# ANOTAÇÃO DE ERROS EM SUMÁRIOS AUTOMÁTICOS

Vários autores (Friedrich et al., 2014; Kaspersson et al., 2012; Pitler et al., 2010) citam que os principais erros em sumários automáticos são relacionados a menções de entidades e violações de não gramaticalidade e redundância (*clause level*). Além desses erros, acredita-se que a existência de informações contraditórias também afeta negativamente os sumários. A seguir, apresentam-se os principais erros de acordo com a literatura.

## I. Erros relacionados a menções de entidades

**Primeira menção sem explicação** (1M-EXP): é atribuída a primeira menção de uma entidade nomeada para a qual falta uma referência clara para o leitor. Para identificar esse erro, não deve ser utilizado conhecimento de mundo. Se aparecer, por exemplo, *Itaú*, sem dizer que é um banco, este deve ser marcado como 1M-EXPL. No exemplo (1), não se sabe o que é Tepco, e, no exemplo (2), falta definição do que é Itaú.

(1) *A* <e TYPE=1M-EXP>*Tepco*</e> *inicialmente declarou que o tremor não havia causado vazamentos, mas, mais tarde, revelou que 1.200 litros de água com materiais radioativos da usina haviam vazado para o mar.*

(2) *Em comparação com a receita obtida nos seis primeiros meses de 2006, de R$ 2,958 bilhões, o lucro do* <e TYPE=1M-EXP>*Itaú*</e> *cresceu 36% neste ano.*

**Menções subsequentes com explicação** (nM+EXP): aplica-se para menções de entidades nomeadas que, mas ainda aparecem com uma introdução explicativa inapropriada. No exemplo (3), explica-se novamente na segunda sentença quem é Leomar Quintanilha. O mesmo ocorre com o exemplo (4), explica-se novamente o que é CET na segunda sentença.

(3) *O presidente do Conselho de Ética do Senado, Leomar Quintanilha (PMDB-TO), disse hoje ser contrário à unificação dos processos contra o senador Renan Calheiros (PMDB-AL) que tramitam na Casa Legislativa.*
<e TYPE=nM+EXP SENT=S3 TEXT= "*O presidente do Conselho de Ética do Senado, Leomar Quintanilha (PMDB-TO)*" > *O presidente do conselho, Leomar Quintanilha (PMDB-TO)*</e>*, disse que é contra a união das representações, mas que vai colocar a proposta em votação.*

(4) *Às 9 horas, a cidade tinha 113 km de lentidão, sendo que a média para o horário é de 82 km, segundo a Companhia de Engenharia de Tráfego (CET).*
*Naquele horário, segundo* <e TYPE=nM+EXP SENT=S4 TEXT= "a Companhia de Engenharia de Tráfego (CET)"> *a CET (Companhia de Engenharia de Tráfego)* </e>*, havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km.*

**Sintagma Nominal Definido sem Referência a Menções Anteriores** (SNdef-REF): Sintagmas Nominais Definidos são geralmente usados no texto para referirem a entidades que já estão presentes no contexto do discurso. Assim, marcam-se os Sintagmas Nominais Definidos que violam esta regra. No exemplo (5), o erro está na última sentença, na qual porta-voz aparece como uma unidade definida.

(5) *Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.*
*Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.*
*<e TYPE=SNdef-REF>O porta-voz</e> informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.*

**Sintagma Nominal Indefinido com Referência a Menções Anteriores** (SNind+REF): Sintagmas Nominais Indefinidos são usados para introduzir novas entidades no discurso. Assim, marcam-se os Sintagmas Nominais Indefinidos com Referencia a Menções Anteriores que violam esta regra. No exemplo (6), o erro está ao chamar um Airbus A320, pois se trata de uma entidade já definida.

(6) *O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13.*
*O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, <e TYPE=SNind+REF SENT=S6 TEXT= "O Airbus-A320"> um Airbus A320</e>, continuou voando, com o reverso direito desligado.*

**Pronome sem Antecedente** (PRO-ANT): Este erro ocorre quando um pronome não tem antecedente sintaticamente possível no sumário, ou seja, não há antecedente que combina em gênero e número. No exemplo (7), o pronome ele aparece na primeira sentença do sumário, sendo impossível saber de quem se fala.

(7) *Internado em um hospital em Buenos Aires, <e TYPE=PRO-ANT> ele </e> teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.*
*"Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".*

**Pronomes com antecedentes Enganosos** (PRO_ENG): ocorre quando uma expressão anafórica refere-se a um antecedente enganoso e seu antecedente correto não está presente no texto. No caso de sumários, pode ser necessário consultar o texto-fonte para identificação do antecedente correto. No exemplo (8), o pronome ele (segunda sentença) parece conectar-se a entidade Kaká (primeira sentença), mas, no texto-fonte, o pronome refere-se ao jogador Robinho que não aparece no sumário.
Além de identificar o tipo de erro, é importante deixar explícito o antecedente enganoso, usando o atributo ANT (antecedente) e colocando entre aspas o antecedente. Quando houver mais de um antecedente enganoso, eles devem aparecer separados por vírgula.

(8) *Aos 27, **Kaká** arriscou de muito longe e **Ronaldinho** colocou o desviou o chute.*
*A 20cm da linha de fundo <e TYPE=PRO_ENG ANT="Kaká, Ronaldinho"> ele </e> deu dois dribles humilhantes no zagueiro equatoriano e cruzou para Elano, que fez o quarto, aos 37.*

**Acrônimos sem Explicação** (ACR-EXP): marcam-se todos os acrônimos que não foram explicados no sumário. Nos exemplos (9) e (10), consideram-se acrônimos sem explicação Deic e PF.
Alguns acrônimos são senso comum, tais como siglas de estados e partidos. Esses casos devem ser identificados com o atributo SC e colocar entre aspas o significado do acrônimo, conforme se vê na anotação do exemplo (10).

(9) *O outro suspeito tem 27 anos, é grafiteiro e, segundo o <e TYPE=ACR-EXP>Deic</e>, tem passagem por roubo, mas já cumpriu a pena.*

(10) *A <e TYPE=ACR-EXP SC="Polícia Federal"> PF </e> não soube informar se esse tipo de recompensa é paga para órgãos policiais.*

II. Erros relacionados a violações de gramaticalidade e redundância

**Informação redundante** (RED): a existência de informações redundantes (total ou parcial) afetam negativamente os sumários.

(11) *Na segunda etapa, a parte concluída será reaberta e a obra passará a ser feita na outra cabeceira.*
*<e TYPE=RED SENT=S11>Na segunda parte, a outra cabeceira será reformada e, na terceira etapa, o centro da pista será reformado.</e>*

(12) *Uma bomba caseira foi jogada contra o prédio do Ministério Público, no centro da capital, mas não deixou feridos.*
*<e TYPE=RED SENT=S12>Uma bomba de fabricação caseira explodiu em frente ao prédio do Ministério Público Estadual e lojas vizinhas também foram atingidas por estilhaços.</e>*

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

(14) *A Receita Federal intensificou a fiscalização* e o resultado foi um aumento do número de contribuintes que caíram na malha fina.

(15) *<e TYPE=RED SENT=S13,S14>Dobrou o número de pessoas físicas autuadas</e>* depois de cair na malha fina até julho, de acordo com a Receita Federal do Brasil.
A expectativa da Receita é que até o final do ano mais de 300 mil contribuintes sejam autuados pela malha fina.

**Contradição** (CONTR): ocorre quando duas sentenças apresentam informações conflitantes.

(16) *O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que **57 pessoas morreram e 128 ficaram feridas** no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.*
*<e TYPE=CONTR SENT=S16> Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas.</e>*

**Sentenças incompletas** (SENT_INC): pode ocorrer na forma de sentenças incompletas, falta de sinais de pontuação ou espaços. Nesse caso, o erro se aplica a toda sentença. No exemplo (15), a última sentença está incompleta (terminando com uma vírgula).

(17) *Como esperado, a atleta Fabiana Murer conquistou a medalha de ouro no salto com vara nos Jogos Pan-Americanos do Rio, nesta segunda-feira, no Estádio João Havelange.*
*<e TYPE=SENT_INC>Murer conquistou o lugar mais alto do pódio com a marca de 4m60, contra 4m40 da norte-americana April Steiner,</e>*

**Sem relacionamento semântico** (SEM_REL): marcam-se sentenças adjacentes que não possuem qualquer relacionamento semântico, i.e., casos em que o leitor imagina o que a sentença x tem a ver com a sentença y. No exemplo (16), não é possível entender porque o candidato Lula foi a pista de dança.

(18) *Após um fim de semana no Norte e Nordeste ao lado de caciques pefelistas adeptos de uma campanha mais ofensiva e com discursos duros contra o presidente Luiz Inácio Lula da Silva, o candidato do PSDB à Presidência, Geraldo Alckmin, deixou ontem a linha "paz e amor" e se curvou à temperatura alta do debate eleitoral.*
*Alckmin acusou Lula de arrogante, de subestimar a inteligência dos brasileiros e relacionou o presidente aos escândalos do mensalão, sanguessuga e ao caso Waldomiro Diniz.*
*No mesmo dia em que Lula se comprometeu a não atacar, o adversário tucano Geraldo Alckmin elevou o tom do seu discurso.*
*Sem citar nominalmente o adversário, Alckmin criticou de novo Lula ao comentar especulações de que o petista, convicto na vitória no primeiro turno, já estaria fazendo planos sobre sua nova equipe ministerial.*
<e TYPE=SEM_REL> Após comer, o candidato foi até a pista de dança e, ao som de Alcione, foi disputado pelas senhoras da velha guarda da escola.</e>
*"Não sou nenhum expert, mas gosto de dançar. Conheci a Lu (sua mulher) assim, num baile em Pinda (Pindamonhangaba, sua cidade natal)".*

**Conectivo/marcador discursivo sem contexto apropriado** (MD): esse erro ocorre em sentenças que possuem marcadores discursivos explícitos (‘mas’, porque’, ‘porém’) que não são mais apropriados no contexto do sumário. Nesses casos, criam-se links entre as sentenças envolvidas e marca-se o conectivo. No exemplo (19), o marcador discursivo contudo não possui ligação com a sentença anteriormente dita.

(19) Em meio ao tráfico de drogas constante, uma praça está quase pronta bem ao lado do fluxo de viciados na cracolândia, na Luz (centro de São Paulo).
<e TYPE=MD CONEC = “Contudo”> *Contudo*, a secretária Municipal de Assistência Social não informou qual seria o prazo de entrega previsto. </e>

III - Outros tipos de erros

**OUTRO**: Caso aconteça algum problema que não está listado em algum dos tipos acima, deve ser anotado com OUTRO, com a explicação do erro no atributo EXPLANATION. O anotador deve colocar a etiqueta OUTRO para a sentença completa ou para um ponto específico da sentença, dependendo do problema.

(20) Além de Rafael Nadal, o torneio contará com mais três atletas classificados entre os 20 melhores do ranking da ATP: o espanhol Nicolás Almagro (11º colocado e tricampeão do Brasil Open), o argentino Juan Mônaco (12º) e o suíço Stanilas Wawrinka (17º).
(21) A organização do **<e TYPE=Outros EXPLANATION="referência em português para termo introduzido em inglês">**Aberto do Brasil 2013**</e>** anunciou na manhã desta terça-feira que o torneio a ser disputado em fevereiro, no ginásio do Ibirapuera, em São Paulo, marcará a volta do espanhol Rafael Nadal às quadras.

# Exemplos de Sumários Anotados

(S1) A lista dos chamados exoplanetas, mundos localizados fora do sistema solar, têm um extraordinário novo membro.
(S2) 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.
(S3)<e TYPE=RED SENT=S2> Usando telescópios do <e TYPE=nM+EXP SENT=S2 TEXT= "Observatório Europeu Austral" > Observatório Europeu Sul (ESO) </e>, <e TYPE=SNind+REF SENT=S2 TEXT= "Astrônomos do Observatório Europeu Austral">astrônomos</e> descobriram um planeta com sete vezes a massa de Júpiter, o mais pesado dos que giram em torno do Sol, e outro, com o dobro desse peso. </e>
(S4) "Este é um par de gêmeos verdadeiramente de destaque, já que cada um tem uma massa de apenas 1% de nosso Sol", declarou <e TYPE=1M-EXP> Jayawardhana</e>.

Figura 1: Exemplo 1

(S1) Em comparação com a receita obtida nos seis primeiros meses de 2006, de R$ 2,958 bilhões, o lucro do <e TYPE=1M-EXP>Itaú</e> cresceu 36% neste ano.
(S2) <e TYPE=nM+EXP>O Itaú, segundo maior banco privado do País,</e> obteve lucro líquido de R$ 4,016 bilhões no primeiro semestre desse ano, superando o rival Bradesco e registrando o maior lucro entre bancos privados nos últimos 20 anos.
(S3) O lucro líquido acumulado de janeiro a junho chegou a R$ 4,016 bilhões, 35,7% acima dos R$ 2,958 bilhões dos primeiros seis meses de 2006 e também superior aos R$ 4,007 bilhões anunciados na véspera pelo Bradesco, líder no ranking de bancos do país.
(S4) As operações de crédito do banco totalizaram R$ 104,82 bilhões em junho, um crescimento de 40% sobre o resultado para o mesmo período no passado.

Figura 2: Exemplo 2

(S1) 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.
(S2) O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.
(S3) <e TYPE=RED SENT=S1> Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.</e>

Figura 3: Exemplo 3

# A tarefa

A tarefa será realizada em grupo e de forma presencial em alguma sala do ICMC/USP, a ser divulgada nas vésperas dos dias de anotação.

Ao todo serão anotados 200 sumários automáticos, dos sistemas baseados em conhecimentos superficial e profundo. Os sistemas cujos sumários serão anotados são RSumm (Ribaldo, 2013), GistSum (Pardo, 2005), RC-4 (Cardoso, 2014) e MTRST-MLAD (Castro, 2015).

Inicialmente, 2 dias serão destinados para treinamento com todos os anotadores para esclarecer os procedimentos para a realização tarefa.

A anotação propriamente dita terá duração de aproximadamente 15-20 dias. Diariamente devem ser anotados 20 sumários, aproximadamente.

Periodicamente, os anotadores anotarão individualmente alguns sumários para fins de cálculo de concordância.

**Material de Trabalho**

- As etiquetas com suas definições e exemplos serão expostas, via projetor, nos dias da anotação;
- Os anotadores também terão em mãos um resumo impresso das etiquetas, com definições e exemplos típicos para cada uma;
- Poderá ser utilizado qualquer editor de texto para a anotação;
- A anotação deverá ser feita e salva no mesmo arquivo recebido por cada anotador, sem modificar o nome e a codificação do mesmo;
- Para cada sumário, um anotador tomará a frente e conduzirá a anotação, de forma que seu raciocínio fique explícito e a anotação possa ser mais afinada.

**Envio dos Sumários**

- Quando necessário, os arquivos devem ser entregues via email (marciosouzadias@gmail.com e paulastm@gmail.com) ou pendrive disponibilizado nos dias da anotação.

**Referências**

Friedrich, A.; Valeeva, M.; Palmer, A. (2014). LQVSumm: A Corpus of Linguistic Quality Violations in Multi-Document Summarization.

Kaspersson, T.; Smith, C.; Danielsson, H.; Jonsson, A. (2012). This also affects the context – errors in extraction based summaries.

Pitler, E.; Louis, A.; Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization.