se um alinhamento em nível sentencial entre dois textos paralelos, um em português do Brasil (PB) e o outro em espanhol.

**Quadro 1:** Exemplo de alinhamento sentencial na tradução automática.

| **Português** | **Espanhol** |
|---|---|
| Há pessoas que evitam ingerir carne e laticínios e, para assegurar o necessário suprimento de ferro e cálcio, substituem esses alimentos por vegetais - em geral os verde-escuros como espinafre, couve-manteiga e brócolis. | Existen personas que evitan ingerir carne y lácteos y, para asegurarse el necesario aporte de hierro y calcio, reemplazan estos alimentos por vegetales -en general los de color verde oscuro, como las espinacas, coles y brócolis. |
| Contudo, segundo Jocelem Mastrodi Salgado, da Escola Superior de Agricultura Luiz de Queiroz (Esalq), nem sempre esses minerais são aproveitados pelo organismo quando se consume verduras: algumas delas contêm substâncias tóxicas que impedem a absorção dos nutrientes. | Con todo, según Jocelem Mastrodi Salgado, de la Escuela Superior de Agricultura Luiz de Queiroz (Esalq), no siempre esos minerales son aprovechados por el organismo cuando se consumen verduras: algunas de éstas contienen sustancias tóxicas que impiden la absorción de los nutrientes. |

Especificamente, observa-se que: (i) a sentença em PB na primeira linha da primeira coluna do Quadro 1 foi alinhada à sentença em espanhol da primeira linha da segunda coluna e (ii) a sentença em PB na segunda linha da primeira coluna foi alinhada à sentença em espanhol da segunda linha da segunda coluna.

Nos exemplos do Quadro 1, diz-se que os alinhamentos são do tipo 1-1. No entanto, uma unidade textual (palavra, sentença ou parágrafo) de um texto-fonte pode ser relacionada a mais de uma unidade do(s) texto(s)-alvo, o que caracteriza os alinhamentos do tipo 1-N. O contrário também ocorre, ou seja, mais de uma unidade do texto-fonte pode ser alinhada a uma única do(s) texto(s)-alvo, caracterizando um

alinhamento do tipo N-1. Na tradução automática, é comum que uma sentença da língua-fonte seja alinhada a mais de uma sentença da língua-alvo. O mesmo é verificado nos alinhamentos de nível lexical, já que uma unidade lexical da língua-fonte (p.ex.: *contudo* em PB) pode ser equivalente a uma unidade complexa (p.ex.: *con todo* em espanhol) ou expressão na língua-alvo e vice-versa. Além disso, nem sempre é possível alinhar segmentos de textos distintos, o que resulta nos alinhamentos do tipo 1-0.

Na aplicação de perguntas e respostas, busca-se alinhar uma pergunta a sua resposta (p.ex.: Soricut e Brill, 2004) e, na simplificação textual, um texto original é alinhado a sua versão simplificada (produzida por humanos) (p.ex.: Specia, 2010). Em ambas as aplicações, os diferentes alinhamentos têm o mesmo objetivo, que é o de possibilitar a aquisição de conhecimento sobre a tarefa a ser automatizada. No caso da simplificação textual, por exemplo, os alinhamentos, como o apresentado no Quadro 2, podem possibilitar o aprendizado de regras de simplificação.

**Quadro 2:** Exemplo de alinhamento sentencial na simplificação textual.

| Sentença original | Sentença simplificada |
|---|---|
| Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea. | Cientistas britânicos detectaram em adultos que células-tronco da medula óssea produziram células do fígado. |

Na sumarização automática (SA), em especial, busca-se alinhar um sumário humano (isto é, um *abstract*) a um ou mais textos-fonte (p.ex.: Marcu, 1999; Hirao *et al*., 2004). Tendo em vista a quantidade de textos-fonte, um sumário resultante do processo de sumarização humana monodocumento é alinhado a um único documento e um sumário resultante do processo de sumarização humana multidocumento é alinhado a mais de um texto-fonte. No Quadro 3, ilustra-se especificamente o alinhamento de um sumário humano multidocumento a dois textos-fontes que tratam do mesmo assunto. No caso, a sentença do sumário está alinhada a uma única sentença de cada texto-fonte, caracterizando, assim, um alinhamento 1-2.

**Quadro 3:** Exemplo de alinhamento sentencial na sumarização automática.

| Sumário | Documentos |
|---|---|
| O Brasil não fará parte do trajeto de 20 países do revezamento da tocha. | A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico. |
| | O Brasil não faz parte do trajeto da tocha olímpica. |

Especificamente na SA, o alinhamento de sumários humanos ou manuais e textos-fonte evidencia a origem das informações que compõem o sumário, permitindo investigar suas características e a forma por meio da qual foram transpostas para o sumário. Desse tipo de investigação, é possível obter estratégias de sumarização humana que podem subsidiar a sumarização automática, tornando-a mais linguisticamente motivada. Tais estratégias, concebidas como regras explícitas, podem ser obtidas por análise manual dos alinhamentos ou por meio de aprendizado de máquina.

Tendo em vista a relevância do alinhamento na SA, procedeu-se ao alinhamento dos sumários humanos multidocumento e dos textos-fonte que compõem o *corpus* CSTNews (Cardoso *et al*., 2011), o qual foi realizado no âmbito dos projetos SUCINTO[1]. A Figura 1 ilustra o alinhamento referente à coleção 31 do CSTNews, que pertence à categoria "esporte" e é composta por dois documentos ou textos-fonte.

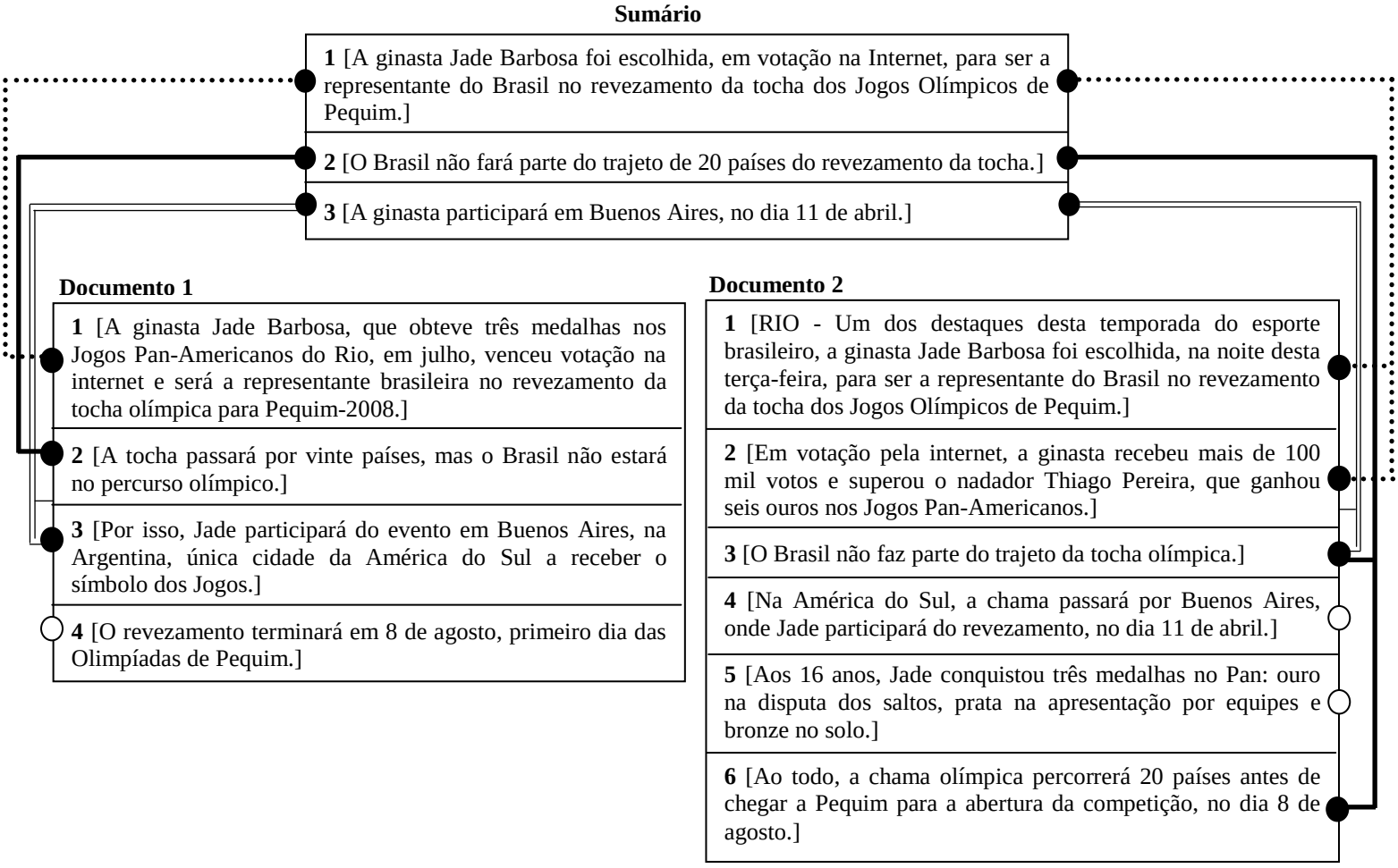


**Figura 1:** Exemplo do alinhamento de um sumário a seus textos-fonte.

Na Figura 1, observa-se, por exemplo, que a sentença (S) 1 do sumário foi alinhada à S1 do documento (D) 1 e à S1 e S2 do D2. No caso da coleção 31, nenhuma sentença do sumário foi alinhada à S4 do D1 e à S4 e S5 do D2.

Neste relatório, descreve-se em detalhes o referido alinhamento do CSTNews, o qual, aliás, será utilizado em dois trabalhos de mestrado em andamento que focam a sumarização automática multidocumento (SAM) (Mani, 2001).

Em um deles, de natureza computacional, o alinhamento manual será utilizado para avaliar o desempenho de uma ferramenta computacional (isto é, um alinhador automático) que buscará simular o processo humano/manual de alinhamento. Tal ferramenta está sendo desenvolvida porque o alinhamento, visto como uma etapa do processo de sumarização, permite a aplicação de diferentes métodos de SAM.

[1] http://www.icmc.usp.br/~taspardo/sucinto

Na outra pesquisa, de natureza linguística, o alinhamento manual do CSTNews permitirá a investigação de estratégias de sumarização humana multidocumento que, uma vez formalizadas, podem subsidiar métodos de SAM. A análise linguística dos alinhamentos manuais pode revelar estratégias de sumarização humana multidocumento quanto à seleção de conteúdo e produção dos sumários.

Para a descrição do alinhamento ora mencionado, este relatório foi organizado em 5 Seções. Na Seção 2, descrevem-se as características do *corpus* multidocumento CSTNews (Cardoso *et al*., 2011). Na Seção 3, descreve-se especificamente o processo de alinhamento dos sumários humanos e dos textos-fonte, destacando-se, sobretudo, as regras utilizadas na tarefa. Na Seção 4, os resultados da tarefa ora descrita são apresentados. Por fim, na Seção 5, algumas considerações finais são feitas.

## 2. O *corpus* CSTNews

O CSTNews[2] é um *corpus* jornalístico composto por 50 coleções ou grupos de textos, sendo que cada coleção versa sobre um tema distinto (Cardoso *et al*., 2011). Os textos são do gênero discursivo "notícias jornalísticas", pertencentes à ordem do relatar (Dolz, Schnewly, 2004; Barbosa, 2001; Lage, 2004). As principais características das "notícias" são: (i) documentar as experiências humanas vividas (domínio social) e (ii) representar pelo discurso as experiências vividas, situadas no tempo (capacidade da linguagem).

Cada coleção do CSTNews contém 2 ou 3 textos sobre um mesmo assunto ou tema compilados de diferentes fontes jornalísticas e seus respectivos sumários humanos e automáticos, além de diversas anotações. Quanto às fontes jornalísticas, ressalta-se que os textos-fonte foram coletados dos principais jornais *online* do Brasil, a saber: *Folha de São Paulo, Estadão, Jornal do Brasil, O Globo* e *Gazeta do Povo*. A coleta manual foi feita durante aproximadamente 60 dias, de agosto a setembro de 2007. As coleções possuem em média 42 sentenças (de 10 a 89) e os sumários humanos multidocumento possuem em média 7 sentenças (de 3 a 14). Ademais, as coleções estão categorizadas pelos rótulos das "seções" dos jornais dos quais os textos foram compilados. Assim, o *corpus* é composto por coleções das seguintes categorias: "esporte" (10 coleções), "mundo" (14 coleções), "dinheiro" (1 coleção), "política" (10 coleções), "ciência" (1 coleção) e "cotidiano" (14 coleções).

Quanto aos sumários humanos multidocumento, ressalta-se que estes foram construídos manualmente de forma abstrativa, ou seja, com reescrita. Consequentemente, a identificação da origem, nos textos-fonte, da informação que constitui um sumário nem sempre é direta. Além disso, a produção dos mesmos foi guiada por uma taxa de compressão de 70%. Consequentemente, os sumários contêm, no máximo, 30% do número de palavras do maior texto-fonte da coleção.

Os dados completos sobre o CSTNews constam da Tabela 1.

[2] O *corpus* foi denominado "CSTNews" porque os textos-fonte estão alinhados no interior de cada coleção pelas relações semântico-discursivas (p.ex.: as relações de *equivalence*, *identity*, *subsumption*, etc.) do modelo linguístico-computacional *Cross-Document Structure Theory* (CST) (Radev, 2000). O alinhamento dos textos-fonte em função dessa teoria tem subsidiado a proposição de métodos de SAM.

**Tabela 1:** Estatísticas do CSTNews.

| Coleção | Categoria | Nº de documentos | Nº de sentenças/ documentos | Nº de sentenças/ sumários |
|---|---|---|---|---|
| **C1** | Mundo | 3 | 24 | 5 |
| **C2** | Política | 3 | 51 | 7 |
| **C3** | Cotidiano | 3 | 50 | 10 |
| **C4** | Cotidiano | 3 | 39 | 5 |
| **C5** | Cotidiano | 2 | 23 | 5 |
| **C6** | Cotidiano | 3 | 36 | 5 |
| **C7** | Ciência | 2 | 23 | 4 |
| **C8** | Esportes | 3 | 25 | 6 |
| **C9** | Política | 3 | 36 | 6 |
| **C10** | Mundo | 3 | 38 | 10 |
| **C11** | Cotidiano | 3 | 56 | 11 |
| **C12** | Mundo | 3 | 34 | 4 |
| **C13** | Mundo | 3 | 37 | 6 |
| **C14** | Mundo | 3 | 25 | 5 |
| **C15** | Mundo | 3 | 26 | 6 |
| **C16** | Política | 3 | 47 | 6 |
| **C17** | Política | 2 | 41 | 6 |
| **C18** | Mundo | 3 | 70 | 9 |
| **C19** | Esportes | 2 | 13 | 4 |
| **C20** | Política | 3 | 42 | 8 |
| **C21** | Cotidiano | 3 | 41 | 3 |
| **C22** | Cotidiano | 3 | 50 | 9 |
| **C23** | Mundo | 2 | 25 | 6 |
| **C24** | Esportes | 3 | 24 | 5 |
| **C25** | Esportes | 3 | 88 | 8 |
| **C26** | Mundo | 3 | 58 | 10 |
| **C27** | Esportes | 3 | 89 | 12 |
| **C28** | Esportes | 3 | 35 | 4 |
| **C29** | Mundo | 3 | 48 | 6 |
| **C30** | Dinheiro | 3 | 46 | 4 |
| **C31** | Esportes | 2 | 10 | 3 |
| **C32** | Mundo | 3 | 66 | 9 |
| **C33** | Cotidiano | 3 | 68 | 13 |
| **C34** | Cotidiano | 3 | 59 | 8 |
| **C35** | Mundo | 3 | 36 | 7 |
| **C36** | Cotidiano | 3 | 74 | 14 |
| **C37** | Cotidiano | 2 | 26 | 5 |
| **C38** | Esportes | 3 | 26 | 3 |
| **C39** | Cotidiano | 3 | 34 | 3 |
| **C40** | Política | 3 | 28 | 4 |
| **C41** | Esportes | 3 | 45 | 6 |
| **C42** | Política | 2 | 39 | 5 |
| **C43** | Política | 3 | 49 | 7 |
| **C44** | Política | 2 | 26 | 9 |
| **C45** | Cotidiano | 3 | 47 | 6 |
| **C46** | Mundo | 3 | 23 | 5 |
| **C47** | Mundo | 3 | 43 | 6 |
| **C48** | Esportes | 2 | 43 | 9 |
| **C49** | Cotidiano | 3 | 23 | 6 |
| **C50** | Política | 3 | 62 | 8 |
| **Total** | — | 140 | 2067 | 331 |
| **Média** | — | 2.8 | 41.34 | 6.62 |

# 3. O alinhamento manual

## 3.1. Caracterização geral

A tarefa em questão foi realizada por 2 anotadores da área de Linguística Computacional durante aproximadamente 2 meses, em reuniões diárias de 1 a 2 horas. Cada pesquisador ficou responsável por alinhar metade das coleções do CSTNews.

O alinhamento, em especial, foi feito em função de duas diretrizes centrais. A primeira delas diz respeito ao nível dos segmentos a serem alinhados e a segunda refere-se ao critério para a identificação das correspondências. Quanto ao nível dos segmentos textuais, optou-se pelo sentencial, posto que as sentenças não unidades de informação bem delimitadas. Sobre o critério de alinhamento propriamente dito, ressalta-se que as correspondências entre os sumários e seus respectivos textos-fonte foram identificadas com base na sobreposição de conteúdo, total ou parcial.

Ao se optar por um alinhamento baseado na sobreposição de conteúdo ou informação, o processo de indexação não se baseia na sobreposição de formas, ou seja, de unidades lexicais. Consequentemente, sentenças que contêm conteúdo em comum, total ou parcial, com baixa sobreposição lexical (*word overlap*), são alinhadas.

No exemplo do Quadro 4, observa-se que a sentença do sumário e a sentença do texto-fonte apresentam sobreposição parcial de conteúdo. Diz-se "parcial" porque a sobreposição refere-se apenas aos trechos negritados. No caso, o trecho "**se preparando para a passagem do furacão**" do sumário expressa uma informação mais genérica que o trecho "**estocaram alimentos, água, lanternas e velas**" do texto-fonte, já que a "estocagem" pode ser interpretada como uma "espécie de preparação" para a chegada do furacão. Tal sobreposição de conteúdo não seria identificada com base exclusivamente nas unidades lexicais, pois as sentenças em questão não apresentam palavras de conteúdo (nome, verbo, adjetivo e advérbio) em comum.

**Quadro 4:** Exemplo de alinhamento com base na sobreposição de conteúdo.

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

A partir das diretrizes gerais, realizou-se, antes do alinhamento propriamente dito, uma fase de treinamento, na qual 2 coleções foram aleatoriamente selecionadas e alinhadas por cada um dos anotadores, individualmente. Na sequência, os alinhamentos foram comparados e os casos de divergência foram discutidos com o intuito de ajustar a concordância entre os linguistas computacionais. Desse treinamento, algumas regras gerais e específicas foram elaboradas, as quais passaram por um processo de refinamento ao longo da indexação. Assim, ao final, gerou-se um manual de alinhamento de sumários humanos multidocumento e textos-fonte, cujas regras são apresentadas na subseção 3.2.

## 3.2. As regras de alinhamento

Ao todo, foram criadas 8 regras, sendo 4 gerais e 4 específicas. A seguir, tais regras são descritas e exemplificadas por alinhamentos reais do CSTNews, destacando-se, sobretudo, os critérios linguísticos que subsidiaram a formulação das mesmas. Para tanto, as sentenças dos documentos ou textos-fonte são referenciadas pela abreviação SD e as sentenças dos sumários são referenciadas pela abreviação SS.

### 3.2.1. Gerais

**REGRA 1:** *Alinhar com base na sobreposição de conteúdo e não de forma*
Essa regra estabelece que o alinhamento seja feito em função da sobreposição de conteúdo entre uma SS e uma ou mais SDs e não em função da ocorrência de unidades lexicais comuns ou mesmo estruturas sintáticas semelhantes. Consequentemente, sentenças que veiculam certo conteúdo em comum por meio de expressões linguísticas (superficiais) diferentes devem ser alinhadas. Além disso, ressalta-se que o conteúdo em comum nem sempre é identificado diretamente, mas sim por meio de inferências.

Em (1) e (2), apresentam-se exemplos de alinhamento com base na Regra 1. No caso de (1), as sentenças compartilham o mesmo conteúdo principal, ou seja, "o número de mortos no acidente aéreo", expresso de forma diferente. Em (2), o alinhamento se deve à inferência, por meio da expressão "*abrir(ão) mão*", de que a SS focaliza o mesmo conteúdo central da SD, no caso, as "*renúncias*" de deputados denunciados.

(1) **SS:** 17 pessoas morreram após a queda de um avião na República Democrática do Congo.
**SD:** Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas. (D2_C1)

(2) **SS:** A expectativa de lideranças da Câmara e do Conselho de Ética é que pouco mais de 10% dos 69 deputados denunciados no relatório parcial da CPI **abrirão mão de seus mandatos**.
**SD:** **As renúncias** têm que ser publicadas até terça-feira, quando o presidente do Conselho de Ética, deputado Ricardo Izar (PTB-SP), vai instaurar os processos de perda de mandato contra os 69 deputados acusados pela CPI dos Sanguessugas de envolvimento com a máfia das ambulâncias. (D1_C16)

**REGRA 2:** *Alinhar com base na sobreposição da informação principal*
Essa regra estabelece que o alinhamento seja feito em função do conteúdo principal veiculado pelas sentenças. Assim, uma SS é alinhada a SDs quando há sobreposição da ideia central, expressa pelo verbo principal.

Em (3), apresenta-se um exemplo em que as sentenças não foram alinhadas em função da Regra 2, apesar da sobreposição dos sujeitos. Os verbos principais "*descobrir*" e "*informar*, que expressam a informação principal de cada sentença, não são similares.

(3) **SS:** Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, **descobriram** um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro, denominado Oph 162225-240515, o primeiro planemo duplo.

**SD:** Os pesquisadores Ray Jayawardhana e Valentin D. Ivanov **informam** a descoberta na edição de quinta-feira do serviço online Science Express, mantido pela revista Science. (D1_C7)

**REGRA 3:** *Alinhar com base na sobreposição de informação secundária*

Essa regra especifica que as sentenças sejam alinhadas diante da sobreposição de conteúdo ou informação secundária. Assim, uma SS deve ser alinhada a uma ou mais SDs não somente pelo conteúdo principal, mas também pelo compartilhamento de informação periférica.

Em (4) e (5), apresentam-se alinhamentos que ilustram a aplicação da Regra 3. Em (4), por exemplo, a SS e a SD foram alinhadas porque compartilham a informação secundária expressa pelos trechos "*giram um ao redor do outro*" e "*giram em torno um do outro*", apesar de não haver sobreposição do conteúdo central. Em (5), o alinhamento se deve ao fato de que a SS e a SD compartilham a causa ("*pagamento de despesas pessoais*") do fato principal veiculado pela SS ("*Renan é alvo de um processo por quebra de decoro*").

(4) **SS:** Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.

**SD1:** Ambos os mundos têm massa semelhante à de outros exoplanetas já catalogados, mas não giram em torno de uma estrela - na verdade, **giram em torno um do outro**. (D1_C7)

(5) **SS:** Renan é alvo de um processo por quebra de decoro acusado de receber recursos da construtora Mendes Junior para **pagamento de despesas pessoais, como aluguel e pensão para a jornalista Mônica Veloso**, com quem tem uma filha.

**SD:** Isso permitiria que os peritos da Polícia Federal pudessem trabalhar durante o período de descanso dos senadores e, no retorno das férias, apresentarem um relatório detalhado sobre o conjunto de documentos - notas fiscais, recibos de vacinação, extratos bancários, guias de transporte de animais - que o senador

apresentou para justificar o **pagamento da pensão informal à jornalista Mônica Veloso**. (D3_C43)

**REGRA 4:** *Alinhar todas as sobreposições de um mesmo conteúdo*
Essa regra estabelece que uma SS seja alinhada sempre que uma SD com sobreposição de conteúdo for identificada, mesmo que a SS já tenha sido alinhada devido ao compartilhamento desse mesmo conteúdo.

Em (7), ilustra-se a aplicação da Regra 4. No caso, a SS, já alinhada a uma sentença do texto-fonte D1 em função do compartilhamento de informação secundária (cf. (4)), alinha-se novamente a duas sentenças distintas do texto-fonte D2 da mesma coleção, posto que a sobreposição de conteúdo fora novamente identificada.

(7) **SS:** Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.

**SD:** Astrônomos do Observatório Europeu Austral, localizado no Chile, anunciaram a descoberta de uma dupla de planetas errantes (sem estrela-mãe) que **giram ao redor deles mesmos** e que vagam livremente pelo espaço. (D2_C7)

**SD:** O fato extraordinário é que **ele não gira em volta de uma estrela, mas em torno de outro corpo frio** com o dobro de sua massa. (D2_C7)

### 3.2.2. Específicas

O conjunto de regras específicas é composto efetivamente por 4 normas, as quais foram elaboradas em função de casos particulares de sobreposição de conteúdo.

A Regra 5, em especial, foi formulada para lidar com um caso específico de contradição, a saber: uma SS e uma ou mais SDs expressam basicamente o mesmo conteúdo, mas diferem quanto a dados numéricos referentes a um mesmo fato.

As demais regras foram formuladas para os casos em que uma SS e uma ou mais SDs expressam o mesmo conteúdo principal, mas diferem quanto ao: (i) grau de generalização (ou especificação) de uma mesma informação (Regra 6) e (iii) grau de assertividade do falante sobre um mesmo fato (Regra 7). A Regra 8, em especial, é o único caso em que, apesar da similaridade de conteúdo, as sentenças, não devem ser alinhadas.

**REGRA 5:** *Alinhar com base na sobreposição da informação principal mesmo diante de dado numérico contraditório*
Essa estabelece que dada SS deva ser alinhada a uma ou mais SDs em função da sobreposição da ideia central mesmo diante de dados numéricos contraditórios, os quais podem, por exemplo, ser referentes à hora de ocorrência de determinado fato.

Em (8), o exemplo em questão ilustra um alinhamento feito com base a Regra 5. Especificamente, a SS e a SD compartilham o conteúdo principal, no caso, ambas registram o fato de "a cidade de São Paulo apresentar pontos de alagamento", mas apresentam informação contraditória sobre o horário em que de tal fato foi observado/registrado. Diante de contradições desse tipo, as sentenças são alinhadas.

(8) **SS:** Às **9h**, a cidade tinha oito pontos de alagamento, sendo dois intransitáveis.
**SD:** O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às **9h30** desta segunda-feira.

**REGRA 6:** *Alinhar com base na sobreposição da informação principal mesmo diante de diferentes graus de generalização*

A Regra 6 prevê que uma SS deva ser alinhada a uma ou mais SDs em função da sobreposição da ideia central mesmo que essa informação seja apresentada com graus de generalização distintos.

Em (9), observa-se que o alinhamento se dá em função do compartilhamento da informação principal entre a SS e a SD, no caso, "o índice de congestionamento (na cidade de São Paulo) acima da média", apesar de a SS especificar essa informação ao registrar (i) as extensões exatas do congestionamento em quilômetros e (ii) os horários de registro dessas extensões.

Em (10), as sentenças do sumário em questão foram alinhadas à do texto-fonte pela Regra 6 tendo em vista que a SS(2) e a SS(3) apresentam informações mais específicas que a SD. No caso, a SS(2) e a SS(3) especificam em porcentagem a "intensificação da fiscalização", conteúdo principal de SD.

Quanto à (11), ressalta-se que é a SS que contém a informação mais genérica, ao passo que as SDs dos documentos D2 e D3 contêm a informação mais específica.

(9) **SS:** A Companhia de Engenharia de Tráfego (CET) anunciou que o índice de congestionamento era de **54 quilômetros** às 8h, **113 km** às 9h e **110 km** meia hora depois, valores bem acima das médias para os horários, que eram de **36, 82 e 76 quilômetros** respectivamente, mas não havia registro de acidentes graves, apesar de haver feridos.
**SD:** Com o asfalto molhado, o trânsito ficou mais lento e **o congestionamento ficou o dobro da média.** (D3_C4)

(10) **SS(2):** O balanço divulgado mostra que as autuações **cresceram 316,5%** nos sete primeiros meses deste ano e chegaram a R$ 1,339 bilhão.
**SS(3):** Foram autuados 208.471 contribuintes, um **crescimento de 104,47%** em relação ao mesmo período do ano passado.
**SD:** BRASÍLIA - A Receita Federal **intensificou a fiscalização** e o resultado foi um aumento do número de contribuintes que caíram na malha fina (D2_C34).

(11) **SS:** A Receita Federal **intensificou a fiscalização** sobre as declarações das pessoas físicas neste ano.

**SD:** Balanço da fiscalização, divulgado nesta segunda-feira pela Receita mostra que as autuações cresceram 316,5% nos sete primeiros meses deste ano e **chegaram a R$ 1,339 bilhão**. (D2_C34).

**SD:** O volume de recursos recolhido com multas **passou de R$ 326,1 milhões para R$ 1,339 bilhão**. (D3_C34)

**REGRA 7:** *Alinhar sentenças com sobreposição da informação principal e diferença no grau de assertividade*

A Regra 7 prevê que uma SS deva ser alinhada a uma ou mais SDs em função da sobreposição da ideia central mesmo que tais sentenças apresentem diferentes graus de assertividade do falante com relação à informação principal que está sendo veiculada. No exemplo em (12), as sentenças em questões foram alinhadas devido à sobreposição da informação central (no caso, "a autoria das ações criminosas"), mesmo verificando-se que a SS apresenta maior grau de assertividade do falante quanto ao fato principal que a SD. Na SD, o menor grau de assertividade é identificado pela ocorrência do verbo auxiliar modal "poder" na SD ("**podem ter sido ordenadas"**).

(12) **SS: As ações são atribuídas à facção criminosa Primeiro Comando da Capital (PCC)**, que já comandou outros ataques em duas ocasiões.

**SD: As ações criminosas podem ter sido ordenadas pelos líderes do Primeiro Comando da Capital (PCC)**, que haviam prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo.

**REGRA 8:** *Não alinhar sentenças com sobreposição da informação principal quando uma expressar um todo e a outra uma parte do todo.*

Essa regra prevê que uma SS e uma ou mais SDs não devam ser alinhadas caso haja diferença de intensidade ou quantidade referente à informação principal comum a elas.

Em (13), ilustra-se um caso em que as sentenças não foram alinhadas com base na Regra 8. Em (13), verifica-se que a SS e a SD apresentam informação principal similar, no caso, "internação do senador". No entanto, a SS apresenta um sintagma adverbial que indica a quantidade de vezes em que o fato principal ocorreu, no caso, "**por três vezes**". A SD, por sua vez, apresenta o sintagma adverbial "**em abril**", que indica a pontualidade da "internação". Assim, vê-se que a SS expressa a repetição de um mesmo fato (sequência), ao passo que a SD descreve uma das ocorrências do fato, resultando no não alinhamento das sentenças.

(13) **SS:** Somente neste ano, o senador **se internou por três vezes** no Incor.

**SD: Em abril, o senador foi internado no InCor** com insuficiência cardíaca.

Na Seção 4, apresentam-se os resultados do alinhamento dos sumários humanos multidocumento e dos textos-fonte que compõem as coleções do *corpus* jornalístico CSTNews segundo as regras descritas nesta Seção. Os resultados englobam especificamente a quantificação dos diferentes tipos de alinhamento e a exemplificação de alguns dos casos.

## 4. Resultados

De um modo geral, destaca-se que, das 331 sentenças que compõem a coleção dos 50 sumários do *corpus* CSTNews, 230 (aproximadamente 70%) foram alinhadas a mais de uma sentença dos textos-fonte. Tal fato justifica-se por se tratar de sumários multidocumento, ou seja, versões condensadas de coleções de textos que se estendem sobre um mesmo assunto ou tema. Além disso, todas as sentenças dos sumários foram alinhadas, exceto duas delas. O não alinhamento de tais sentenças justifica-se pelo fato de que ambas apresentam informação que não está efetivamente presente nos textos-fonte, tendo sido inseridas nos sumários por meio de inferência feita pelos humanos produtores dos resumos. Um desses casos de não alinhamento ocorreu na coleção C25 do CSTNews, que pertence à categoria "esporte", e cujos textos versam sobre "jogos das seleções brasileiras de vôlei e futebol durante o pan-americano de 2007". No sumário humano multidocumento da C25, a sentença "**Neste domingo, o esporte brasileiro alegrou a torcida verde-amarelo**" não foi alinhada a nenhuma sentença dos 3 documentos que compõem a coleção, pois a informação nela contida não está explícita nos textos-fonte, tendo sido inferida ou deduzida pelo humano produtor do sumário.

Quanto aos textos-fonte, ressalta-se que, do total de 2067 sentenças que compõem as 50 coleções, 877 (42,43%) foram alinhadas, sendo que a mesma sentença de um sumário pode ter sido alinhada a mais de uma sentença dos textos-fonte. Todos os tipos de alinhamento resultantes podem ser vistos na Tabela 2 e na Figura 2.

**Tabela 2:** Quantificação numérica dos tipos de alinhamento.

| **Quant.** | **Tipos de alinhamento** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1-0 | 1-1 | 1-2 | 1-3 | 1-4 | 1-5 | 1-6 | 1-7 | 1-8 | 1-9 | 1-10 | 1-11 | 1-12 |
| | 2 | 71 | 90 | 67 | 36 | 37 | 13 | 5 | 5 | 1 | 1 | 2 | 1 |

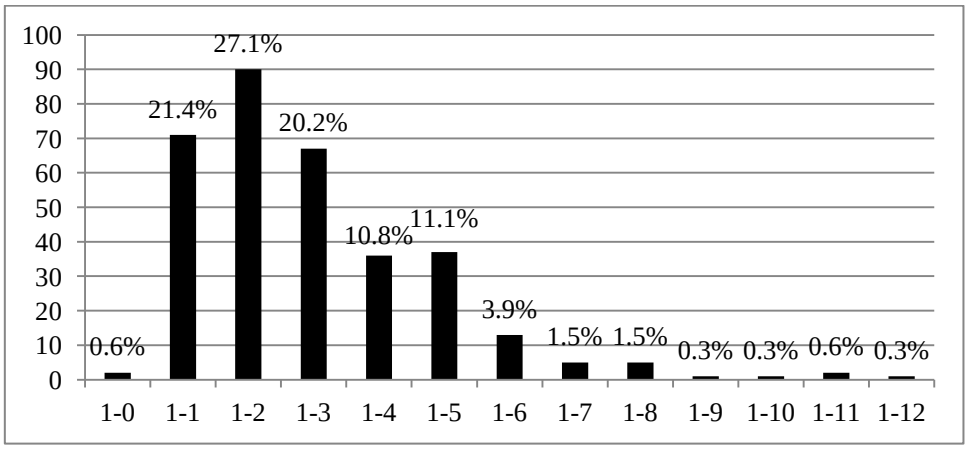


**Figura 1:** Quantificação percentual dos tipos de alinhamento.

A quantidade e porcentagem de sentenças alinhadas por coleção estão na Tabela 3.

**Tabela 3:** Quantidade numérica e percentual das sentenças alinhadas por coleção.

| Coleção (C) | Nº de sentenças por C | Nº de sentenças alinhadas | % de sentenças alinhadas |
|---|---|---|---|
| **C1** | 24 | 16 | 66,66 |
| **C2** | 51 | 22 | 43,13 |
| **C3** | 50 | 18 | 36 |
| **C4** | 39 | 16 | 41,02 |
| **C5** | 23 | 11 | 47,82 |
| **C6** | 36 | 18 | 50 |
| **C7** | 23 | 12 | 52,17 |
| **C8** | 25 | 16 | 64 |
| **C9** | 36 | 18 | 50 |
| **C10** | 38 | 19 | 50 |
| **C11** | 56 | 22 | 39,28 |
| **C12** | 34 | 15 | 44,11 |
| **C13** | 37 | 18 | 48,64 |
| **C14** | 25 | 19 | 76 |
| **C15** | 26 | 17 | 65,38 |
| **C16** | 47 | 16 | 34,04 |
| **C17** | 41 | 13 | 31,70 |
| **C18** | 70 | 25 | 35,71 |
| **C19** | 13 | 7 | 53,84 |
| **C20** | 42 | 15 | 35,71 |
| **C21** | 41 | 7 | 17,07 |
| **C22** | 50 | 15 | 30 |
| **C23** | 25 | 12 | 48 |
| **C24** | 24 | 13 | 54,16 |
| **C25** | 88 | 31 | 35,22 |
| **C26** | 58 | 28 | 48,27 |
| **C27** | 89 | 43 | 48,31 |
| **C28** | 35 | 17 | 48,57 |
| **C29** | 48 | 13 | 27,08 |
| **C30** | 46 | 16 | 34,78 |
| **C31** | 10 | 7 | 70 |
| **C32** | 66 | 29 | 43,93 |
| **C33** | 68 | 29 | 42,64 |
| **C34** | 59 | 29 | 49,15 |
| **C35** | 36 | 18 | 50 |
| **C36** | 74 | 25 | 33,78 |
| **C37** | 26 | 10 | 38,46 |
| **C38** | 26 | 10 | 38,46 |
| **C39** | 34 | 13 | 38,23 |
| **C40** | 28 | 10 | 35,71 |
| **C41** | 45 | 14 | 31,11 |
| **C42** | 39 | 12 | 30,76 |
| **C43** | 49 | 12 | 24,48 |
| **C44** | 26 | 17 | 65,38 |
| **C45** | 47 | 22 | 46,80 |
| **C46** | 23 | 10 | 43,47 |
| **C47** | 43 | 11 | 25,58 |
| **C48** | 43 | 22 | 51,16 |
| **C49** | 23 | 19 | 82,60 |
| **C50** | 62 | 30 | 48,38 |
| **Totais** | 2067 | 877 | 42,43 |

Com base na Tabela 2 e na Figura 2, observa-se que: (i) 2 sentenças dos sumários não foram alinhadas (1-0), o que resulta da inserção de informação no sumário que não está presente nos textos-fonte; (ii) 71 sentenças dos sumários foram alinhadas a 1 sentença dos textos-fonte (1-1); (iii) 90 sentenças dos sumários foram alinhadas a 2 sentenças dos textos-fonte (1-2), e assim por diante.

A seguir, ilustra-se o único caso de alinhamento do tipo 1-12. Esse alinhamento foi feito na coleção C27, pertencente à categoria "esporte" e composta por 3 notícias que relatam "a goleada aplicada pela seleção brasileira de futebol sobre o Equador pelas eliminatórias para a Copa do Mundo-2010". No caso, a sentença do sumário "O jogo contou com belas atuações de craques como Ronaldinho e Kaká" foi alinhada, no total, a 12 sentenças distintas que compõem os textos-fonte. Esses alinhamentos foram feitos basicamente em função da Regra 6, pois a SS e as 12 SDs compartilham a mesma ideia central com diferentes graus de generalização. No caso, todas as 12 SDs apresentam detalhes sobre a informação generalizada na SS.

Na Figura 2, apresenta-se o caso de alinhamento do tipo 1-12 em questão.

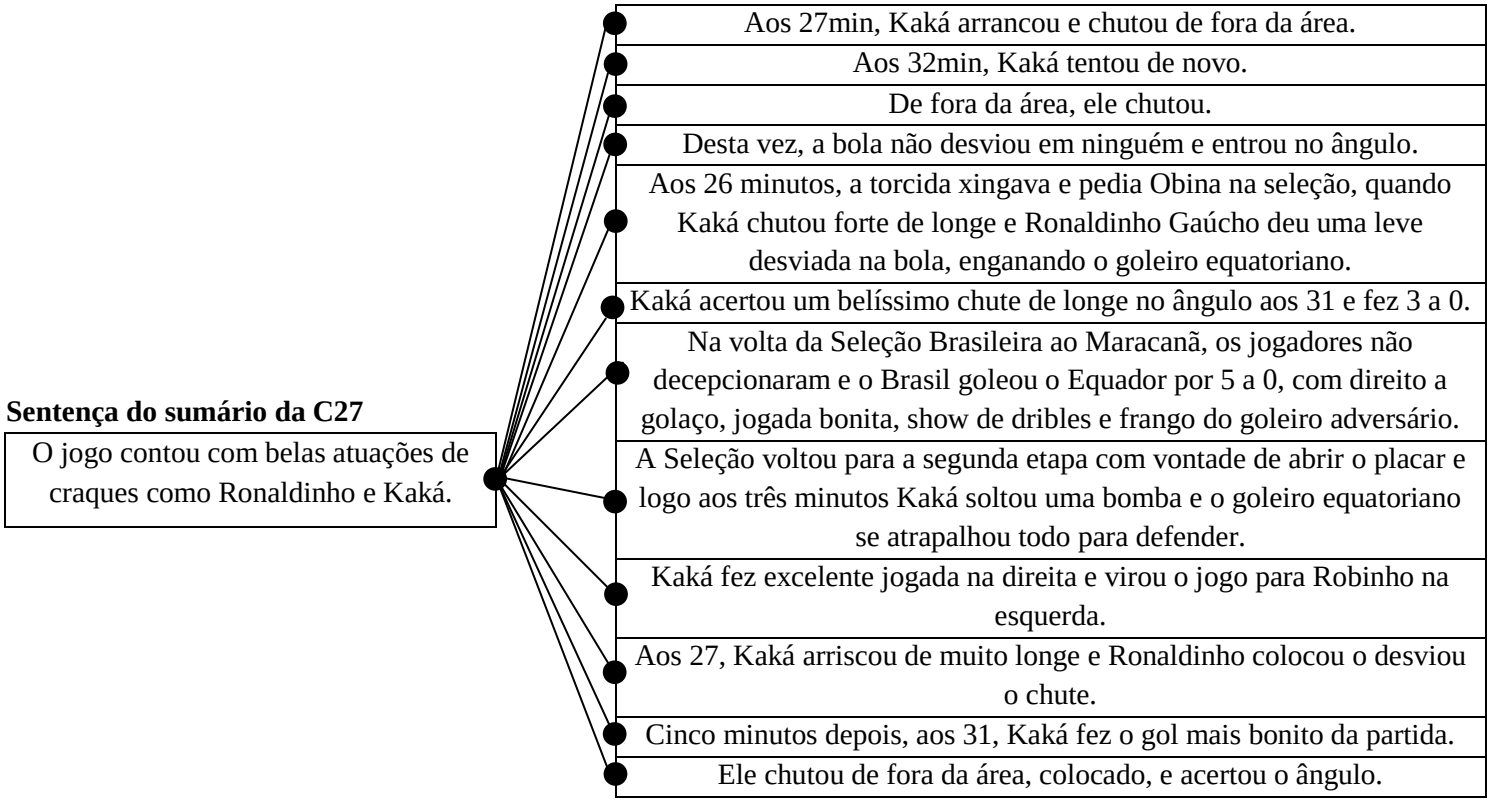


**Figura 2:** Exemplo de alinhamento do tipo 1-12.

Outro exemplo de alinhamento um-para-muitos foi identificado na coleção C4, pertencente à categoria "cotidiano" e composta por 3 notícias que relatam o fato de "a cidade de São Paulo apresentar pontos de alagamento". No caso, a sentença 4 do sumário foi alinhada, no total, a 10 sentenças dos textos-fonte em função basicamente da Regra 6, assim como no exemplo anterior. A Figura 3 ilustra o alinhamento 1-10.

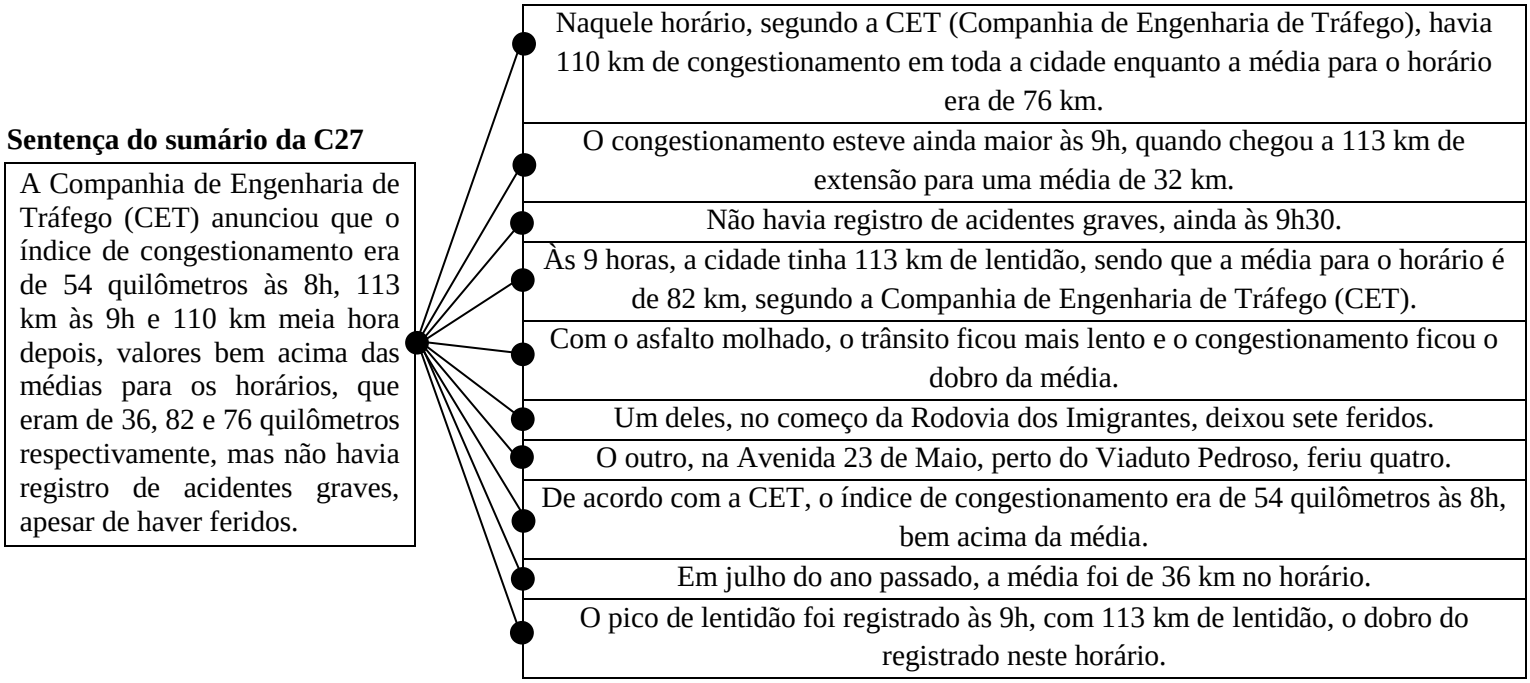


**Figura 3.** Exemplo de alinhamento do tipo 1-10.

A confiabilidade dos alinhamentos ora descritos foi calculada por meio da medida de concordância *kappa* (Cohen, 1960; Carletta, 1996). O valor dessa medida varia de 0 a 1, sendo que 0 indica a não concordância entre os anotadores e 1 indica total concordância entre eles. Com o intuito de calcular a concordância, selecionou-se aleatoriamente 1 coleção por semana durante o período que englobou as últimas 5 semanas de anotação. No total, 5 coleções rotuladas por "mundo", "esporte", "dinheiro", "cotidiano" e "política" foram utilizadas nessa tarefa. Especificamente, a cada semana, os pesquisadores alinhavam individualmente uma dessas coleções e comparavam os resultados de cada alinhamento para verificar a concordância. A concordância *kappa* resultante foi de 0.831, valor alto, o que pode significar que a tarefa de alinhamento é bem definida e relativamente pouco subjetiva. Vale ressaltar que, da comparação entre os alinhamentos individuais, gerava-se um terceiro alinhamento, resultante do consenso entre os pesquisadores, o qual passava a ser efetivamente o alinhamento oficial de cada um das 5 coleções utilizadas para a concordância.

Quanto à forma de disponibilização, o alinhamento ora apresentado segue o formato de anotação XML (*Extensible Markup Language*). No Quadro 5, o alinhamento do sumário multidocumento e dos textos-fonte da coleção C31 do CSTNews está representado em XML. No esquema em questão, existem três blocos de codificação, um para cada sentença do sumário. Esses blocos são delimitados por "<align 'número da sentença'>" e "</align>". O primeiro bloco do esquema XML descreve o alinhamento da sentença 1 do sumário (de <align SENT="1"> até </align>). Nesse bloco, a sentença 1 (SENT="1") foi alinhada à: (i) SENT="1" do D(ocumento)1, (ii) SENT="1" do D2 e (iii) SENT="2" do D2. Além da informação sobre a sentença e o texto-fonte, o esquema XML prevê a especificação do tipo de alinhamento (TYPE="none") e do anotador (JUDGE="veronica").

**Quadro 5:** Exemplo da representação em XML do alinhamento.

```
<align SENT="1">
        <DOC="D1_C31_Folha.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
</align>
<align SENT="2">
        <DOC="D1_C31_Folha.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="6" TYPE="none" JUDGE="veronica"/>
</align>
<align SENT="3">
        <DOC="D1_C31_Folha.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
</align>
```

## 5. Considerações finais

Neste relatório técnico, descreveu-se o processo de alinhamento dos sumários humanos multidocumento e de seus respectivos texto-fonte presentes no *corpus* jornalístico CSTNews. O alinhamento em questão foi realizado manualmente por dois anotadores. Como resultado, disponibiliza-se o referido alinhamento para que futuras pesquisas linguístico-computacionais possam ser realizadas a partir do novo tipo de anotação do CSTNews.

Quanto às dificuldades encontradas nesta tarefa, destaca-se a necessidade de conhecimento de domínio para realizar o alinhamento dos sumários e dos textos de certas coleções, sobretudo dos que compõem as coleções da categoria "política".

Como trabalho futuro, destaca-se a possibilidade de incluir, no esquema XML de cada coleção, a informação sobre o tipo do alinhamento. Atualmente, ao atributo TYPE que compõe os esquemas XML de cada alinhamento, está associado o valor nulo "none". Esse atributo, no entanto, pode ser preenchido, por exemplo, com a informação de que se trata de um alinhamento, por exemplo, com contradição (cf. Regra 5) ou mesmo com grau distinto de assertividade (cf. Regra 6) ou generalização (cf. Regra 7). Com a especificação do tipo de alinhamento, será possível determinar quais os tipos que um alinhador automático acerta com mais frequência.

## Referências


Barbosa, J. P. Notícia (Coleção trabalhando com os gêneros do discurso: relatar). São Paulo: FTD, 2001

Cardoso, P.C.F.; Maziero, E.G.; Castro Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105. October 26, Cuiabá/MT, Brazil.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, v. 22, n. 2, pp. 249-254.

Caseli, H.M. (2003), "Alinhamento sentencial de textos paralelos português-inglês". School: Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP), Fevereiro, 2003.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, v. 20, n. 1, pp. 37-46.

Dolz, J.; Schneuwly, B. *Gêneros orais e escritos na escola*. Trad. Roxane Rojo e Glaís Sales Cordeiro. São Paulo: Mercado de Letras, 2004.

Gale, W.A. and Church, K.W. (1991). A program for aligning sentences in bilingual corpora. In the *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berkley, pp. 177-184.

Gale, W.A. and Church, K.W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, v. 19, n. 3, pp. 75-102.

Hirao, T.; Suzuki, J.; Isozaki, H.; Maeda, E. (2004). Dependency-based Sentence Alignment for Multiple Document Summarization. In the *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*, pp. 446-452.

Lage, N. Linguagem jornalística. 7 ed. São Paulo: Editora Ática, 2004

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Marcu, D. (1999). The automatic construction of large-scale corpora for summarization research. In the *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, pp. 137–144.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross- document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, pp. 74-83.

Soricut, R. and Brill, E. (2004). Automatic Question Answering: Beyond the Factoid. In the *Proceedings of HLT-NAACL*, pp. 57-64.

Specia, L. (2010). Translating from Complex to Simplified Sentences. In the *Proceedings of PROPOR*, pp. 30-39.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In the *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 523-530. Toulouse, France, July.