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

Anotação de Aspectos Textuais em Sumários do Córpus CSTNews

Relatório Técnico do NILC

NILC-TR-13-01

Relatório Técnico do ICMC

Nº 394

Amanda P. Rassi[1], Andressa C. I. Zacarias[1], Erick G. Maziero[2], Jackson W. C. Souza[1], Márcio S. Dias[2], Maria Lúcia R. Castro Jorge[2], Paula C. F. Cardoso[2], Pedro P. Balage Filho[2], Renata T. Camargo[1], Verônica Agostini[2], Ariani Di Felippo[1], Eloize R. M. Seno[3], Lucia H. M. Rino[4], Thiago A. S. Pardo[2]

[1]Departamento de Letras, Universidade Federal de São Carlos
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[3]Instituto Federal de São Paulo
[4]Departamento de Computação, Universidade Federal de São Carlos

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

Resumo

Neste relatório técnico discorre-se sobre a anotação de aspectos textuais nos sumários manuais multidocumento do Córpus CSTNews. Esse córpus foi construído principalmente com vistas à Sumarização Automática Multidocumento. Ele é composto por coleções de textos jornalísticos provenientes de agências de notícias on-line conhecidas do Brasil. Especificamente, cada coleção contém em média 3 textos sobre um mesmo assunto, e cada texto advém de uma agência distinta. A partir do CSTNews, vários subcórpus foram construídos. Um deles é composto pelos sumários manuais elaborados para cada uma das coleções, ou seja, existe um sumário manual multidocumento para cada coleção. A anotação manual de aspectos foi feita para esses sumários multidocumento. Os aspectos em foco nessa anotação indicam diferentes tipos de informação que podem ser veiculados por um texto. Podem, por exemplo, referir-se a papéis semânticos como agente (quem), objeto (o que), modo (como), tempo (quando), etc., e, muitas vezes são dependentes do assunto, ou categoria à qual o texto pertence (p.ex.: esporte, mundo, etc.). Essa anotação dos sumários multidocumento do Córpus CSTNews visa trazer mais informatividade às tarefas automáticas, para melhorar sua qualidade. Para a Sumarização Automática Multidocumento, os aspectos podem indicar estruturas padrão (*templates*) para a modelagem de critérios de seleção e organização do conteúdo nos sumários.

Este trabalho conta com o apoio
das agências FAPESP, CAPES e CNPq

1. INTRODUÇÃO

No Processamento de Línguas Naturais (PLN) buscam-se várias propostas para refinar os modelos de Sumarização Automática (SA), área que visa à produção de sumários a partir de uma ou mais fontes textuais (Mani, 2001; Nenkova and McKeown, 2011). Os métodos atuais para a SA podem ser baseados em modelos linguísticos, estatísticos ou híbridos dos anteriores, para lidar com a seleção e a organização de conteúdo. Em qualquer caso, são inúmeros os problemas de desempenho dos programas computacionais atuais, dentre os quais se destacam a ausência de organização coerente da informação, a inclusão de segmentos de informação irrelevantes, a exclusão de segmentos relevantes e a quebra da coesão textual e do encadeamento referencial apropriado. A falta de coesão textual muitas vezes é a responsável pela má formação, ou incoerência, do sumário final; a quebra do encadeamento referencial ou a exclusão de informações importantes pode afetar ou deturpar a mensagem original. A própria inclusão de informação irrelevante nos sumários pode resultar na repetição desnecessária de conteúdo, com claro prejuízo para a manutenção da ideia principal dos textos originais ou, mesmo, para sua compreensão.

Esses problemas de desempenho se devem, em sua maioria, à complexidade de se processar textos automaticamente: consideram-se diversos modelos ou parâmetros de decisão para dar conta da variedade de fenômenos linguísticos e, assim, elaborar a tarefa de SA.

Na SA monodocumento, por exemplo, em que se toma por base somente um texto-fonte, são clássicos o uso da frequência de palavras ou de sua distribuição no texto, como medidas relativas de significância (Luhn, 1958), e a identificação de palavras sinalizadoras da importância do conteúdo textual, como parâmetros de seleção computacional de sentenças com maior potencial de transmitir ao leitor a substância de um documento (Edmundson, 1969). Também se configurou como clássica a necessidade de se restringir domínios ou gêneros textuais para melhorar os resultados, sobretudo quanto se trata da SA extrativa, isto é, aquela em que os sumários são produzidos pela mera seleção de segmentos inalterados dos textos-fonte (Pollock e Zamora, 1975). Especialmente para textos jornalísticos, sejam eles informativos ou genéricos[1], as características típicas que são relevantes para a SA incluem a presença de palavras-chave, que são comumente associadas às mais frequentes no texto-fonte (Luhn, 1958), informações contidas em títulos, subtítulos ou no início do texto-fonte, dentre outras (Cremmins, 1996; Endress-Niggemeyer, 1998).

Na Sumarização Automática Multidocumento (SAM), que só recentemente passou a ser amplamente investigada, esses critérios permanecem atuais (ver, por exemplo, Mani, 2001; Jorge e Pardo, 2010), mesmo com diferente roupagem para o tratamento dos problemas de desempenho. Tendo em vista a produção de um sumário a partir de vários textos sobre um mesmo assunto, o alinhamento de segmentos de textos a partir de fontes diversas e o tratamento da redundância passam a ser problemas adicionais para a qualidade dos sumários na SAM.

As restrições de domínio e palavras ou segmentos sinalizadores da importância de conteúdo que, no passado, serviram para assegurar a qualidade dos

[1] Sumários informativos e genéricos são aqueles que incorporam o conteúdo principal do texto–fonte e, assim, podem substituir sua leitura. Os genéricos contemplam especialmente audiências de leitores mais gerais e pouco específicas.

modelos automáticos permaneceram como fatores importantes para se estabelecerem critérios para a modelagem dos sumarizadores multidocumento. Entretanto, recentemente se reacendeu o interesse pela identificação de outras características textuais que pudessem melhorar a SA mono ou multidocumento. Sumários manuais, produzidos por pessoas (e não máquinas), são comumente usados para esse fim.

Também a sumarização guiada (Owczarzak e Dang, 2011; Makino *et al*., 2011), visando à construção de sistemas orientados pelo significado, incorpora essas características. Essa modalidade de SAM guiada por aspectos foi proposta na TAC 2010 (*Text Analysis Conference*) para projetar sumários curtos e coerentes produzidos segundo categorias e aspectos pré-definidos. As categorias, segundo sua definição, indicam o assunto ou domínio do texto (p.ex.: esporte, mundo, etc.). Os aspectos podem ser específicos para cada categoria ou genéricos; neste caso indicam sua validade, ou abrangência, para um leque maior de domínios.

A identificação de aspectos textuais pode ser útil tanto para a determinação de informações relevantes dos textos-fonte, quanto para a identificação de restrições estruturais durante a construção dos seus sumários (Genest *et al*., 2009). Essas restrições, por sua vez, podem ser descritas por meio de esquemas de organização do conteúdo (os ditos *templates*), os quais podem servir para a sumarização dirigida por categorias, daí o termo 'sumarização guiada'. Esquemas de uma mesma categoria poderão incluir o mesmo tipo de fatos, descrito pelos aspectos relacionados a ela, o que a tornará previsível e, portanto, controlável para os processos automáticos.

Muitos estudos foram desenvolvidos seguindo esses princípios da TAC 2010. Steinberger *et al*. (2010) realizaram análises semânticas profundas para a modelagem de aspectos visando a SA multilíngue. Makino et al. (2011) e Li et al. (2011) compilaram aspectos de sumários da Wikipédia. Barrera et al. (2011) criaram um sistema de perguntas e respostas com base na identificação de aspectos para diferentes categorias. Mesmo antes da TAC, alguns trabalhos já apresentavam abordagens semelhantes, por exemplo, White et al. (2001) propuseram *templates* com base em aspectos para sumários de textos de desastres e Zhou et al. (2005) estudaram os aspectos presentes em sumários biográficos.

As categorias previstas na TAC 2010 são as seguintes: (1) *Acidentes e desastres naturais*, (2) *Ataques*, (3) *Saúde e segurança*, (4) *Recursos naturais ameaçados* e (5) *Julgamentos e investigações*. As regularidades relatadas a partir da modelagem de aspectos para cada uma dessas categorias incluem, por exemplo (Barrera et al., 2011): aspectos WHY (razão) e DAMAGES (danos) para a categoria (1) e aspectos WHAT (o que), IMPORTANCE (importância), THREATS (ameaças) e COUNTERMEASURES (providências) para a categoria (4). Alguns aspectos são gerais e se aplicam a várias categorias; outros são estritamente dependentes da categoria (ou seja, indicam uma dependência de domínio). Por exemplo, o aspecto THREATS é definido para a categoria *Acidentes e desastres naturais* e aponta especificamente ameaças a recursos naturais. Claramente aspectos dependentes de domínio devem ser recenseados para cada categoria de um dado córpus.

Neste relatório, descreve-se a anotação manual de aspectos textuais presentes nos sumários manuais multidocumento do córpus em português denominado CSTNews (Aleixo e Pardo, 2008; Cardoso *et al*., 2011a)[2], o qual foi construído com vistas à investigação da SA mono e multidocumento.

[2] Disponível para download em www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html.

Esse córpus é composto por 50 coleções de textos-fonte, cada uma englobando de 2 a 3 textos sobre um mesmo assunto (ou notícia jornalística). Cada texto de uma mesma coleção foi compilado de uma agência de notícia on-line amplamente conhecida no Brasil. Além dos textos-fonte crus (ou seja, sem nenhum tipo de anotação), cada coleção do CSTNews possui: (i) sumários manuais ou humanos monodocumento (também denominados *abstracts*[3]) para cada um dos textos do CSTNews, num total de 140 sumários; (ii) um sumário manual multidocumento para cada coleção, isto é, um sumário produzido por um especialista humano a partir de todos os textos de cada coleção; (iii) um sumário automático multidocumento para cada coleção, produzido por um sistema computacional baseado em um método particular de SA; (iv) um extrato manual multidocumento para cada coleção, ou seja, um extrato produzido por humanos competentes em português, a partir de segmentos inalterados dos textos-fonte); (v) versões anotadas, em nível discursivo, de cada um dos textos-fonte com base na *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987) e na *Cross-document Structure Theory* (CST) (Radev, 2000); dentre outras anotações. Os sumários manuais e automáticos multidocumento, assim como os extratos manuais multidocumento, totalizam 50 textos em cada caso.

A RST e a CST são teorias voltadas, respectivamente, à SA mono e multidocumento. A RST também pode ser usada para a SA multidocumento (veja, por exemplo, Cardoso et al., 2011b), apesar de ter sido explorada mais profundamente, até o momento, para a SA monodocumento (veja, por exemplo, Boguraev e Kennedy, 1997; Marcu, 1997a; 1997b; 2000; Uzêda et al., 2010). Neste caso, ela fornece um modelo de relevância para distinguir proposições originárias de segmentos a incluir nos sumários-alvo a partir de uma única fonte. Já a anotação CST permite lidar com os fenômenos multidocumento, identificando similaridades e diferenças entre os textos e, assim, determinar os segmentos mais apropriados para se incluir no sumário (veja, por exemplo, Zhang et al., 2002; Afantenos et al., 2008; Jorge e Pardo, 2010; Jorge et al., 2011).

Ao reproduzir a tarefa da TAC para a SAM de textos em português, os aspectos foram anotados manualmente por um grupo de falantes nativos da língua. Uma etapa preliminar desse trabalho foi apresentado no ELC'2012 (veja, e.g., Camargo et al., 2012; Jorge et al., 2012; Rassi et al., 2012; Zacarias et al., 2012).

A perspectiva de gerar mais conhecimento a partir das características textuais de sumários humanos multidocumento se insere no âmbito do Projeto SUSTENTO (*Generation of linguistic knowledge for Multi-document Summarization*)[4]. Já a aplicação desse conhecimento para melhorar a qualidade de sumários automáticos se insere no trabalho desenvolvido no Projeto SUCINTO[5](*Summarization for clever information access*). O SUSTENTO e o SUCINTO fornecem, assim, o contexto de interesse para a anotação de aspectos textuais aqui relatada, para adquirir mais conhecimento sobre a tarefa humana de produção de sumários multidocumento e, com isso, trazer maior informatividade à tarefa automática.

Para apresentar a anotação em questão, estruturou-se este relatório em 7 seções. Na Seção 2, discorre-se sobre as características do córpus em relação aos aspectos e categorias, para, em seguida, apresentar a metodologia de anotação e a

[3] Os sumários abstrativos, ao contrário dos extrativos, são produzidos por meio da reescrita do material linguístico presente nos textos-fonte.
[4] Projeto FAPESP 2012/13246-5/ CNPq 483231/2012-6.
[5] Projeto FAPESP 2012/03071-3, www.icmc.usp.br/pessoas/taspardo/sucinto/.

síntese da distribuição de aspectos. Na Seção 3, apresentam-se os resultados da anotação para todas as categorias consideradas e discutem-se casos de representatividade dos aspectos nos sumários do Córpus CSTNews. Na Seção 4, apresenta-se uma síntese global da anotação, cujos principais problemas são apresentados na Seção 5. Possíveis padrões de organização de conteúdo a partir dos aspectos, que podem guiar a SAM de textos em português, são apresentados na Seção 6. Esses padrões darão origem aos *templates* antes mencionados. Por fim, discutem-se alguns desdobramentos da anotação de aspectos para a SAM na Seção 7.

2. CATEGORIAS E ASPECTOS NO CÓRPUS CSTNews

A classificação de aspectos no CSTNews se baseia na suposição de que há dois critérios para se classificar textos de publicações como jornais e revistas (Nenkova e Louis, 2008): (i) nível de coesão e (ii) assunto. Essa classificação permite usar métodos mais orientados pelo significado para o PLN e buscar uma análise semântica mais profunda das fontes de conhecimento. Com base em (i) e (ii), os textos podem descrever um único evento ou múltiplos eventos, discutir um único assunto, prover informações sobre uma única pessoa ou ainda fornecer diferentes opiniões sobre um assunto. Para Nenkova e Louis (2008), o textos que veiculam informações sobre um único evento, assunto ou pessoa constituem categorias coesivas e os demais, que descrevem múltiplos eventos ou fornecem diferentes opiniões, categorias não-coesivas. Nenkova e Louis (2008) consideram que as categorias coesivas levam a melhor qualidade das aplicações computacionais do que categorias não-coesivas. Por essa razão, grandes categorias relacionadas ao assunto principal de cada texto do córpus foram consideradas na anotação.

As categorias no CSTNews diferem das delineadas na TAC 2010, pois são indicadas pelas seções jornalísticas das quais foram extraídos os textos e elas não correspondem diretamente às categorias originais da TAC. No entanto, reconhecem-se assuntos correlatos mesmo dentro das seis categorias consideradas: elas são ligadas aos cadernos 'Cotidiano', 'Esporte', 'Mundo', 'Política', 'Dinheiro' e 'Ciência' dos diversos jornais on-line no Brasil (*Folha de São Paulo*, *Estadão*, *O Globo*, *Gazeta do Povo* e *Jornal do Brasil*). No caderno 'Cotidiano', por exemplo, pode haver artigos sobre ataques ou julgamentos (categorias 2 ou 5 da TAC) ou saúde (categoria 3 da TAC); nos cadernos 'Ciência' ou 'Mundo', pode haver menção a acidentes e desastres naturais (categoria 1 da TAC); no caderno 'Ciência, recursos naturais ameaçados (categoria 4 da TAC) podem estar em foco. Ignorando a dispersão de categorias da TAC por vários cadernos do Córpus CSTNews, tomaram-se os próprios títulos dos cadernos como rótulos das categorias do córpus.

A quantidade de coleções de textos por categoria é mostrada na Tabela 1. As categorias 'Dinheiro' e 'Ciência' não apresentaram volume suficiente de notícias afins publicadas em fontes variadas, daí sua exclusão do conjunto de sumários que foram anotados com aspectos.

Tabela 1. Quantidade de coleções textuais por categoria.

Categoria	Quantidade de coleções de textos
Mundo	14
Cotidiano	14
Política	10
Esporte	10
Dinheiro	1
Ciência	1

Já que foram identificadas categorias distintas daquelas da TAC no CSTNews, resolveu-se verificar quais seriam os aspectos ocorrentes, embora sob a luz do rol completo já indicado na TAC também. Descrevem-se as fases de anotação a seguir.

2.1. A metodologia de anotação

Para a anotação dos aspectos, partiu-se do rol fornecido na TAC 2010. São 16 etiquetas distintas para aspectos que podem referir-se a conceitos ou objetos (OBJs), os quais são identificados como entidades a serem anotadas. Exemplos de etiquetas são: WHAT, para fatos ou eventos; WHEN, para datas; WHERE, para locativos; WHO, para *agente* e *paciente*, termos relacionados às etiquetas WHO_AGENT e WHO_AFFECTED; WHY, para causas ou justificativas; PLEAD, para asserções referentes a apelações (processuais) ou reações a cobranças; e PERPETRATOR, para agentes de um ataque. Para o CSTNews, esse rol foi refinado em função das categorias diferentes sugeridas nos textos. Esse refinamento envolveu tanto a exclusão de algumas etiquetas originais, quanto a inserção de novas etiquetas de interesse para o CSTNews. Quanto à sua denominação, optou-se por manter a terminologia da TAC, quando coincidentes, e também por mantê-las em inglês, para efeito de divulgação.

Na busca pela correspondência com etiquetas da TAC ou na identificação de novas etiquetas, foi necessário fazer uma definição clara de cada um dos aspectos, visando não só a consistência de anotação, como também a documentação de todo o processo. Para isso, recorreu-se ao seu contexto de ocorrência no córpus. Mesmo as etiquetas originais não foram definidas para a primeira anotação sugerida na TAC 2010. Na verdade, de acordo com os relatos à época, a anotação se pautou principalmente na intuição ou na interpretação dos leitores, para o reconhecimento das funcionalidades dos aspectos considerados. Por entender que esse procedimento *ad-hoc* prejudicaria a uniformidade de anotação do CSTNews, buscou-se a elaboração de definições que fosse consensual entre os anotadores. Como resultado, a definição dos aspectos constitui uma contribuição genuína para o PLN, em geral, e para a SA, em particular. Além disso, se considerado seu teor semântico, pode também ser usada em outros contextos que demandem conhecimento profundo da semântica de eventos ou categorias. Embora as definições tenham sido baseadas no português, elas são aplicáveis a outras línguas também.

2.2. Fases de anotação

A tarefa de anotação do CSTNews teve início em Maio de 2012 e estendeu-se até Dezembro. Inicialmente, constituiu-se o grupo de voluntários para a anotação manual do córpus de sumários manuais multidocumento. Ressalta-se que a decisão de anotar os sumários e não os respectivos textos-fonte se deveu ao alto grau de

complexidade da tarefa e aos interesses atuais dos Projetos SUSTENTO e SUCINTO, mais voltados, respectivamente, para a caracterização linguística dos sumários e identificação de seus padrões estruturais, para subsidiar a tarefa automática.

O grupo de voluntários foi dividido em 4 subgrupos compostos por 3 ou 4 linguistas computacionais, havendo um pesquisador sênior em cada subgrupo, como responsável pela coordenação da tarefa de anotação. Foram consideradas somente as 4 categorias com mais textos (Tabela 1), razão para a existência de 4 subgrupos: cada um ficou responsável pela anotação completa de uma categoria. Assim, as categorias mais representativas consideradas são: 'Mundo', 'Cotidiano', 'Política' e 'Esporte'.

A anotação propriamente dita foi feita em duas etapas, uma preliminar e outra final. Na primeira, consideraram-se instruções bastante flexíveis, pois não havia ainda um manual de anotação a ser seguido e admitia-se que cada grupo definisse seu próprio rol de etiquetas a partir das definidas na TAC. Neste caso, elas poderiam até ser redefinidas ou os anotadores poderiam ainda criar novas etiquetas. Os sumários das categorias menos representativas – 'Dinheiro' e 'Ciência' – foram usados como referência de anotação consensual nessa fase preliminar. Todos os membros dos subgrupos fizeram anotação dos dois sumários desses cadernos. Para a anotação final, os aspectos foram classificados em micro e macroaspectos (cf. será explicado mais tarde) e suas definições foram elaboradas, para servir de base para a anotação real do córpus.

2.2.1. A anotação preliminar

Nessa fase, os subgrupos iniciaram suas atividades pela análise das questões propostas na TAC 2010 (Owczarzak e Dang, 2011) e dos sumários manuais a serem anotados. A análise dos sumários visou identificar aspectos semelhantes aos da TAC, aspectos que ocorressem mais frequentemente no córpus ou alguma ordem preferencial, quando ocorressem vários aspectos em um mesmo sumário. Com base nessa análise, cada subgrupo poderia, de forma independente, redefinir suas categorias em função dos aspectos, criar novos aspectos ou ainda relacionar seus aspectos a outras teorias ou modelos de discurso, ontologias ou teorias semânticas, tais como a Teoria de Casos (Fillmore, 1968) ou os modelos semântico-cognitivos de Jackendoff (1983).

Especificamente, os subgrupos buscaram distinguir os seguintes casos, mesmo conscientes de que não necessariamente haveria respostas para todos eles:

i. Se alguns aspectos que já constavam da relação original também estavam presentes no CSTNews.
ii. Estando presentes, se eles seriam definidos como na proposta original.
iii. Se diferentes aspectos poderiam ser associados a um modelo de significância baseado em sua frequência nos sumários/córpus.
iv. Se a ordem dos aspectos seria determinante ou típica de uma dada categoria, podendo haver, inclusive, mais de uma ordem.
v. Se haveria aspectos mais difíceis de se encontrar, reconhecer ou, mesmo, se haveria discordância sobre suas definições, entre os anotadores.
vi. Se haveria aspectos correlacionados ou redundantes, que pudessem levar ou à subsunção de um por outro ou à generalização de definições entre eles, resultando em um único aspecto mais representativo.

vii. Se seria possível identificar relações entre os aspectos que pudessem subsidiar a sua organização ou estruturação. Neste caso, buscou-se identificar, por exemplo, relações de generalização e especificação entre os aspectos de forma a organizá-los em uma estrutura taxonômica (p.ex.: WHO_AGENT e WHO_AFFECTED podem ser vistos como aspestos mais específicos de WHO).
viii. Se o conhecimento do domínio seria necessário para a anotação de aspectos e em que nível da representação do conhecimento a anotação se faria possível.
ix. Em vez de se adotarem aspectos mais genéricos, se não seria importante manter as especificidades de domínio, para os aspectos em foco.
x. Finalmente, se haveria um modo de correlacionar a anotação de aspectos e a especificação das categorias com outros trabalhos da área, por exemplo, rotulagem de papéis semânticos (*semantic role labeling*), reconhecimento de entidades nomeadas (*named entity recognition*) ou análise baseada em gêneros (*genre analysis*).

A análise dos sumários também se pautou em outros trabalhos, sobretudo para que as conclusões pudessem ser escalonadas. Com esse objetivo, estabeleceu-se que novas coleções textuais seriam anotadas sempre que necessário, para confirmar hipóteses ou certificar as definições e o próprio uso dos aspectos.

Além das diferentes experiências exploratórias geradas pelos subgrupos sobre os aspectos textuais ocorrentes em sumários em português, essa anotação preliminar serviu para estabelecer as diretrizes notacionais para a especificação dos aspectos que foram usadas na anotação final. Essas diretrizes consistiram da adoção da sentença como unidade textual mínima para identificar os aspectos e das definições dos formatos de anotação e de armazenamento e nomeação dos arquivos correspondentes aos sumários anotados.

A sentença foi escolhida como unidade de análise por ser bem delimitada e veicular uma ideia completa. Desse modo, as unidades de informação embutidas em cada sentença podem veicular aspectos interrelacionados, e cada sentença de um sumário multidocumento pode ser anotada com vários aspectos. Estes, por sua vez, podem também ser associados a diversos níveis da estrutura discursiva: enquanto alguns são identificados localmente, isto é, entre um segmento e outros na mesma sentença, outros extrapolam o contexto de uma sentença, isto é, um segmento de uma sentença remete a outros segmentos no texto. Marcadores linguísticos presentes nos segmentos textuais também podem indicar as fronteiras de definição dos aspectos e, portanto, servem mesmo para identificá-los.

Também foi relevante distinguir aspectos relacionados ao tópico principal de um texto, daqueles relativos a informações secundárias. Neste caso, os aspectos são indicados pelo sufixo EXTRA. Por padrão, o tópico principal sempre é anotado com o aspecto WHAT. Logo, quando há uma mudança de tópico, a(s) sentença(s) correspondente(s) é anotada com WHAT_EXTRA. Analogamente, todos os seus outros aspectos interdependentes recebem esse sufixo.

O formato de anotação preestabelecido é o seguinte: cada sentença é delimitada por colchetes e todos os aspectos que nela ocorrem são agrupados ao fim, delimitados entre si pela barra de divisão (/). Assim, o formato genérico da anotação para n aspectos é dado por […<sentença>…]**aspecto_1/aspecto_2/…/aspecto_n.** Sempre que possível manteve-se a ordem das etiquetas igual à ordem dos segmentos sentenciais que as indicam. Os nomes das etiquetas foram mantidos em inglês para

facilitar a identificação das similaridades com os aspectos sugeridos na TAC 2010. Um exemplo de sentença anotada é dado abaixo:

[A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.]
WHO_AGENT/WHAT/SCORE/WHERE/CONSEQUENCE/SITUATION

Seguindo a prática usual em tarefas de anotação, os sumários manuais anotados com aspectos são salvos em arquivos texto (.txt) separadamente. Para a nomeação desses arquivos, acrescenta-se à nomeação original de cada sumário do CSTNews o sufixo *_aspects*, o que indica claramente que o texto está anotado com aspectos.

Essa experiência preliminar serviu de base para os procedimentos reais de identificação e definição dos aspectos no córpus, na fase de anotação final.

2.2.2. A anotação final

2.2.2.1. Identificação dos aspectos no córpus

A anotação real do córpus de sumários manuais multidocumento demandou repetidas discussões sobre a identificação dos segmentos representativos dos objetos conceituais em estudo e a determinação dos aspectos correspondentes em cada caso. Manteve-se a unidade mínima para efeito de anotação como a unidade sentencial. O esforço na busca de consistência de anotação era esperado para ambas as tarefas, já que identificar ou delimitar segmentos conceituais a partir da superfície textual é de ordem subjetiva e inerente à interpretação textual. Esse esforço foi ainda mais relevante porque não havia critérios claros ou, mesmo, definições formais de todos os aspectos, cuja normatização tornou o grupo dos projetos SUSTENTO e SUCINTO pioneiro nesse campo.

Inicialmente, decidiu-se tomar como base as relações sintagmáticas, com seus correspondentes descritos por papéis sintáticos. Assim é que se recorreu às seguintes etiquetas para os aspectos derivados dessas relações: WHO, WHERE, WHEN e WHAT. A seguir, considerou-se a estrutura argumental de cada sentença, cujas relações pudessem indicar os conceitos assinalados por unidades simples de significado (estas poderiam corresponder às unidades elementares de discurso definidas por Mann e Thompson na RST). Foram, então, identificadas ocorrências de justificativas, causas ou consequências, objetivos ou comparações, levando às etiquetas correspondentes WHY, CONSEQUENCE, GOAL e COMPARISON.

Até então, o nível intrassentencial fora explorado, porém, alguns conceitos ultrapassaram esse nível, fazendo-se necessário considerar também o nível intersentencial, na medida em que conceitos delineados intrassentencialmente estabeleciam vínculos com conceitos de outras sentenças, fenômeno natural para a construção da trama discursiva. A etiqueta WHAT é um exemplo que assinala a relação conceitual de um componente no nível intersentencial: ela indica o tópico principal do texto e norteia as escolhas de conteúdo do texto todo.

A necessidade de identificação de segmentos textuais em diversos níveis estruturais para a determinação do aspecto correspondente resultou na classificação dos aspectos em *microestruturais* e *macroestrutur*ais. Os primeiros emergem do texto com base em segmentos que, em geral, compõem uma sentença (nese caso, são segmentos intrassentenciais); os segundos referem-se, em geral, a uma sentença

completa e, assim, emergem da combinação de seus segmentos informativos (os quais compõem a ideia que ela incorpora). Desse modo, as etiquetas correspondentes resultam da verificação dessa distinção de granularidade dos segmentos que dão origem aos aspectos.

O Quadro 1 indica todos os aspectos encontrados no Córpus CSTNews, segundo essa classificação. São 10 macroaspectos e 7 microaspectos ao todo. Além disso, todos eles podem se desdobrar em etiquetas EXTRA, bastando, para isso, que o segmento textual correspondente não se refira ao tópico principal do texto, que já é anotado com WHAT, como mencionado antes. Dessa forma, a cada etiqueta do Quadro 1 pode ser acrescentado o sufixo EXTRA. Finalmente, apesar de se distinguir a diferença de granularidade entre os aspectos, verificou-se que nem sempre sua identificação mantinha uma regularidade de associação com segmentos inter ou intrassentenciais. Os aspectos marcados com um asterisco (*) no quadro indicam isso, ou seja, podem se revelar em nível micro ou macrotextual.

Os aspectos GOAL e SITUATION foram classificados como macroaspectos por serem mais recorrentes em anotações das ideias gerais sugeridas por uma sentença. Já WHY e HOW aparecem sinalizados por segmentos locais, referentes a unidades que compõem uma sentença, daí considerá-los microaspectos. Essa classificação geral não é uniforme entre as categorias. Para 'Cotidiano' e 'Política', por exemplo, HOW aparece como macroetiqueta, mas, em alguns casos da categoria 'Esporte', ela é microetiqueta. Estudos posteriores deverão servir para verificar se, de fato, haverá uma regularidade que permita classificá-las melhor como macro ou microaspecto.

Quadro 1. Aspectos no córpus (no macro e micro níveis).

Macroaspectos	Microaspectos
1. COMMENT	1. SCORE
2. COMPARISON	2. WHEN
3. CONSEQUENCE	3. WHERE
4. COUNTERMEASURES	4. WHO_AFFECTED
5. DECLARATION	5. WHO_AGENT
6. GOAL*	6. WHY *
7. HISTORY	7. HOW *
8. PREDICTION	
9. SITUATION *	

2.2.2.2. A definição dos aspectos

Os Quadros 2 e 3 apresentam as definições de cada micro e macroaspecto, respectivamente, acompanhadas de exemplos prototípicos para cada categoria do CSTNews. Categorias não ilustradas indicam a não ocorrência de casos no córpus, para o aspecto correspondente. Aspectos com sufixos EXTRA não são reproduzidos nesses quadros, pois suas definições permanecem as mesmas.

Quadro 2. Microaspectos do Córpus CSTNews.

Microaspectos	Definição e exemplo
HOW	*O modo como um fato/evento ocorre.*
	Cotidiano: Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças.
	Esporte: Fabiana conseguiu o ouro em três tentativas.
	Mundo: ---
	Política: Em cada um dos turnos, precisa de 308 votos favoráveis.
SCORE	*O resultado numérico de um fato/evento (score, tempo, distância, etc., sobretudo relativo a esportes).*
	Cotidiano: **---**
	Esporte: A seleção brasileira, sob direção de Dunga, conquistou o oitavo título da Copa América, goleando a Argentina por 3 a 0.
	Mundo: ---
	Política: ---
WHEN	*A data/período de tempo (estritamente temporal) de ocorrência de um fato/evento.*
	Cotidiano: Um homem suspeito de ter roubado o relógio Rolex do apresentador de televisão Luciano Huck foi detido na quarta-feira, 16, em Taboão da Serra, na Grande São Paulo.
	Esporte: A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos.
	Mundo: Antes de chegar à Jamaica, Dean matou ao menos nove pessoas nas ilhas de Santa Lúcia, Dominica, República Dominicana e Haiti, no Caribe.
	Política: O senador João Pedro (PT-AM), relator da segunda representação contra Renan Calheiros (PMDB-AL) no Senado, confirmou que vai apresentar nesta quarta-feira, 26, na reunião do Conselho de Ética, pedido de sobrestamento (suspensão temporária) das investigações sobre o caso.
WHERE	*A localização geográfica ou física de um fato/evento.*
	Cotidiano: Uma nova série de ataques criminosos foi registrada na madrugada desta segunda-feira, dia 7, em São Paulo e municípios do interior paulista.
	Esporte: A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.
	Mundo: Um acidente envolvendo dois trens, ao norte do Cairo, deixou por volta de 80 mortos e 165 feridos, segundo fontes policiais e médicas.
	Política: Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.
WHO_AGENT	*A entidade (pessoa ou organização) responsável por causar/provocar a ocorrência de um fato/evento.*
	Cotidiano: O Ministério Público Federal apreendeu nesta terça-feira, 7, os registros dos últimos cinco anos do livro de ocorrências da torre de controle do Aeroporto de Congonhas, zona sul de São Paulo, durante um mandado de busca e apreensão.

	Esporte: A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.
	Mundo: Um atirador matou ao menos 30 pessoas em dois diferentes locais da Universidade Técnica da Virgínia, em Blacksburg (Virgínia), nesta segunda-feira, no pior ataque a tiros contra um campus universitário da história dos Estados Unidos.
	Política: O ministro da Fazenda, Guido Mantega, apresentou nesta terça-feira a proposta do governo em troca do apoio do PSDB na votação da PEC (Proposta de Emenda Constitucional) que prorroga a cobrança da CPMF (Contribuição Provisória sobre Movimentação Financeira) até 2011.
WHO_AFFECTED	*A entidade (pessoa ou organização) que sofre os efeitos de um fato/evento.*
	Cotidiano: Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças.
	Esporte: A ginasta Jade Barbosa foi escolhida em votação na Internet, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim.
	Mundo: 17 pessoas morreram após a queda de um avião na República Democrática do Congo.
	Política: Na segunda representação, Renan é acusado de trabalhar para reverter dívida de R$100 milhões da Schincariol junto ao INSS.
WHY	*Uma explicação do porquê um fato/evento acontece (ou aconteceu).*
	Cotidiano: O crescimento nas autuações de contribuintes que caíram na malha fina se deu porque os auditores passaram a contar com programas mais modernos de computadores que analisam todas as irregularidades fiscais dos contribuintes, inclusive de anos anteriores, e não mais por grupos de infrações.
	Esporte: Maradona voltou a ter problemas de saúde no fim de semana e foi internado novamente em um hospital em Buenos Aires.
	Mundo: O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta após se chocar com uma montanha, prejudicado pelo mau tempo.
	Política: 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.

Quadro 3. Macroaspectos do Córpus CSTNews.

Macroaspectos	Definição e exemplo
COMMENT	*Um comentário do autor sobre um fato/evento.*
	Cotidiano: O presidente deu grande ênfase ao fim do protecionismo agrícola, que enriquece os ricos e empobrece os pobres.
	Esporte: Neste domingo, o esporte brasileiro alegrou a torcida verde-amarelo.
	Mundo: ---
	Política: ---
COMPARISON	*Dados ou estatísticas diferentes comparando duas ou mais entidades.*
	Cotidiano: Foram autuados 208.471 contribuintes, um crescimento

	de 104,47% em relação ao mesmo período do ano passado.
	Esporte: ---
	Mundo: ---
	Política: Quando se compara com uma pesquisa sem a lista oficial dos candidatos, Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%.
CONSEQUENCE	*Um fato/evento causado por outro fato/evento.*
	Cotidiano: A Secretaria da Fazenda também foi atingida por uma bomba.
	Esporte: A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca.
	Mundo: O furacão Dean passou pela costa sul da Jamaica, inundando a capital e espalhando árvores e telhados.
	Política: 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. Em alguns casos, os parlamentares estão sendo abandonados pelos partidos, especialmente por ser ano eleitoral
COUNTERMEASURES	*Medidas que visam solucionar/antecipar/impedir problemas relacionados a um fato/evento.*
	Cotidiano: Segundo informações do jornalista Ricardo Noblat, o presidente Luiz Inácio Lula da Silva mandou a FAB colocar dois aviões à disposição da família do senador.
	Esporte: Para evitar o segundo cartão amarelo, o treinador fez a substituição do jogador que estava muito nervoso. (exemplo que não consta do CSTNews)
	Mundo: Foi decretado estado de emergência preventiva no local.
	Política: Ocorre hoje a votação da PEC. Entretanto, a oposição passará o dia tentando obstruir os trabalhos em plenário com o único objetivo de retardar a votação e dificultar a tarefa governista.
DECLARATION	*Um discurso ou fala de alguém ou de uma fonte por citação direta ou indireta.*
	Cotidiano: Segundo um informante da delegacia, os dois teriam vendido o acessório de luxo avaliado em cerca de R$10 mil.
	Esporte: 'O Barcelona jogou o Mundial para valer no ano passado. Nós faremos o mesmo', disse o meia, após a partida em Stamford Bridge. (exemplo que não consta do CSTNews)
	Mundo: Segundo o jornal 'Choson Sinbo', mais de 7 mil casas foram destruídas ou danificadas, e quase 16 mil hectares de terra cultivada foram inundados.
	Política: **'**Eu não moverei uma palha contra eles [oposição] porque vocês moverão um paiol inteiro', afirmou o presidente Luiz Inácio Lula da Silva, candidato à reeleição pelo PT, sobre os ataques de seus adversários
GOAL	*Finalidade/razão para um fato/evento que irá acontecer.*
	Cotidiano: O objetivo das buscas é garantir a apreensão dos registros de ocorrências que contêm informações sobre as falhas no controle de tráfego aéreo.

	Esporte*: Boca entra em campo para ganhar após 5 partidas.* (exemplo que não consta do CSTNews)
	Mundo: A Operação Farrapos, da Polícia Federal, com o objetivo de desarticular uma quadrilha internacional de tráfico de drogas, prendeu 14 dos 17 suspeitos, após 2 anos de investigações.
	Política: Entretanto, a oposição passará o dia tentando obstruir os trabalhos em plenário com o único objetivo de retardar a votação e dificultar a tarefa governista.
HISTORY	*Informação de contexto sobre uma história/um passado relacionado ao fato/evento.*
	Cotidiano: ACM já tinha sofrido infarto em 1989 e já tinha recebido três pontes de safena.
	Esporte: A equipe brasileira já conquistou cinco vezes a Liga Mundial.
	Mundo: Este foi o maior acidente ferroviário egípcio desde 2002, após o incêndio de um trem que deixou 376 mortos.
	Política: Segundo a pesquisa CNI/Ibope, realizada em julho para o primeiro turno da eleição presidencial, Luiz Inácio Lula da Silva teria 44% dos votos contra 25% de Geraldo Alckmin e 11% de Heloisa Helena. Esta pesquisa foi a primeira da série CNI/Ibope com a lista oficial dos candidatos à Presidência, fornecido pelo TSE.
PREDICTION	*Informação sobre a factibilidade de fatos/eventos futuros (podendo, inclusive, ser um evento com ocorrência certa).*
	Cotidiano: Esse trabalho permitirá avaliar os riscos aos quais estão expostos os passageiros e tripulantes de aeronaves e tomar medidas necessárias para aumentar a segurança no setor aéreo.
	Esporte: O próximo confronto será contra os rivais mais perigosos, a seleção de Cuba.
	Mundo: Na sexta-feira, choveu muito acima do esperado e há previsão de mais tempestades hoje.
	Política: Com 2 pontos percentuais para mais e para menos, os resultados assegurariam vitória de Lula no primeiro turno.
SITUATION	*Uma ocasião em que ocorreu um fato/evento. Envolve uma transação, um campeonato, um compromisso ou outros tipos de situação em uma data ou local inespecíficos.*
	Cotidiano: O presidente Luiz Inácio Lula da Silva afirmou nesta segunda-feira, durante o programa de rádio 'Café com o Presidente', que vai anunciar obras de infra-estrutura e saneamento que transformarão o Brasil em um 'verdadeiro canteiro de obras'.
	Esporte: A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos.
	Mundo: Nesta batalha, 15 soldados israelenses morreram ao serem atingidos por um míssel.
	Política: Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.
WHAT	*Um fato/evento descrito no texto.*
	Cotidiano: Após quase 24 horas de tensão, terminou no fim da manhã desta quarta-feira a rebelião na Central de Custódia de Presos

	de Justiça (CCJP) no Maranhão.
	Esporte: A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca.
	Mundo: 17 pessoas morreram após a queda de um avião na República Democrática do Congo.
	Política: Ocorre hoje a votação da PEC (Proposta de Emenda Constitucional) que prorroga a cobrança da CPMF (Contribuição Provisória sobre Movimentação Financeira) até 2011 e mantém a alíquota de 0,38%.

2.3. Exemplos de anotação

Os textos abaixo ilustram os sumários manuais das categorias 'Ciência' (coleção 7) e 'Dinheiro' (coleção 30) (cf. Tabela 1), anotados com aspectos por todos os anotadores.

[Astrônomos têm denominado objetos menores que anãs marrons que não estão presos a nenhum sistema estelar de planemos, localizados nos arredores de regiões de formação de estrelas.]**WHO_AGENT_EXTRA/WHAT_EXTRA/WHERE_EXTRA**
[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.]**HOW/WHO_AGENT/WHAT**
['Este é um par de gêmeos verdadeiramente de destaque, já que cada um tem uma massa de apenas 1% de nosso Sol', declarou Jayawardhana.]**DECLARATION/ WHO_AGENT**
['Sua mera existência é uma surpresa e sua origem e seu futuro são um mistério', acrescentou.]**DECLARATION**

[O Itaú obteve no primeiro semestre de 2007 o maior lucro já registrado por um banco privado do país nos últimos vinte anos.]**WHO_AGENT/WHEN/WHAT/HISTORY**
[O lucro líquido acumulado chegou a R$ 4.016 bilhões, 35,7% acima dos R$ 2.958 bilhões do período em 2006 e também superior aos R$ 4.007 bilhões anunciados na véspera pelo Bradesco, líder no ranking de bancos do país.]**COMPARISON/DECLARATION/WHO_AGENT_EXTRA**
[Segundo cálculos da consultoria Economática, o resultado só perde para os R$ 4.032 bilhões registrados pelo Banco do Brasil no ano passado.]**WHO_AGENT_EXTRA/DECLARATION/COMPARISON**
[O resultado reflete ganhos não-recorrentes, sendo o principal a venda da participação do banco no Serasa e da sede do BankBoston em São Paulo e constituição de provisão para créditos de liquidação duvidosa excedente ao mínimo requerido de forma a permitir a absorção de eventuais aumentos de inadimplência ocasionados por forte reversão do ciclo econômico em situações de stress.]**HOW**

Como se pode notar, os aspectos são sempre indicados por letras maiúsculas e aparecem após o término da sentença, a qual é delimitada por colchetes. A informação principal no primeiro texto não aparece na primeira sentença, mas nas seguintes. No segundo texto, ela já aparece na primeira sentença. Todas as etiquetas não marcadas com o sufixo EXTRA sempre se referem à informação principal. Por esse motivo, pode-se depreender que o segundo texto inteiro não muda o foco, enquanto o primeiro já se inicia com uma informação periférica. A marca principal da informação central, em toda a anotação, é dada pela etiqueta WHAT: o assunto sobre o qual o texto versa.

3. A REPRESENTATIVIDADE DOS ASPECTOS NO CÓRPUS

Nesta seção apresentamos a síntese da distribuição dos aspectos para cada categoria, que é dada pela sua representatividade, medida por sua frequência de ocorrência na categoria. Nota-se que todos os grupos seguiram as mesmas definições apresentadas nos Quadros 2 e 3. Nos quadros que seguem, a indicação de aspectos seguidos de (e) assinala que os mesmos podem apresentar o sufixo EXTRA.

Em cada caso exibem-se as etiquetas que ocorrem nos sumários específicos de uma categoria, sua representatividade (medida por sua frequência de ocorrência em todos os sumários) e informações sobre sua distribuição relativa no córpus (especialmente considerando-se sumários multidocumento, essa distribuição abrange os textos da cada coleção que serviram de fonte para o sumário). São apresentadas também as posições de ocorrência de cada etiqueta, medidas pela posição da sentença em que ela aparece. Consideram-se as sentenças numeradas sequencialmente de 1 em diante. Cada tabela de distribuição por categoria abrange o número máximo de sentenças na categoria. Logo, nem todos os aspectos de um sumário ocuparão todas as posições ilustradas, já que não será todo sumário que conterá esse número máximo de sentenças e, tampouco, nem todo aspecto ocorrerá em todos os sumários.

Considerou-se importante registrar esse tipo de distribuição – por posição em cada sumário e por coleção de documentos – por dois motivos: i) a posição de um aspecto, em textos jornalísticos, pode indicar sua relevância; ii) a ocorrência significativa de um aspecto em uma certa coleção pode indicar que seu conteúdo é relevante para a modelagem computacional. No caso (i), é sabido, p.ex., que as informações que remetem ao tópico principal, em textos jornalísticos, aparecem, em geral, logo no início, nas chamadas 'sentenças lead'. Assim, a distribuição de aspectos por sentenças pode indicar quais esquemas serviriam de modelos para a geração automática de sumários (questão discutida na Seção 6). No caso (ii), as distribuições dos aspectos mais expressivos em algumas coleções podem indicar que elas devem ser recuperadas, para análises mais profundas.

3.1. Cotidiano

A categoria 'Cotidiano' compreende 14 coleções de textos, e portanto, 14 sumários manuais multidocumento. Esses textos descrevem temas variados, tais como: urbanismo, violência, trânsito, meteorologia e outros acontecimentos de ordem nacional. Os aspectos encontrados nessa categoria são apresentados no Quadro 4. Pode-se notar que, com exceção de SCORE, os outros 16 aspectos ocorrem nesses

textos, porém, não necessariamente denotam micro ou macroestruturas conforme a classificação apresentada na seção anterior. Sua reclassificação está presente nesse quadro.

Quadro 4. Aspectos existentes na categoria 'Cotidiano'.

Macroaspectos	Microaspectos
1. WHAT(e)	1. WHEN(e)
2. CONSEQUENCE	2. WHERE(e)
3. COUNTERMEASURES	3. WHO_AFFECTED(e)
4. DECLARATION	4. WHO_AGENT(e)
5. GOAL	5. WHY(e)
6. HISTORY	6. HOW(e)
7. PREDICTION	
8. SITUATION	
9. COMMENT	
10. COMPARISON	

As Tabelas 2 e 3 mostram o número de ocorrências dos micro e macroaspectos, respectivamente, e sua representatividade na coleção. Na Tabela 2 pode ser observado que os microaspectos WHEN (22%), WHERE (20%) e WHO_AGENT (38%) tiveram alta frequência, considerando o número de ocorrências juntamente com o seu respectivo extra. Dada a variedade de assuntos na categoria 'Cotidiano' e a frequência encontrada, acredita-se que as informações que essas etiquetas carregam são essenciais para essa categoria. A frequência de WHO_AFFECTED (10%) se explica pelo fato de parte dos sumários descreverem acidentes, desastres naturais e violência, ou seja, eventos que envolvem mortos e feridos. Os aspectos HOW e WHY tiveram baixa frequência, totalizando 6% e 4%, respectivamente.

Tabela 2. Representatividade dos microaspectos para a categoria 'Cotidiano'.

Microaspectos	Número de ocorrências	Representatividade
WHO_AGENT	22	21%
WHO_AGENT_EXTRA	18	17%
WHEN	13	12%
WHERE	12	11%
WHEN_EXTRA	11	10%
WHO_AFFECTED	9	9%
WHERE_EXTRA	9	9%
HOW	5	5%
WHY	2	2%
WHY_EXTRA	2	2%
WHO_AFFECTED_EXTRA	1	1%
HOW_EXTRA	1	1%
Total	**105**	**100%**

Na Tabela 3, observa-se que os macroaspectos WHAT (33%), DECLARATION (24%) e CONSEQUENCE (19%) foram os mais frequentes. WHAT e WHAT_EXTRA aparecem como esperado, entre os aspectos mais recorrentes, pois representam, respectivamente, eventos principais e secundários das notícias. Quanto a DECLARATION, sua frequência se deve ao fato de as agências de notícias geralmente

descreverem as declarações dos personagens envolvidos nos eventos. CONSEQUENCE também é significativamente frequente em textos do 'Cotidiano': quase metade dos textos dessa categoria relata consequências de eventos relacionados a acidentes (casuais ou propositais), ataques criminosos e fortes chuvas, os quais são acompanhados, por exemplo, de medidas tomadas após os acidentes. As ocorrências desses relatos de acidentes ou ataques é menos frequente quando agregam as medidas que foram tomadas após esses eventos, as quais são expressas pela macroetiqueta COUNTERMEASURES (2% de representatividade, ante 21% da mais representativa). Caso inédito é o de HISTORY: das 10 ocorrências, 7 estão em um mesmo sumário (C34). Diante disso, não se pode considerar que sua frequência seja expressiva o suficiente, para tipificar os textos da categoria 'Cotidiano'. Os demais macroaspectos também são inexpressivos nessa categoria.

Tabela 3. Representatividade dos macroaspectos para a categoria 'Cotidiano'.

Macroaspectos	Número de ocorrências	Representatividade
DECLARATION	28	24%
WHAT_EXTRA	25	22%
CONSEQUENCE	22	19%
WHAT	13	11%
HISTORY	10	9%
COMMENT	6	5%
COUNTERMEASURES	4	3%
SITUATION	4	3%
PREDICTION	1	1%
GOAL	1	1%
COMPARISON	1	1%
Total	115	100%

O sumário abaixo (coleção C4, categoria 'Cotidiano') descreve a situação do trânsito em São Paulo, depois de fortes chuvas. Os aspectos da primeira sentença se mostram como um padrão na categoria 'Cotidiano'.

[A forte chuva em São Paulo complicava o trânsito na manhã desta segunda-feira, 16.]**WHAT/WHERE/WHEN**
[Às 9h, a cidade tinha oito pontos de alagamento, sendo dois intransitáveis.]**WHEN/CONSEQUENCE**
[O risco de que novos alagamentos surgissem fez com que o Centro de Gerenciamento de Emergência (CGE) Prefeitura colocasse a cidade em estado de atenção.]**COUNTERMEASURES**
[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.]**DECLARATION/CONSEQUENCE/WHO_AFFECTED/WHEN_EXTRA**
[O estado de atenção foi suspenso às 9h25.]**WHAT_EXTRA/WHEN_EXTRA**

As Tabelas 4 e 5 mostram, respectivamente, a distribuição dos micro e macroaspectos por posição (sentença) que ocupam nos textos analisados. As

posições refletem as posições das sentenças em cada texto. Nem todos os sumários possuem 14 sentenças; a média por sumário é de 7 sentenças, na categoria 'Cotidiano'. Assim, são exibidas frequências dos aspectos contidos na primeira sentença do texto (S1), na segunda sentença (S2), etc. Pode ser observado na Tabela 4 que os aspectos WHERE e WHEN aparecem com mais frequência na primeira sentença dos sumários. Para os outros microaspectos não foi encontrado um padrão de distribuição. No máximo, há 14 sentenças nos sumários do 'Cotidiano'.

Tabela 4. Distribuição dos microaspectos por posição, na categoria 'Cotidiano'.

Microaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	Total
WHO_AGENT	7	3	4	1	3	0	1	1	1	0	1	0	0	0	22
WHO_AGENT_EXTRA	2	1	3	5	0	1	1	2	1	1	0	0	1	0	18
WHEN	12	1	0	0	0	0	0	0	0	0	0	0	0	0	13
WHERE	10	1	0	0	0	1	0	0	0	0	0	0	0	0	12
WHEN_EXTRA	1	1	0	1	2	0	1	0	1	0	1	1	1	1	11
WHO_AFFECTED	2	2	1	3	1	0	0	0	0	0	0	0	0	0	9
WHERE_EXTRA	1	0	0	2	1	2	1	1	1	0	0	0	0	0	9
HOW	0	1	1	1	1	1	0	0	0	0	0	0	0	0	5
WHY	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
WHY_EXTRA	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
WHO_AFFECTED_EXTRA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
HOW_EXTRA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Total de ocorrências	36	11	9	14	7	6	4	4	6	1	3	1	2	1	105

Pode ser observado na Tabela 5 que o aspecto WHAT foi bastante expressivo nas primeiras sentenças dos sumários. Embora o aspecto DECLARATION também teve boa frequência na primeira sentença, não se pode afirmar que seja um padrão recorrente. Na categoria 'Cotidiano', isso acontece devido a alguns sumários começarem com declarações de indivíduos ou organizações. Para os outros aspectos, não é encontrado um padrão de distribuição.

Tabela 5. Distribuição dos macroaspectos por posição, na categoria 'Cotidiano'.

Macroaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	Total
DECLARATION	6	3	4	4	2	0	2	3	2	1	1	0	0	0	28
WHAT_EXTRA	1	3	3	3	6	3	4	1	0	0	0	0	1	0	25
CONSEQUENCE	0	6	2	2	1	2	2	3	2	1	1	0	0	0	22
WHAT	10	0	1	1	0	1	0	0	0	0	0	0	0	0	13
HISTORY	0	2	1	0	0	1	0	0	1	1	1	1	1	1	10
COMMENT	0	0	0	1	1	2	0	0	0	1	0	1	0	0	6
COUNTERMEASURES	0	0	1	1	1	0	0	1	0	0	0	0	0	0	4
SITUATION	4	0	0	0	0	0	0	0	0	0	0	0	0	0	4
GOAL	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
PREDICTION	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
COMPARISON	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Total de ocorrências	21	15	14	12	11	9	8	8	5	4	3	2	2	1	115

As Tabelas 6 e 7 mostram a distribuição de todos os aspectos por coleção, na categoria 'Cotidiano'. São 14 coleções de textos ao todo. Pode-se notar que alguns aspectos ocorrem mais em algumas coleções e outros não ocorrem em coleção nenhuma. Como já dito, os microaspectos mais frequentes foram WHO_AGENT e WHEN, sendo que os mesmos só não ocorrem em uma coleção de textos cada (C22 e

C3, respectivamente, para 0 ocorrências quando remetem a informações principais ou secundárias, caso do sufixo EXTRA). O aspecto WHERE, apesar de sua relevante frequência, não apareceu em todas as coleções.

Tabela 6. Distribuição dos microaspectos nas coleções da categoria 'Cotidiano'.

croaspectos	C3	C4	C5	C6	C11	C21	C22	C33	C34	C36	C37	C39	C45	C49	Total
WHO_AGENT	0	0	1	5	2	0	0	7	1	0	1	1	1	3	22
WHO_AGENT_EXTRA	3	1	1	0	0	2	0	1	1	4	0	0	3	2	18
WHEN	0	2	1	1	1	1	1	1	1	1	1	1	0	1	13
WHERE	1	1	1	0	1	1	1	0	0	2	1	1	1	1	12
WHEN_EXTRA	0	1	0	1	2	0	0	0	0	6	0	0	1	0	11
WHO_AFFECTED	1	1	1	0	0	0	0	0	0	1	3	0	1	1	9
WHERE_EXTRA	0	0	0	0	4	1	0	0	0	2	0	0	1	1	9
HOW	0	0	0	0	1	3	0	0	0	0	1	0	0	0	5
WHY	0	0	0	0	0	0	1	0	0	1	0	0	0	0	2
WHY_EXTRA	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
WHO_AFFECTED_EXTRA	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
HOW_EXTRA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Total de ocorrências	6	6	5	7	12	8	3	9	4	19	7	3	8	8	105

Na Tabela 7, observa-se que WHAT ocorre na maioria das coleções, o que se explica por esse macroaspecto descrever o evento ou fato principal. No entanto, nas coleções C5, C6 e C49, ele não aparece, mas está implícito na identificação do fato principal como DECLARATION. Este aspecto é atribuído a textos que relatam eventos que geralmente reportam as falas dos envolvidos.

Tabela 7. Distribuição dos macroaspectos nas coleções da categoria 'Cotidiano'.

Macroaspectos	C3	C4	C5	C6	C11	C21	C22	C33	C34	C36	C37	C39	C45	C49	Total
DECLARATION	3	1	2	5	0	2	0	5	1	3	0	0	3	3	28
WHAT_EXTRA	2	1	2	0	2	2	0	3	2	4	1	0	3	3	25
CONSEQUENCE	3	2	0	0	7	2	6	0	2	0	0	0	0	0	22
WHAT	2	1	0	0	1	1	1	1	1	1	1	1	2	0	13
HISTORY	0	0	0	0	1	0	0	0	1	7	0	0	1	0	10
COMMENT	0	0	1	0	0	0	1	4	0	0	0	0	0	0	6
COUNTERMEASURES	0	1	0	0	0	0	2	0	0	1	0	0	0	0	4
SITUATION	0	0	1	1	0	0	0	1	0	0	0	0	0	1	4
GOAL	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
PREDICTION	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
COMPARISON	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Total de ocorrências	10	6	6	6	11	7	10	14	8	16	2	3	9	7	115

3.2. Esporte

A categoria 'Esporte' compreende 10 coleções de textos e, consequentemente, 10 sumários humanos multidocumento. Dessas 10 coleções, 7 se caracterizam pelo relato de eventos de natação, vôlei, futebol e atletismo. As outras 3 coleções não englobam efetivamente notícias sobre eventos esportivos, mas sim de fatos correlatos, como o 'estado de saúde de um jogador de futebol'. A distribuição das coleções em função dos diferentes esportes é mostrada na Figura 1.

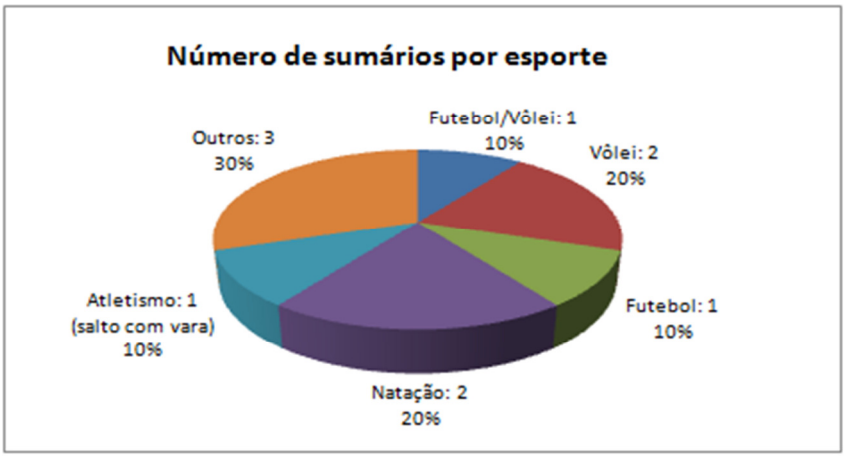

Figura 1. Distribuição das coleções de textos na categoria 'Esporte'.

O Quadro 5 exibe os 16 aspectos que ocorrem nos sumários da categoria 'Esporte' reorganizados nas categorias micro e macro. O únioco ausente em esportes é COUNTERMEASURES.

Comparando o Quadro 5 com o Quadro 1, tem-se que HOW, tipicamente microaspecto no CSTNews, ocorreu como macroaspecto na categoria 'Esporte'. De forma similar, SITUATION, tipicamente macro, ocorreu como microaspecto na categoria 'Esporte'. As variações e as dificuldades de anotação são explicitadas na Seção 4.

Quadro 5. Aspectos existentes na categoria 'Esporte'.

Macroaspectos	Microaspectos
1. WHAT[(e)]	1. WHERE[(e)]
2. CONSEQUENCE [(e)]	2. WHEN [(e)]
3. PREDICTION	3. WHO_AGENT [(e)]
4. HISTORY	4. WHO_AFFECTED [(e)]
5. COMMENT [(e)]	5. WHY [(e)]
6. HOW	6. SCORE [(e)]
7. COMPARISION	7. SITUATION [(e)]
8. DECLARATION	
9. GOAL	

Nas Tabelas 8 e 9, mostram-se o número de ocorrências e a representatividade dos micro e macroaspectos. Nota-se que alguns microaspectos são muito mais frequentes do que outros. Enquanto WHO_AGENT (23%), WHO_AGENT_EXTRA (15%) e WHEN (12%) correspondem, juntos, a quase 50% das ocorrências, WHY (2%), WHO_AFFECTED_EXTRA (1%) e WHY_EXTRA (1%) chegam, juntos, a apenas 4%. Essa distribuição revela que os textos relatam eventos esportivos em que o feito ou desempenho de atletas (ou times) têm relevância, sendo os atletas os atuantes das ações, daí a alta incidência de WHO_AGENT (mesmo em sua forma EXTRA). A indicência de datas (WHEN) e ocasiões (SITUATION) também é significativa nesses textos. Em geral, tratam da agenda de jogos e do próprio campeonato.

A baixa incidência de WHY e WHY_EXTRA indica que as notícias sobre eventos esportivos comumente não fornecem justificativas para os feitos ou desempenho dos atletas.

Tabela 8. Representatividade dos microaspectos para a categoria 'Esporte'.

Microaspectos	Número de Ocorrências	Representatividade
WHO_AGENT	19	23%
WHO_AGENT_EXTRA	13	15%
WHEN	10	12%
SCORE	6	7%
SITUATION	6	7%
WHO_AFFECTED	5	6%
SITUATION_EXTRA	5	6%
WHERE	5	6%
WHEN_EXTRA	4	5%
WHERE_EXTRA	4	5%
SCORE_EXTRA	3	4%
WHY	2	2%
WHO_AFFECTED_EXTRA	1	1%
WHY_EXTRA	1	1%
Total	**84**	**100%**

Quanto aos macroaspectos, HOW é o mais frequente (20%), revelando que a forma por meio da qual se dá o resultado de um jogo ou o desempenho de um atleta/time é informação constante em notícias esportivas. COMMENT é o segundo mais frequente (19%) e isso revela que também é comum a presença da opinião do autor da notícia a respeito do evento esportivo.

Os macroaspectos WHAT_EXTRA (16%) e WHAT (12%) também são frequentes, pois representam, respectivamente, os eventos secundários e principais das notícias. WHAT_EXTRA é mais frequente que WHAT porque as notícias de esporte tendem a veicular muita informação que não se refere ao tópico principal do texto. Por exemplo, em uma notícia sobre a conquista de uma medalha de ouro, relata-se também o desempenho dos demais atletas que disputaram a mesma prova.

Ainda sobre os mais frequentes, ressalta-se que a representatividade de CONSEQUENCE (9% das ocorrências) reflete a existência, em notícias esportivas, de menções a consequências de resultados de jogos ou competições, por exemplo.

Tabela 9. Representatividade dos macroaspectos para a categoria 'Esporte'.

Macroaspectos	Número de Ocorrências	Representatividade
HOW	17	20%
COMMENT	16	19%
WHAT_EXTRA	15	16%
WHAT	10	12%
CONSEQUENCE	8	9%
PREDICTION	5	6%
HISTORY	4	5%
COMMENT_EXTRA	4	5%
DECLARATION	2	2%
CONSEQUENCE_EXTRA	2	2%
COMPARISION	1	1%
GOAL	1	1%
Total	**85**	**100%**

Alguns macroaspectos têm frequência mediana, como PREDICTION (5%), HISTORY (4%) e COMMENT_EXTRA (4%), refletindo que as notícas esportivas não veiculam com frequência compromissos futuros do atleta/time, retrospectivas de desempenho e opiniões sobre os fatos secundários. A baixa frequência de COMPARISION e GOAL (1%) evidencia que dados comparativos entre atletas/time e o objetivo dos mesmos em uma competição são raramente mencionados.

Nas Tabelas 10 e 11, apresenta-se a distribuição dos micro e macroaspectos em função da presença dos mesmos nas diferentes sentença.

Tabela 10. Distribuição dos microaspectos por posição, na categoria 'Esporte'.

Microaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Total
WHO_AGENT	6	3	5	2	2	0	1	0	0	0	0	0	19
WHO_AGENT_EXTRA	0	1	3	3	2	0	0	1	1	1	1	0	13
WHEN	3	0	0	2	2	1	0	0	1	1	0	0	10
SCORE	3	2	2	0	0	0	0	0	0	0	0	0	7
SITUATION	5	1	0	0	0	0	0	0	0	0	0	0	6
WHO_AFFECTED	2	1	1	1	0	0	0	0	0	0	0	0	5
WHERE	4	1	0	0	0	0	0	0	0	0	0	0	5
SITUATION_EXTRA	0	0	2	2	0	1	0	0	0	0	0	0	5
WHEN_EXTRA	0	1	3	0	0	0	0	0	0	0	0	0	4
WHERE_EXTRA	0	1	1	1	0	1	0	0	0	0	0	0	4
SCORE_EXTRA	0	1	0	1	1	0	0	0	0	0	0	0	3
WHY	1	0	0	1	0	0	0	0	0	0	0	0	2
WHO_AFFECTED_EXTRA	0	0	0	0	0	1	0	0	0	0	0	0	1
WHY_EXTRA	0	0	0	0	0	1	0	0	0	0	0	0	1
Total de ocorrências	**24**	**12**	**17**	**13**	**7**	**5**	**1**	**1**	**2**	**2**	**1**	**0**	**85**

Tabela 11. Distribuição das macroaspectos por posição, na categoria 'Esporte'.

Macroaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Total
HOW	1	1	2	2	2	1	2	2	1	1	1	1	17
COMMENT	4	2	2	1	2	2	2	0	1	0	0	0	16
WHAT_EXTRA	0	4	4	3	2	1	0	0	0	0	0	0	14
WHAT	8	2	0	0	0	0	0	0	0	0	0	0	10
CONSEQUENCE	3	2	3	0	0	0	0	0	0	0	0	0	8
PREDICTION	0	1	1	1	0	1	0	0	1	0	0	0	5
HISTORY	0	0	1	1	1	1	0	0	0	0	0	0	4
COMMENT_EXTRA	0	0	1	0	1	0	0	0	0	1	0	1	4
DECLARATION	0	0	0	1	0	0	0	1	0	0	0	0	2
CONSEQUENCE_EXTRA	0	0	0	1	0	0	0	0	1	0	0	0	2
COMPARISION	1	0	1	0	0	0	0	0	0	0	0	0	2
GOAL	0	0	0	1	0	0	0	0	0	0	0	0	1
Total de ocorrências	**17**	**12**	**15**	**11**	**8**	**6**	**4**	**3**	**4**	**2**	**1**	**2**	**85**

Nota-se que os microaspectos WHO_AGENT, SITUATION, WHERE, SCORE e WHEN ocorrem com mais frequência na primeira sentença. O mesmo é observado quanto ao macroaspecto WHAT, o que indica a característica do gênero jornalístico de apresentar o evento/fato principal logo no início do texto. Os macroaspectos COMMENT e CONSEQUENCE também ocorrem com frequência na S1. O sumário apresentado na sequência, pertencente à coleção 8 (C8) da categoria 'Esporte', ilustra a ocorrência dos aspectos na primeira sentença.

[A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.]**WHO_AGENT/WHAT/SCORE/WHERE/CONSEQUENCE/SITUATION**
[Amanhã as equipes voltarão a se enfrentar, no mesmo local.]**WHEN_EXTRA/ WHO_AGENT_EXTRA/PREDICTION/WHERE_EXTRA**
[Com o resultado, o Brasil está na liderança do grupo B, perto da classificação para a próxima fase do campeonato.]**WHO_AGENT/CONSEQUENCE**
[A seleção brasileira ainda enfrentará portugueses e finlandeses na fase de classificação.]**WHO_AGENT/PREDICTION/SITUATION_EXTRA**
[A equipe brasileira já conquistou cinco vezes a Liga Mundial.]**WHO_AGENT/ HISTORY**
[A fase final deste ano acontecerá na Rússia.]**PREDICTION/WHERE_EXTRA**

As Tabelas 12 e 13 mostram, respectivamente, a distribuição dos micro e macroaspectos por coleção de textos.

Tabela 12. Distribuição dos microaspectos por coleção, para a categoria 'Esporte'.

Microaspectos	C8	C38	C41	C24	C25	C27	C28	C31	C48	C19	Total
WHO_AGENT	4	2	1	2	5	1	2	1	1	0	19
WHO_AGENT_EXTRA	1	1	3	1	0	4	0	0	3	0	13
WHEN	0	1	1	0	1	5	0	0	2	0	10
SCORE	1	1	1	1	1	0	1	0	1	0	7
SITUATION	1	1	2	1	0	0	0	0	1	0	6
WHO_AFFECTED	0	0	0	0	0	0	0	1	0	4	5
SITUATION_EXTRA	1	1	2	0	0	0	1	0	0	0	5
WHERE	1	0	0	0	0	1	1	0	1	1	5
WHEN_EXTRA	1	1	1	0	0	0	0	1	0	0	4
WHERE_EXTRA	2	0	0	0	0	0	0	1	1	0	4
SCORE_EXTRA	0	0	0	3	0	0	0	0	0	0	3
WHY	0	0	0	0	0	0	0	0	0	2	2
WHO_AFFECTED_EXTRA	0	0	1	0	0	0	0	0	0	0	1
WHY_EXTRA	0	0	1	0	0	0	0	0	0	0	1
Total de ocorrências	**12**	**8**	**13**	**8**	**7**	**11**	**5**	**4**	**10**	**7**	**85**

Tabela 13. Distribuição dos macroaspectos por coleção, para a categoria 'Esporte'.

Macroaspectos	C8	C38	C41	C24	C25	C27	C28	C31	C48	C19	Total
HOW	0	0	0	1	2	11	0	1	2	0	17
COMMENT	0	1	1	0	4	6	1	0	3	0	16
WHAT_EXTRA	0	1	4	3	0	0	0	1	2	3	14
WHAT	1	1	1	1	1	1	1	1	1	1	10
CONSEQUENCE	2	1	2	1	1	0	1	0	0	0	8
PREDICTION	3	0	0	0	0	0	0	1	1	0	5
HISTORY	1	0	0	0	2	0	1	0	0	0	4
COMMENT_EXTRA	0	0	0	1	0	2	0	0	1	0	4
DECLARATION	0	0	0	0	0	0	0	0	1	1	2
CONSEQUENCE_EXTRA	0	0	1	0	0	1	0	0	0	0	2
COMPARISION	0	0	0	1	0	0	1	0	0	0	2
GOAL	0	0	0	0	0	0	1	0	0	0	1
Total de ocorrências	**7**	**4**	**9**	**8**	**10**	**21**	**6**	**4**	**11**	**5**	**85**

A notável representatividade de WHO_AGENT e WHAT (e ambas as versões EXTRA) revela que notícias sobre esportes relatam um fato ou evento – daí suas etiquetas WHAT ou WHAT_EXTRA – em que atletas e times (anotados com WHO_AGENT ou WHO_AGENT_EXTRA) apresentam o desempenho mencionado. WHAT_EXTRA é mais frequente que WHAT (8% contra 6%) porque a notícia traz mais informação que não está associada ao tópico principal. Por exemplo, a informação sobre uma aquisição de uma medalha de ouro por um atleta comumente é agregada à informação sobre o desempenho dos outros atletas que também competiram. Os aspectos HOW e COMMENT também são frequentes (representatividade de 10% e 9%, respectivamente). Isso mostra que o modo como se chega a um resultado de um jogo, ou a um certo desempenho de um atleta ou time, e a opinião do autor do texto são informações recorrentes nas notícias sobre esportes. A alta frequência de HOW, no entanto, se deve a várias ocorrências em um único sumário. Os aspectos com frequência média, como CONSEQUENCE (5%), SCORE (4%) e SITUATION (4%),

sugerem que as notícias trazem menos frequentemente um resultado, suas consequências ou o local em que os jogos ou torneios ocorreram. Esses três aspectos, no entanto, são bem comuns na maioria dos sumários da categoria 'Esporte'. Vale notar ainda que SCORE é um aspecto típico, exclusivo, dessa categoria.

3.3. Mundo

A categoria 'Mundo' compreende 14 coleções de textos e, portanto, 14 sumários multidocumento. As coleções podem ser divididas em textos sobre acidentes, desastres naturais, ataques e decisões legais e políticas. A distribuição das coleções nesses temas é mostrada na Figura 2.

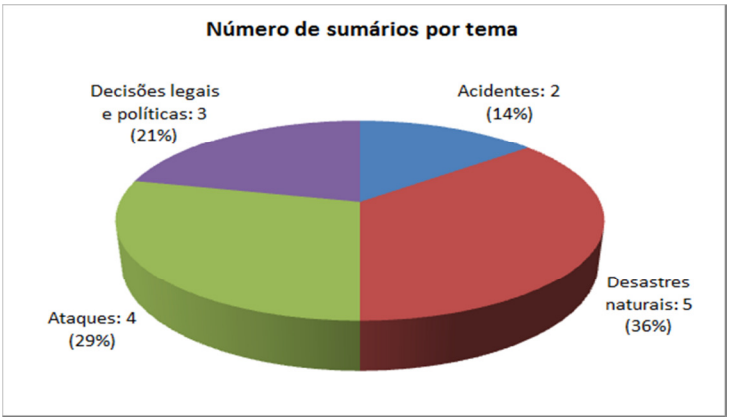

Figura 2. Distribuição das coleções de textos na categoria 'Mundo'.

O Quadro 6 exibe os aspectos que ocorrem nesses sumários. Dos 17 previstos, 13 aparecem em 'Mundo'. Os ausentes são: COMMENT, HOW, COMPARISON e SCORE. Entende-se a ausência de SCORE por ser uma etiqueta muito específica da categoria 'Esporte', dentre as quatro sob enfoque na tarefa de anotação.

Nesse quadro, os aspectos também foram reclassificados como micro ou macroaspectos, segundo sua forma de ocorrência. Por exemplo, apesar de haver macroaspectos WHY, a maioria dos WHY são microapectos na categoria 'Mundo'. Similarmente, SITUATION sempre ocorre como um microaspecto. Na Seção 4, em que se discutem as exceções e dificuldades encontradas, os casos variantes são comentados.

Quadro 6. Aspectos existentes na categoria 'Mundo'.

Macroaspectos	Microaspectos
1. WHAT(e)	1. WHEN(e)
2. CONSEQUENCE(e)	2. WHERE(e)
3. COUNTERMEASURES(e)	3. WHO_AFFECTED(e)
4. DECLARATION(e)	4. WHO_AGENT(e)
5. HISTORY	5. WHY(e)
6. PREDICTION(e)	6. GOAL(e)
	7. SITUATION

As Tabelas 14 e 15 mostram o número de ocorrências e a representatividade de cada aspecto na categoria 'Mundo'. Dentre os microaspectos, WHERE corresponde a mais de 14%, enquanto GOAL e SITUATION não chegam a 2%. Também é interessante notar

que WHO_AFFECTED é mais frequente que sua contraparte WHO_AGENT. Isso pode ser explicado porque muitos dos textos de 'Mundo' abordam acidentes, desastres e ataques, o que leva a correlacionar a eles mais as vítimas do que os agentes causadores desses eventos.

Tabela 14. Representatividade dos microaspectos para a categoria 'Mundo'.

Microaspectos	Número de Ocorrências	Representatividade
WHERE	17	14,78%
WHO_AGENT_EXTRA	17	14,78%
WHO_AFFECTED	16	13,91%
WHEN_EXTRA	12	10,43%
WHY	12	10,43%
WHEN	10	8,70%
WHO_AFFECTED_EXTRA	10	8,70%
WHERE_EXTRA	8	6,96%
WHO_AGENT	5	4,35%
WHY_EXTRA	2	1,74%
GOAL	2	1,74%
GOAL_EXTRA	2	1,74%
SITUATION	2	1,74%
Total	**115**	**100%**

Dentre os macroaspectos, WHAT e WHAT_EXTRA são os aspectos mais usuais, já que remetem aos eventos e fatos principais narrados. Devido à natureza dos textos, com muitos acidentes, desastres e ataques, os aspectos COUNTERMEASURES e CONSEQUENCE são também frequentes, representando, respectivamente, as várias medidas ou ações de socorro, precauções ou reconstruções decorrentes dos problemas relatados, em si. É interessante notar que DECLARATION e HISTORY ainda superam CONSEQUENCE: no contexto mais frequente, de desastres, ataques, etc., é comum que haja agências de notícias ou porta-vozes oficiais (representantes de seções governamentais, por exemplo) fazendo declarações sobre o ocorrido e prestando informações oficiais.

Tabela 15. Representatividade dos macroaspectos para a categoria 'Mundo'.

Macroaspectos	Número de Ocorrências	Representatividade
WHAT_EXTRA	27	26,21%
WHAT	16	15,53%
COUNTERMEASURES	14	13,59%
DECLARATION	13	12,62%
HISTORY	13	12,62%
CONSEQUENCE	10	9,71%
PREDICTION	5	4,85%
CONSEQUENCE_EXTRA	2	1,94%
COUNTERMEASURES_EXTRA	1	0,97%
DECLARATION_EXTRA	1	0,97%
PREDICTION_EXTRA	1	0,97%
Total	**107**	**100%**

As Tabelas 16 e 17 mostram a distribuição de micro e macroaspectos em função de sua posição nos sumários. Pode-se ver que WHERE, WHO_AFFECTED, WHEN e WHO_AGENT costumam ocorrer com mais frequência na primeira sentença. WHERE é a que mais ocorre em primeiras sentenças dos sumários, com 13 ocorrências, no total. WHY, por sua vez, costuma ocorrer mais bem distribuído, aparecendo principalmente no meio do sumário. Note, por exemplo, que sua maior ocorrência (4, no total) está em sentenças que ocupam a quarta posição nos sumários. Já os macroaspectos WHAT e WHAT_EXTRA ocorrem massivamente no início dos sumários, pois, como mencionado, o evento/fato principal é a informação veiculada logo no início dos textos jornalísticos.

Tabela 16. Distribuição dos microaspectos por posição, na categoria 'Mundo'.

Microaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total
WHERE	13	1	0	1	1	1	0	0	0	0	17
WHO_AGENT_EXTRA	0	1	3	5	2	3	2	1	0	0	17
WHO_AFFECTED	11	4	0	1	0	0	0	0	0	0	16
WHEN_EXTRA	0	3	3	1	2	1	1	1	0	0	12
WHY	3	1	0	4	2	2	0	0	0	0	12
WHEN	7	1	0	0	1	1	0	0	0	0	10
WHO_AFFECTED_EXTRA	0	2	3	2	2	1	0	0	0	0	10
WHERE_EXTRA	0	2	3	1	1	1	0	0	0	0	8
WHO_AGENT	5	0	0	0	0	0	0	0	0	0	5
GOAL	2	0	0	0	0	0	0	0	0	0	2
GOAL_EXTRA	0	0	0	1	1	0	0	0	0	0	2
WHY_EXTRA	0	1	1	0	0	0	0	0	0	0	2
SITUATION	1	0	0	1	0	0	0	0	0	0	2
Total de Ocorrências	**42**	**16**	**13**	**17**	**12**	**10**	**3**	**2**	**0**	**0**	**115**

Tabela 17. Distribuição dos macroaspectos por posição, na categoria 'Mundo'.

Macro-Aspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total
WHAT_EXTRA	0	5	9	6	3	2	0	2	0	0	27
WHAT	14	0	0	0	1	1	0	0	0	0	16
COUNTERMEASURES	0	1	2	1	5	1	2	2	0	0	14
DECLARATION	2	2	2	3	1	1	0	0	2	0	13
HISTORY	2	0	2	1	2	2	2	0	2	0	13
CONSEQUENCE	4	1	3	1	0	1	0	0	0	0	10
PREDICTION	0	1	1	0	0	1	1	0	1	0	5
CONSEQUENCE_EXTRA	0	1	1	0	0	0	0	0	0	0	2
COUNTERMEASURES_EXTRA	0	0	0	0	0	0	0	0	0	1	1
DECLARATION_EXTRA	0	0	0	0	1	0	0	0	0	0	1
PREDICTION_EXTRA	0	0	0	0	0	0	0	0	1	0	1
Total de Ocorrências	**22**	**11**	**20**	**12**	**13**	**9**	**5**	**4**	**6**	**1**	**103**

O sumário abaixo, da coleção C1 do córpus, é prototípico desse tipo de organização de micro e macroaspectos, com os principais aspectos ocorrendo na primeira sentença.

[17 pessoas morreram após a queda de um avião na República Democrática do Congo.]**WHO_AFFECTED/WHAT/WHY/WHERE**
[14 dessas vítimas eram passageiros e três membros da tripulação, todos de nacionalidade russa.]**WHO_AFFECTED**
[Nenhuma vítima sobreviveu.]**CONSEQUENCE**
[O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta após se chocar com uma montanha, prejudicado pelo mau tempo.]**WHERE/WHY**
[O avião também levava cargas e minerais.]**WHAT_EXTRA**

As Tabelas 18 e 19 mostram as distribuições dos micro e macroaspectos por coleção, na categoria 'Mundo'. Pode-se notar que alguns aspectos ocorrem mais em algumas coleções e outros não ocorrem em coleção alguma.

Tabela 18. Distribuição dos microaspectos nas coleções da categoria 'Mundo'.

Micro-Aspectos	C1	C10	C12	C13	C14	C15	C18	C23	C26	C29	C32	C35	C46	C47	Total
WHERE	2	1	1	1	1	2	3	1	1	1	1	0	1	1	17
WHO_AGENT_EXTRA	0	3	2	1	0	0	0	0	0	2	0	4	1	4	17
WHO_AFFECTED	2	1	1	1	1	1	2	1	1	2	1	1	1	0	16
WHEN_EXTRA	0	2	0	0	0	0	1	3	3	1	2	0	0	0	12
WHY	2	1	1	2	1	3	0	1	0	0	0	0	0	1	12
WHEN	0	1	1	1	0	2	3	0	0	0	0	0	1	1	10
WHO_AFFECTED_EXTRA	0	3	0	1	0	0	2	1	0	0	0	2	1	0	10
WHERE_EXTRA	0	0	0	0	0	0	0	0	3	1	0	2	0	2	8
WHO_AGENT	0	1	0	0	0	0	1	0	0	1	0	1	0	1	5
GOAL	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2
GOAL_EXTRA	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2
WHY_EXTRA	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2
SITUATION	0	1	0	1	0	0	0	0	0	0	0	0	0	0	2
Total de Ocorrências	**6**	**14**	**6**	**8**	**3**	**8**	**13**	**7**	**8**	**9**	**4**	**12**	**6**	**11**	**115**

O microaspecto WHERE não ocorre em apenas uma coleção, C35, assim como WHO_AFFECTED, que não ocorre na coleção C47. Já os aspectos GOAL, GOAL_EXTRA, WHY_EXTRA e SITUATION ocorrem em apenas duas, com frequência 1 em cada uma delas.

Tabela 19. Distribuição dos macroaspectos nas coleções da categoria 'Mundo'.

Macro-aspectos	C1	C10	C12	C13	C14	C15	C18	C23	C26	C29	C32	C35	C46	C47	Total
WHAT_EXTRA	1	4	2	1	1	0	2	4	3	3	2	1	2	1	27
WHAT	1	1	1	1	1	1	3	1	1	1	1	1	1	1	16
COUNTERMEASURES	0	0	0	0	1	1	2	1	4	0	1	0	2	2	14
DECLARATION	0	0	3	1	2	0	1	0	1	3	0	0	0	2	13
HISTORY	0	1	0	2	1	0	2	1	0	1	1	4	0	0	13
CONSEQUENCE	1	0	1	0	0	0	0	1	2	0	3	0	2	0	10
PREDICTION	0	0	0	0	0	0	0	1	2	1	1	0	0	0	5
CONSEQUENCE_EXTRA	0	0	0	0	0	0	0	1	0	0	1	0	0	0	2
COUNTERMEASURES_EXTRA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
DECLARATION_EXTRA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
PREDICTION_EXTRA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Total de ocorrências	**3**	**9**	**7**	**5**	**6**	**2**	**10**	**10**	**13**	**9**	**10**	**6**	**7**	**6**	**103**

Já o macroaspecto WHAT ocorre em todos os grupos de textos. Isso também reforça as análises anteriores, de que os textos jornalísticos todos indicam seu assunto principal. Alguns macroaspectos 'EXTRA' aparecem raramente, como PREDICTION_EXTRA e DECLARATION_EXTRA, que ocorreram apenas uma vez em apenas um grupo de textos. O macroaspecto WHAT_EXTRA é o mais frequente, pois um texto, além do assunto principal, traz, na maioria das vezes, assuntos relacionados ao principal.

3.4. Política

A categoria 'Política' compreende somente 10 coleções de documentos e, portanto, 10 sumários manuais multidocumento para análise e etiquetação de aspectos. A distribuição de macro e microetiquetas para essa categoria é indicada no Quadro 7.

Quadro 7. Aspectos existentes na categoria 'Política'.

Macroaspectos	Microaspectos
1. COMPARISON	1. HOW(e)
2. CONSEQUENCE	2. WHEN(e)
3. COUNTERMEASURES	3. WHERE
4. DECLARATION	4. WHO_AFFECTED(e)
5. GOAL(e)	5. WHO_AGENT(e)
6. HISTORY	6. WHY(e)
7. PREDICTION	
8. SITUATION(e)	
9. WHAT(e)	

Pode-se notar, comparando-a com o Quadro 1, que nessa categoria não ocorrem os aspectos COMMENT e SCORE. Vários dos aspectos também remetem a informações que não são diretamente relacionadas ao tópico principal dos sumários analisados, caso do sufixo EXTRA adicionado à etiqueta e anotato com (e) no quadro.

Na categoria 'Política', a etiqueta WHO_AFFECTED representa pessoas, organizações ou instituições que sofrem uma acusação ou são afetadas pela ação de outrem. WHO_AGENT pode ser a pessoa, grupo ou instituição que executa alguma ação, ou pode ser considerado sinônimo de PERPETRATOR, originalmente usado para a categoria (2) na TAC, 'Ataques'. Em muitos textos de política, esse aspecto indica agentes de ataques verbais ou atitudes de políticos adversários que estão em campanha ou não.

A ocorrência de DECLARATION, em conjunto com WHO_AGENT em política, indica marcas claras de embate político: DECLARATION geralmente indica um ponto de vista político por meio de uma declaração da pessoa que fala (WHO_AGENT). O trecho abaixo mostra um exemplo anotado com essas etiquetas nos dois níveis de análise linguística - o macro e o microestrutural, respectivamente. Ele foi extraído da coleção C40 do CSTNews.

[O presidente do Senado, Renan Calheiros, disse que a decisão do procurador geral da República, Antonio Fernando de Souza, de investigar as denúncias contra ele, atende a um pedido que fez há cerca de um mês.]**WHO_AGENT/WHAT/DECLARATION**

Com exceção de GOAL e WHO (mais especificamente, WHO_AGENT e WHO_AFFECTED), as definições das demais etiquetas equivalem às da TAC. Os macroaspectos coincidentes com os da TAC são COMPARISON e CONSEQUENCE, muito embora na TAC não houvesse distinção de níveis macro e microestrutural entre as etiquetas. Os demais aspectos foram especificamente identificados para a categoria 'Política'.

A Tabela 20 mostra a representatividade de cada microaspecto no córpus, de forma isolada. Já a Tabela 21 apresenta a representatividade dos macroaspectos.

Tabela 20. Representatividade dos microaspectos para a categoria 'Política'.

Microaspectos	Número de ocorrências	Representatividade
WHO_AGENT_EXTRA	31	36%
WHO_AFFECTED_EXTRA	13	15%
WHO_AGENT	9	10%
WHY_EXTRA	9	10%
WHEN	7	8%
WHEN_EXTRA	7	8%
WHO_AFFECTED	5	6%
HOW	3	3%
WHY	2	2%
TOTAL	**86**	**100%**

Tabela 21. Representatividade dos macroaspectos para a categoria 'Política'.

Macroaspectos	Número de Ocorrências	Representatividade
WHAT_EXTRA	51	51%
DECLARATION	16	16%
WHAT	12	12%
COMPARISON	5	5%
PREDICTION	5	5%
CONSEQUENCE	3	3%
GOAL_EXTRA	3	3%
HISTORY	2	2%
SITUATION	2	2%
GOAL	1	1%
COUNTERMEASURES	0	0%
TOTAL	**100**	**100%**

A quantidade relativamente superior da microetiqueta WHO_AGENT, quer quando se trata do evento principal, quer quando se trata de eventos secundários (totalizando 46% dos casos de microetiquetas), é esperada em textos sobre política: eles tratam de descrição de eventos envolvendo figurantes ativos da política nacional ou de organizações, os quais são mencionados como agentes (de embates, discursos ou decisões políticas). A superioridade de WHO_AGENT_EXTRA em relação a WHO_AGENT (ocorrendo praticamente três vezes mais) indica também um fenômeno recorrente nesses textos: é comum que haja um figurante envolvido em um evento principal, o qual é descrito, em geral, na primeira sentença. Esta, por sua vez, dá origem à descrição ou menção a uma sequência de eventos secundários. Nesses casos, todos os figurantes aparecerão como WHO_AGENT_EXTRA. WHO_AFFECTED_EXTRA supera mesmo os casos de WHO_AGENT, o que leva a crer que, é mais frequente a menção a um objeto/pessoa paciente quando se detalham eventos ou embates políticos do que quando se comentam eventos ou embates principais, no texto.

São expressivas também em política as ocorrências de WHEN (em ambas as formas, principal ou EXTRA) e WHY. Porém, este aspecto só ocorre expressivamente quando ligado a eventos secundários: WHY_EXTRA ocorre em 8% dos casos; em eventos principais, só em 2% deles. Novamente, isso sugere que mencionar causas de algum evento ou embate político só é relevante quando se entra em detalhes no texto. Em geral, WHY está associado a justificativas ou explicações que figurantes políticos apresentam – e, neste caso, são reproduzidas pelo autor do texto – no gênero jornalístico. Já a ocorrência de WHEN pode ser explicada segundo a ótica de que eventos políticos ocorrem sobretudo relacionados a ocasiões específicas (reunião, campanha pré-eleitoral, comício, etc.), marcadas por WHEN.

A Tabela 20 indica que, no discurso político, WHO_AGENT e WHEN são os únicos aspectos que ocorrem expressivamente relacionados ao tópico principal do texto, enquanto, ligados a tópicos secundários, além desses mesmos aspectos, ocorrem expressivamente também WHY e WHO_AFFECTED.

A distribuição de macroaspectos (Tabela 21) mostra que formas secundárias de WHAT (51% de WHAT_EXTRA) são muito mais frequentes que as principais (12% de WHAT). Isso sugere que, na elaboração do assunto principal, é muito comum se fazer menção a assuntos secundários, no discurso político. Elas são até mais frequentes que as etiquetas ligadas aos assuntos principais, como mostram as

ocorrências de DECLARATION (16% dos casos), COMPARISON e PREDICTION (ambas com 5% dos casos). A alta frequência de DECLARATION indica, sobretudo, a tendência de se exibir nos sumários os discursos ou falas de alguma personalidade no contexto político brasileiro. Os demais aspectos, incluindo COMPARISON e PREDICTION, ocorrem de forma inexpressiva no córpus de política (totalizam aproximadamente 22% dos casos). Vale notar que apesar de COUNTERMEASURES aparecer com frequência nula, isso se dá devido à aproximação numérica para valores inteiros de frequência. Na verdade, o valor 0 indica a representatividade insignificante desse aspecto no córpus: ele ocorre em duas sentenças de um único sumário multidocumento (coleção C20). É notável que, dentre os inexpressivos, COMPARISON ocorra com a mais alta frequência. Em geral, no discurso político esse aspecto refere-se à comparação de desempenho entre candidatos a eleições ou estatísticas de preferências eleitorais.

As Tabelas 22 e 23 mostram, respectivamente, a distribuição dos micro e macroaspectos em relação a sua posição em todos os sumários. São, no máximo, 9 sentenças em cada sumário.

Tabela 22. Distribuição dos microaspectos por posição, na categoria 'Política'.

Microaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	Total
WHO_AGENT_EXTRA	0	5	5	4	7	6	2	1	1	31
WHO_AFFECTED_EXTRA	0	4	3	2	1	2	1	0	0	13
WHO_AGENT	8	0	0	0	1	0	0	0	0	9
WHY_EXTRA	0	2	0	2	1	3	0	1	0	9
WHEN	6	0	0	0	1	0	0	0	0	7
WHEN_EXTRA	0	0	3	1	1	0	0	2	0	7
WHO_AFFECTED	5	0	0	0	0	0	0	0	0	5
HOW	0	0	0	0	0	1	1	1	0	3
WHY	2	0	0	0	0	0	0	0	0	2
Total de ocorrências	**21**	**11**	**11**	**9**	**12**	**12**	**4**	**5**	**1**	**86**

Tabela 23. Distribuição dos macroaspectos por posição, na categoria 'Política'

Macroaspectos	S1	S2	S3	S4	S5	S6	S7	S8	S9	Total
WHAT_EXTRA	0	10	10	9	8	7	4	2	1	51
DECLARATION	3	3	2	2	3	3	0	0	0	16
WHAT	10	0	0	1	1	0	0	0	0	12
COMPARISON	1	1	1	0	0	1	1	0	0	5
CONSEQUENCE	0	1	0	0	1	1	0	0	0	3
COUNTERMEASURES	0	0	0	0	0	0	0	0	0	0
PREDICTION	0	1	0	2	1	0	1	0	0	5
GOAL	1	0	0	0	0	0	0	0	0	1
GOAL_EXTRA	0	0	0	1	2	0	0	0	0	3
HISTORY	0	0	0	1	1	0	0	0	0	2
SITUATION	1	0	0	0	1	0	0	0	0	2
Total de ocorrências	**16**	**16**	**13**	**16**	**18**	**12**	**6**	**2**	**1**	**100**

As distribuições acima mostram que etiquetas principais são predominantes em primeiras sentenças, enquanto as secundárias são predominantes nas demais posições de um sumário. Isso confirma o fato de sentenças iniciais indicarem a notícia 'lead'. Sentenças seguintes, no caso, apresentam assuntos secundários ou complementares à notícia principal. Essa distribuição fica evidente quando se analisa a frequência de

WHAT, com 10 ocorrências nas primeiras sentenças dos sumários. Ela indica que, em 83% dos casos, a informação principal é apresentada na primeira sentença. Já sua forma secundária (WHAT_EXTRA) nunca ocorre nessa posição, embora seja altamente frequente a partir das segundas sentenças.

Nota-se, ainda, que a microetiqueta WHO_AGENT_EXTRA (Tabela 22) aparece predominantemente a partir de terceiras sentenças (e nunca em posições que poderiam ser correspondentes à notícia 'lead'). Informações desse tipo seguem expressivamente até o fim do documento: ela ocorre só 5 vezes no começo dos textos, mas 26 vezes a partir da terceira sentença deles. É importante notar, ainda, que, das 21 ocorrências de microetiquetas em primeiras sentenças, as mais expressivas são WHO_AGENT (38% dos casos), WHEN (29%) e WHO_AFFECTED (24%). Esse fato pode levar à definição de um *template* particular para a SAM, como veremos na Seção 6. Verifica-se, ainda, que a microetiqueta HOW é a única que tem todas as suas ocorrências somente no final dos sumários (no caso geral, nas 3 últimas sentenças). Já a microetiqueta WHY comporta-se exatamente ao contrário: suas 2 ocorrências verificam-se em primeiras sentenças. No entanto, devido a sua baixa frequência, dificilmente ela será relevante para se determinar um padrão de organização de conteúdo.

A análise da distribuição de macroaspectos (Tabela 23) também indica que, das 100 ocorrências, a maioria está predominantemente distribuída a partir das segundas sentenças dos sumários, com exceção de WHAT, que ocorre em 83% dos casos em primeiras sentenças. Ressalta-se, ainda, que este caso é bastante peculiar: os demais macroaspectos que ocorrem em primeiras sentenças (somente 4, dos 11 que aparecem nos sumários de política), totalizam 6 ocorrências, o que é bastante inexpressivo, quando comparados a WHAT: todos juntos chegam a 37% dos casos relativos a primeiras sentenças.

As Tabelas 24 e 25 mostram, respectivamente, a distribuição dos micro e macroaspectos nas 10 coleções da categoria 'Política'.

Tabela 24. Distribuição dos microaspectos nas coleções da categoria 'Política'.

Microaspectos	C2	C9	C16	C17	C20	C40	C42	C43	C44	C50	Total
WHO_AGENT_EXTRA	0	2	3	4	2	2	3	5	6	4	31
WHO_AGENT	0	1	1	2	0	1	1	1	1	1	13
WHEN	1	1	0	1	1	0	1	0	1	1	9
WHY	0	1	1	0	0	0	0	0	0	0	9
WHEN_EXTRA	0	0	1	0	1	0	2	0	1	2	7
WHO_AFFECTED	1	1	0	0	0	0	1	1	1	0	7
WHO_AFFECTED_EXTRA	5	3	3	0	0	1	0	1	0	0	5
HOW	0	0	0	0	3	0	0	0	0	0	3
WHY_EXTRA	0	0	2	2	0	0	1	2	1	1	2
Total de ocorrências	**7**	**9**	**11**	**9**	**7**	**4**	**9**	**10**	**11**	**9**	**86**

Tabela 25. Distribuição dos macroaspectos nas coleções da categoria 'Política'.

Macroaspectos	C2	C9	C16	C17	C20	C40	C42	C43	C44	C50	Total
WHAT_EXTRA	6	5	5	4	3	3	4	6	8	7	51
DECLARATION	1	0	0	4	0	2	0	2	4	3	16
WHAT	1	1	1	2	2	1	1	1	1	1	12
COMPARISON	5	0	0	0	0	0	0	0	0	0	5
CONSEQUENCE	0	0	3	0	0	0	0	0	0	0	3
COUNTERMEASURES	0	0	0	0	0	0	0	0	0	0	0
PREDICTION	1	0	2	0	0	0	0	2	0	0	5
GOAL	0	0	1	0	0	0	0	0	0	0	1
GOAL_EXTRA	0	0	0	0	1	0	0	1	0	1	3
HISTORY	1	1	0	0	0	0	0	0	0	0	2
SITUATION	0	0	0	1	0	0	1	0	0	0	2
Total de ocorrências	**15**	**7**	**12**	**11**	**6**	**6**	**6**	**12**	**13**	**12**	**100**

A distribuição de microaspectos e macroaspectos (Tabelas 24 e 25) mostram que alguns aspectos apresentam um comportamento com pouca variação enquanto outros apresentam grande variação em relação aos diferentes tipos de sumários que compõem o córpus. Para os microaspectos, se destacam os aspectos WHO_AGENT_EXTRA, WHO_AGENT e WHEN, que apresentaram um comportamento similar em todos os sumários do córpus. Para os macroaspectos, destacam-se os aspectos WHAT_EXTRA e WHAT. O aspecto COUNTERMEASURES não foi evidenciado no córpus.

Observa-se, ainda, que algumas coleções apresentam ditribuições particulares. Por exemplo, C2 apresenta um número significativo de WHO_AFFECTED_EXTRA (5 ocorrências), se comparada com as ocorrências nas demais coleções. Essa mesma coleção também contém todas as ocorrências do macroaspecto COMPARISON (total de 5). Em particular, o sumário multidocumento de C2 reporta o resultado de uma pesquisa de opinião sobre as eleições próximas e, assim, o texto trata de comparações entre as chances dos candidatos envolvidos na pesquisa.

Com base nas distribuições apresentadas nesta seção, foi possível derivar padrões de organização de conteúdo gerais para cada categoria, que são descritos na Seção 6.

4. SÍNTESE DE OCORRÊNCIA DOS ASPECTOS PARA TODAS AS CATEGORIAS

A Tabela 26 mostra a representatividade dos aspectos para cada categoria. Os aspectos são ordenados pelo número total de ocorrências no córpus e correspondente frequência, que indica a representatividade de cada um. A média de ocorrências (Média) e o desvio padrão (DesvPad) também são indicados.

Tabela 26. Representatividade dos aspectos principais para cada categoria.

(a) Cotidiano

Aspecto	Quantidade	Frequência	Média	DesvPad
DECLARATION	28	13%	2.0	1.7
WHAT_EXTRA	25	11%	1.8	1.2
CONSEQUENCE	22	10%	1.6	2.3
WHO_AGENT	22	10%	1.6	2.0
WHO_AGENT_EXTRA	18	8%	1.3	1.3
WHAT	13	6%	0.9	0.6
WHEN	13	6%	0.9	0.5
WHERE	12	5%	0.9	0.5
WHEN_EXTRA	11	5%	0.8	1.6
HISTORY	10	5%	0.7	1.8
WHERE_EXTRA	9	4%	0.6	1.1
WHO_AFFECTED	9	4%	0.6	0.8
COMMENT	6	3%	0.4	1.0
HOW	5	2%	0.4	0.8
COUNTERMEASURES	4	2%	0.3	0.6
SITUATION	4	2%	0.3	0.5
WHY	2	1%	0.1	0.3
WHY_EXTRA	2	1%	0.1	0.3
COMPARISON	1	0%	0.1	0.3
GOAL	1	0%	0.1	0.3
HOW_EXTRA	1	0%	0.1	0.3
PREDICTION	1	0%	0.1	0.3
WHO_AFFECTED_EXTRA	1	0%	0.1	0.3
Total	**220**			

(b) Esporte

Aspecto	Quantidade	Frequência	Média	DesvPad
WHO_AGENT	19	11%	1.9	1.4
HOW	17	10%	1.7	3.2
COMMENT	16	9%	1.6	2.0
WHAT_EXTRA	14	8%	1.4	1.4
WHO_AGENT_EXTRA	13	8%	1.3	1.4
WHAT	10	6%	1	0.0
WHEN	10	6%	1	1.5
CONSEQUENCE	8	5%	0.8	0.7
SCORE	7	4%	0.7	0.5
SITUATION	6	4%	0.6	0.7
PREDICTION	5	3%	0.5	0.9
SITUATION_EXTRA	5	3%	0.5	0.7
WHERE	5	3%	0.5	0.5
WHO_AFFECTED	5	3%	0.5	1.2
COMMENT_EXTRA	4	2%	0.4	0.7
HISTORY	4	2%	0.4	0.7
WHEN_EXTRA	4	2%	0.4	0.5
WHERE_EXTRA	4	2%	0.4	0.7

SCORE_EXTRA	3	2%	0.3	0.9
COMPARISON	2	1%	0.2	0.4
CONSEQUENCE_EXTRA	2	1%	0.2	0.4
DECLARATION	2	1%	0.2	0.4
WHY	2	1%	0.2	0.6
GOAL	1	1%	0.1	0.3
WHO_AFFECTED_EXTRA	1	1%	0.1	0.3
WHY_EXTRA	1	1%	0.1	0.3
Total	**170**			

(c) Mundo

Aspecto	Quantidade	Frequência	Média	DesvPad
WHAT_EXTRA	27	12%	1.9	1.2
WHERE	17	8%	1.2	0.7
WHO_AGENT_EXTRA	17	8%	1.2	1.5
WHAT	16	7%	1.1	0.5
WHO_AFFECTED	16	7%	1.1	0.5
COUNTERMEASURES	14	6%	1	1.1
DECLARATION	13	6%	0.9	1.1
HISTORY	13	6%	0.9	1.1
WHEN_EXTRA	12	6%	0.9	1.1
WHY	12	6%	0.9	0.9
CONSEQUENCE	10	5%	0.7	1.0
WHEN	10	5%	0.7	0.9
WHO_AFFECTED_EXTRA	10	5%	0.7	1.0
WHERE_EXTRA	8	4%	0.6	1.0
PREDICTION	5	2%	0.4	0.6
WHO_AGENT	5	2%	0.4	0.5
CONSEQUENCE_EXTRA	2	1%	0.1	0.3
GOAL	2	1%	0.1	0.3
GOAL_EXTRA	2	1%	0.1	0.3
SITUATION	2	1%	0.1	0.3
WHY_EXTRA	2	1%	0.1	0.3
COUNTERMEASURES_EXTRA	1	0%	0.1	0.3
DECLARATION_EXTRA	1	0%	0.1	0.3
PREDICTION_EXTRA	1	0%	0.1	0.3
Total	**218**			

(d) Política

Aspecto	Quantidade	Frequência	Média	DesvPad
WHAT_EXTRA	51	27%	5.1	1.6
WHO_AGENT_EXTRA	31	17%	3.1	1.6
DECLARATION	16	9%	1.6	1.6
WHO_AFFECTED_EXTRA	13	7%	1.3	1.7
WHAT	12	6%	1.2	0.4
WHO_AGENT	9	5%	0.9	0.5
WHY_EXTRA	9	5%	0.9	0.8
WHEN	7	4%	0.7	0.5

WHEN_EXTRA	7	4%	0.7	0.8
COMPARISON	5	3%	0.5	1.5
PREDICTION	5	3%	0.5	0.8
WHO_AFFECTED	5	3%	0.5	0.5
CONSEQUENCE	3	2%	0.3	0.9
GOAL_EXTRA	3	2%	0.3	0.5
HOW	3	2%	0.3	0.9
HISTORY	2	1%	0.2	0.4
SITUATION	2	1%	0.2	0.4
WHY	2	1%	0.2	0.4
GOAL	1	1%	0.1	0.3
Total	**186**			

A Tabela 27 agrega todos os casos que ocorrem no córpus, sem distinção por micro e macroaspectos ou por categoria. O número médio de ocorrências de cada aspecto no córpus, assim como seu desvio padrão, também são exibidos.

Tabela 27. Representatividade dos aspectos principais no córpus.

Aspecto	Quantidade	Frequência	Média	DesvPad
WHAT_EXTRA	**117**	15%	2.67	2.50
WHO_AGENT_EXTRA	**79**	10%	1.88	2.27
DECLARATION	**59**	7%	1.24	1.46
WHO_AGENT	**55**	7%	1.51	2.91
WHAT	**51**	6%	1.24	1.35
CONSEQUENCE	**43**	5%	1.04	1.80
WHEN	**40**	5%	1.02	1.57
WHO_AFFECTED	**35**	4%	0.82	1.02
WHEN_EXTRA	**34**	4%	0.78	1.22
WHERE	**34**	4%	0.80	0.90
HISTORY	**29**	4%	0.67	1.30
HOW	**25**	3%	0.86	2.87
WHO_AFFECTED_EXTRA	**25**	3%	0.53	1.05
COMMENT	**22**	3%	0.78	2.51
WHERE_EXTRA	**21**	3%	0.51	1.01
COUNTERMEASURES	**18**	2%	0.37	0.80
WHY	**18**	2%	0.41	0.73
PREDICTION	**16**	2%	0.43	0.95
SITUATION	**14**	2%	0.41	0.95
WHY_EXTRA	**14**	2%	0.31	0.58
COMPARISON	**8**	1%	0.20	0.78
SCORE	**7**	1%	0.29	1.03
GOAL	**5**	1%	0.12	0.33
GOAL_EXTRA	**5**	1%	0.10	0.30
SITUATION_EXTRA	**5**	1%	0.20	0.78
COMMENT_EXTRA	**4**	1%	0.16	0.65
CONSEQUENCE_EXTRA	**4**	1%	0.12	0.39
SCORE_EXTRA	**3**	0%	0.12	0.59
COUNTERMEASURES_EXTRA	**1**	0%	0.02	0.14

DECLARATION_EXTRA	**1**	0%	0.02	0.14
HOW_EXTRA	**1**	0%	0.02	0.14
PREDICTION_EXTRA	**1**	0%	0.02	0.14
Totais	**794**	**100%**		

5. PROBLEMAS DE ANOTAÇÃO DOS SUMÁRIOS DAS QUATRO CATEGORIAS

Nesta seção, apresentam-se os problemas e decisões gerais de anotação mais significativos a partir da anotação de cada categoria de sumários multidocumento. Os seguintes casos são considerados, alguns já mencionados neste relatório (os três primeiros foram considerados problemáticos e foram tratados isoladamente para cada categoria):

i. Os aspectos não ocorrem uniformemente entre as categorias.
ii. Nem todos os aspectos elencados para todos os sumários do CSTNews (veja Quadro 1) ocorrem em cada categoria.
iii. A classificação em macro e microetiquetas também não é consenso: em algumas categorias os aspectos estão associados, mais frequentemente, a macroestruturas, mas em outras, estão associados a microestruturas.
iv. Na maioria das ocorrências do aspecto DECLARATION, o aspecto WHO_AGENT também aparece. Do ponto de vista prático, isso resultou em considerar, para esses casos, uma implicação lógica entre um segmento DECLARATION e um segmento WHO_AGENT: DECLARATION ⇒ WHO_AGENT. No entanto, como não foi possível generalizar essa regra, as duas etiquetas são mantidas no córpus, quando pertinente.
v. Caso diferente ocorre com macroetiquetas que já envolvem o aspecto WHAT. DECLARATION é significativamente representativa desse problema: uma declaração em si é um evento e representa um WHAT mais especializado, não sendo necessário explicitar essa etiqueta WHAT na anotação dessa declaração.
vi. Finalmente, após uma etapa preliminar de anotação de uma categoria, foram modificadas as próprias definições originais para alguns casos, visando aprimorar sua clareza em função das evidências fornecidas pelos dados textuais. Um dos principais aprimoramentos consistiu em agregar mais de um aspecto, dos originalmente sugeridos, em um único, resultando na criação de uma nova etiqueta que, em geral, é mais abrangente – isso consiste em um processo de generalização. Por exemplo, em 'Cotidiano', DAMAGE e WHAT_AFFECTED foram incorporados a CONSEQUENCE; PERPETRATOR passou a ser representado genericamente por WHO_AGENT. Outro aprimoramento consistiu no processo de especificação: uma etiqueta genérica foi desmembrada, como WHO, que deu origem a WHO_AGENT e WHO_AFFECTED.

Em cada seção abaixo são relatadas as variações específicas para cada categoria do córpus, fazendo-se menção aos casos acima.

5.1. Cotidiano

O conjunto de aspectos identificado para sumários de 'Cotidiano' resulta parcialmente distinto do conjunto utilizado na anotação preliminar, para essa mesma

categoria (Zacarias *et al*., 2012). Aspectos já considerados nessa primeira etapa foram identificados, porém, com variações significativas em sua conceitualização, o que levou à sua redefinição particular para o contexto do Cotidiano. Os aspectos DAMAGE e WHAT_AFFECTED, por exemplo, foram incorporados a CONSEQUENCE, e o aspecto PERPETRATOR, a WHO_AGENT (caso vi). Na anotação preliminar também foi criado o aspecto IMPORTANCE, relativo à menção, no texto, à importância de um fato principal. Na redefinição do processo, IMPORTANCE foi incorporado a PREDICTION (caso vi), como exemplificado a seguir (C39):

> [Esse trabalho permitirá avaliar os riscos aos quais estão expostos os passageiros e tripulantes de aeronaves e tomar medidas necessárias para aumentar a segurança no setor aéreo.]**PREDICTION**

Devido à decisão de se marcar o aspecto WHO_AGENT junto ao aspecto DECLARATION (caso iv), outros aspectos de carga semântica similar foram desconsiderados em situações específicas. No exemplo a seguir (C49), pode-se dizer que Lula (agente da declaração) também foi afetado pela atitude do público e, portanto, deveria ser marcado com WHO_AFFECTED. Porém, como essa essa informação não está diretamente explicitada no texto, optou-se por anotar somente o WHO_AGENT.

> [Lula admitiu que ficou chateado com a atitude do público, mas garantiu que sua relação com o povo do Rio de Janeiro não será alterada.]**WHO_AGENT/DECLARATION**

Outra dificuldade relaciona-se a estabelecer o fato principal em alguns sumários. A sentença a seguir é a primeira sentença do sumário do grupo C5 e ela sugere a ocorrência dupla de declarações. Por essa razão, decidiu-se considerar as duas declarações igualmente em relevo e, portanto, aludindo a um único fato principal: 'a indicação de Solange Vieira para um cargo'. Entretanto, a segunda declaração poderia ser a única relevante, pois ela é marcada por um verbo claramente associado a declarações ('informar que'), mais taxativo do que o primeiro ('indicar que'). Este pode ser um problema de se identificar, para mesmas etiquetas, seus distintos graus de relevância no texto (caso i).

> [Corriam boatos de que o ministro da defesa, Nelson Jobim, indicaria a economista Solange Vieira para um cargo de diretoria na Agência Nacional de Aviação Civil (Anac), mas, em jantar que celebrou os 50 anos da Rede RBS em Brasília, na noite de desta terça-feira, informou que a mesma será a nova presidente do órgão.] **DECLARATION/WHO_AGENT/SITUATION/WHEN/WHO_AFFECTED/WHERE**

5.2. Esporte

Além da dificuldade de se enquadrar os aspectos dos sumários sobre esportes no rol de aspectos preliminar (cf. Jorge et al., 2012) e, mesmo, de se identificarem alguns dos aspectos originais para esta categoria (caso ii), também houve dificuldade em classificá-los em micro e macroaspectos (caso iii).

Alguns aspectos foram mais detalhados e outros foram generalizados a partir da anotação preliminar. WHO, por exemplo, foi desmembrado em WHO_AGENT e WHO_AFFECTED; RESULT foi especificado sob o nome de SCORE, com o objetivo de capturar os resultados específicos de competições esportivas. Já CHAMPIONSHIP e SCHEDULE foram generalizados para SITUATION e PREDICTION, respectivamente.

Quanto à classificação como micro e macroaspectos, observaram-se as seguintes diferenças em relação à classificação geral apresentada no Quadro 1: o aspecto HOW, tipicamente de nível microestrutural no CSTNews, ocorre como macroaspecto na categoria 'Esporte', como os dois exemplos abaixo ilustram (dos grupos C27 e C24, respectivamente). Em ambos os casos o aspecto HOW emerge da relação entre as sentenças ilustradas e as primeiras sentenças dos respectivos sumários, que expressam os conteúdos principais, marcados com o aspecto WHAT. No primeiro exemplo, HOW indica o modo como a vitória da seleção brasileira de futebol frente à equatoriana se deu; no segundo, indica o modo como se deu a conquista da medalha de ouro por Fabiana Murer. Entretanto, o segundo exemplo pode ser analisado sob outra ótica: a expressão 'conseguir o ouro' já remete ao evento principal (conquista da medalha de ouro) e, assim, o foco em como Fabiana Murer conseguiu essa conquista passa a ser microestrutural. O aspecto HOW passa a ser indicado, portanto, pelo sintagma preposicional 'em três tentativas' e não mais pela sentença inteira.

O jogo contou com belas atuações de craques como Ronaldinho e Kaká.]**HOW/COMMENT**

[Fabiana conseguiu o ouro em três tentativas.]**WHO_AGENT/HOW**

De forma similar, SITUATION, tipicamente macroaspecto no CSTNews, ocorre também como microaspecto na categoria 'Esporte', como mostram os dois exemplos a seguir (ambos de C8):

[A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei06.] **WHO_AGENT/WHAT/SCORE/WHERE/CONSEQUENCE/SITUATION**

[A seleção brasileira ainda enfrentará portugueses e finlandeses na fase de classificação.]**WHO_AGENT/PREDICTION/SITUATION_EXTRA**

Eventos desse tipo não haviam sido etiquetados na anotação preliminar dos sumários. Com a definição de SITUATION (cf. Quadro 3, *ocasião em que ocorreu um fato/evento*), foi possível delimitar unidades de informação relativas a esse conceito e identificar, assim, a correspondência com esse aspecto.

Caso peculiar envolve ainda os aspectos WHO_AGENT e WHO_AGENT_EXTRA, cuja anotação nem sempre foi simples, principalmente diante de sujeitos compostos e coletivos, os quais incorporam diferentes conotações, mas podem estar interrelacionados, conforme a discussão baseada nos dois grupos de exemplos a seguir.

As primeiras duas sentenças são a 2ª. e a 3ª. sentenças do sumário do grupo C27, respectivamente. Na primeira, vê-se que o sujeito composto – 'as seleções de vôlei e futebol' – expressa a informação rotulada por WHO_AGENT. Na segunda, tem-se um sujeito simples ('a seleção de vôlei de Bernardinho'), que poderia ser associado a WHO_AGENT_EXTRA. Neste caso, decidiu-se por anotar o sujeito simples também como WHO_AGENT, pois essa sentença também se refere ao fato principal descrito na sentença anterior, identificado por WHAT.

[As seleções de vôlei e futebol conquistaram a Liga Mundial e a Copa América, respectivamente.]**WHO_AGENT/WHAT**

[A seleção de vôlei de Bernardinho manteve sua hegemonia mundial derrotando a Rússia, mesmo perdendo o primeiro set.] **WHO_AGENT/CONSEQUENCE/COMMENT**

Nas outras duas sentenças ilustradas, de C38, há um sujeito coletivo na primeira sentença – 'a equipe de revezamento 4x200 metros livres' – mas, na segunda, há um sujeito simples – 'Thiago Pereira' – o qual é um membro da coleção mencionada na primeira (i.e., é um integrante da equipe). Diferentemente do grupo anterior, em que há uma correferência explícita com parte do sujeito composto, levando à identificação de um único aspecto, a decisão sobre sujeitos coletivos ou simples foi diferente: o sujeito coletivo foi anotado com WHO_AGENT, mas o sujeito simples, com WHO_AGENT_EXTRA. Isso se justifica principalmente pela segunda sentença veicular informações que não são relativas à informação principal (WHAT) da primeira sentença. Deve-se, portanto, à existência do WHAT_EXTRA atribuído à segunda sentença, como se vê abaixo.

[A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos.]**WHO_AGENT/WHAT/WHEN/SITUATION**

[Pouco antes, Thiago Pereira havia conquistado a segunda medalha de ouro brasileira no dia na final dos 400m medley, superando o norte-americano Robert Margalis e o canadense Keith Beavers]**WHEN_EXTRA/WHO_AGENT_EXTRA/WHAT_EXTRA/SITUATION_EXTRA**

5.3. Mundo

Os principais problemas encontrados na anotação dos sumários da categoria 'Mundo' foram relativos à segmentação, mas também aos problemas recorrentes antes mencionados: categorização dos aspectos em micro e macroaspectos (caso iii) e definição ou reconhecimento de alguns aspectos que emergem dos textos (casos vi e ii, respectivamente).

Para a recuperação do contexto conceitual visando à identificação dos aspectos, não se considerou inicialmente a segmentação sentencial. Os segmentos foram delimitados simplesmente a partir do conceito que sugeriam. É dessa forma que se pode reconhecer, na sentença abaixo (C12), os aspectos correspondentes aos

seis segmentos assinalados. Esse estilo de segmentação e etiquetação é mais informativo, por indicar claramente qual o segmento textual que dá origem ao aspecto e também por permitir um tratamento mais refinado da anotação. A indicação contígua dos aspectos nos segmentos textuais seria ainda mais clara, pois já enfatizaria a informação relevante para se reconhecer um certo aspecto e também seguiria a ordem já definida sintaticamente na superfície textual.

A anotação ao fim da sentença não oferece essa clareza de identificação dos aspectos, o qual se torna ainda mais complexo quando um mesmo aspecto é sugerido mais de uma vez na mesma sentença, pois sua etiqueta aparece apenas uma vez ao fim da sentença. Isso dificulta a identificação tanto da ordem de ocorrência dos aspectos na superfície textual quanto dos segmentos que os veiculam.

> [Ao menos 549 pessoas morreram, 3.043 ficaram feridas e outras 295 ainda estão desaparecidas em consequência das enchentes que atingiram a Coréia do Norte em julho, segundo um jornal japonês pró-Pyongyang.]**WHO_AFFECTED/WHAT/WHY/WHERE/WHEN/DECLARATION**

Também na categoria 'Mundo' houve aspectos redefinidos, quer por generalizações, quer por maior detalhamento (caso vi). Houve também alguns uniformizados meramente por sua renomeação, como DAMAGES, renomeado para CONSEQUENCE, e SOURCE, renomeado para DECLARATION. WHO também foi desmembrado em WHO_AGENT e WHO_AFFECTED, como na categoria 'Esporte'. Outros aspectos foram ignorados em 'Mundo", como SCORE e COMMENT, pois não são comuns a essa categoria.

A identificação de micro e macroaspectos também não se deu de modo uniforme em 'Mundo' (caso i). Os dois exemplos abaixo (C32 e C23, respectivamente) mostram que há ocorrências de microaspectos que, em sua maioria, são considerados macroestruturais (Quadro 3) e que, na mesma categoria, essa classificação é variável. O aspecto CONSEQUENCE, no primeiro caso, refere-se a apenas uma parte da sentença anotada (sublinhada), sendo considerado, portanto, um microaspecto. No segundo caso, ele é macroestrutural.

> [Um terremoto, de 6.8 graus na escala Ritcher, atingiu a região de Niigata, no Japão, causando um incêndio e vazamento de materiais radioativos na maior usina de energia nuclear do mundo.]**WHAT/WHERE/CONSEQUENCE**

> [A chuva torrencial que atinge o Reino Unido encobriu estradas e milhares de pessoas estão sem fornecimento de eletricidade e de água potável em decorrência da pior enchente nos últimos 60 anos no país.] **WHAT/WHERE/CONSEQUENCE/WHO_AFFECTED/WHY/HISTORY**

GOAL é outro aspecto que, no caso genérico, é um macroaspecto, mas em 'Mundo' aparece como microaspecto (C35).

[A Operação Farrapos, da Polícia Federal, com o objetivo de desarticular uma quadrilha internacional de tráfico de drogas, prendeu 14 dos 17 suspeitos, após 2 anos de investigações.]**WHAT/WHO_AGENT/GOAL/WHO_AFFECTED**

O mesmo ocorre com HISTORY no sumário C23:

[A chuva torrencial que atinge o Reino Unido encobriu estradas e milhares de pessoas estão sem fornecimento de eletricidade e de água potável em decorrência da pior enchente nos últimos 60 anos no país.]**WHAT/WHERE/CONSEQUENCE/WHO_AFFECTED/WHY/HISTORY**

Também a coordenação das orações abaixo nesse mesmo sumário indica que somente a segunda oração incorpora uma predição, tornando PREDICTION um microaspecto em 'Mundo'.

[Na sexta-feira, choveu muito acima do esperado e há previsão de mais tempestades hoje.]**WHEN_EXTRA/WHAT_EXTRA/PREDICTION**

SITUATION muitas vezes é assinalado por adjuntos adverbiais, como indica o segmento 'Nesta batalha' (C10). Neste caso, ele é claramente um microaspecto.

[Nesta batalha, 15 soldados israelenses morreram ao serem atingidos por um míssel.]**SITUATION/WHO_AFFECTED_EXTRA**

De forma inversa, há aspectos tipicamente microestruturais que aparecem como macroestruturais em 'Mundo'. É o caso de WHY, no sumário C15:

[Inicialmente, acreditou-se que a explosão foi causada por vazamento de um botijão de gás.]**WHY**

Embora os exemplos anteriores mostrem a diversidade para se classificar vários micro e macroaspectos, em relação à classificação conceitual apresentada nos quadros 2 e 3, os aspectos WHAT e COUNTERMEASURES sempre apareceram como macroestruturais.

Em relação à clareza de definição de alguns aspectos também houve dificuldades. GOAL e WHY, por exemplo, são muito semelhantes e sua diferenciação é muito subjetiva em muitos casos. O mesmo se dá com CONSEQUENCE e PREDICTION: no mesmo sumário C23 já ilustrado e parcialmente reproduzido abaixo, o segmento sublinhado pode ser interpretado como uma consequência das muitas chuvas (menção feita na 1ª. sentença) ou como uma predição de transbordamento dos rios.

... [Na sexta-feira, choveu muito acima do esperado e há previsão de mais tempestades hoje.]**WHEN_EXTRA/WHAT_EXTRA/PREDICTION**

[Os dois maiores rios do Reino Unido, Severn e Tâmisa, ameaçam transbordar nesta segunda, agravando ainda mais a situação.]**WHAT_EXTRA/WHEN_EXTRA/CONSEQUENCE_EXTRA**

5.4. Política

A principal dificuldade encontrada no processo de anotação dos textos dessa categoria foi identificar a informação principal, para, a partir dela, classificar as demais sentenças dos sumários. O texto abaixo indica as três primeiras sentenças do sumário C20, para ilustrar a distinção entre etiquetas principais e secundárias (EXTRA).

[Ocorre hoje a votação da PEC (Proposta de Emenda Constitucional) que prorroga a cobrança da CPMF (Contribuição Provisória sobre Movimentação Financeira) até 2011 e mantém a alíquota de 0,38%.]**WHAT/WHEN**
[Para limpar a pauta da Câmara e garantir a votação, o governo decidiu revogar três das quatro medidas provisórias que estavam com o prazo de aprovação vencido.] **WHO_AGENT_EXTRA/WHAT_EXTRA/COUNTERMEASURES**
[A quarta medida foi aprovada nesta madrugada.] **WHAT_EXTRA/WHEN_EXTRA**

Também foi complexa a delimitação de segmentos superficiais que indicassem os conceitos relativos a cada aspecto. Optou-se por classificá-los em micro ou macroaspectos, como propõe o Quadro 1, justamente em alusão aos segmentos que servem de base para identificá-los. Sempre foi claro, no entanto, que micro ou macroetiquetas poderiam ser associadas tanto à informação principal quanto a informações secundárias em um mesmo texto.

Outro problema encontrado foi que, mesmo nos casos em que o texto se referia indubitavelmente a uma informação diferente do tópico principal – e, portanto, a segmentos secundários cujos aspectos são indicados por EXTRA – os anotadores reconheceram segmentos embutidos que, em vez de referir-se ao tópico secundário correspondente, faziam alusão à informação principal do sumário. Houve, assim, um impasse: apesar de o grande segmento indicar aspectos EXTRA, um ou mais segmentos subordinados a ele deveria ser anotado com etiqueta sem esse sufixo.

Os analistas dos sumários de 'Política' optariam por manter a associação explicitada textualmente e, neste caso, a algumas etiquetas seria acrescentado o sufixo, mas a outras não. Entretanto, foi consenso entre todos os anotadores do córpus que casos de subordinação dessa natureza deveriam seguir a regra genérica do grande segmento, ou seja, todos os aspectos vinculados a um segmento secundário deveriam carregar o sufixo EXTRA. Assim, mesmo um segmento referente ao tópico

principal, se embutido em um segmento secundário, passou a ser anotado com sufixo EXTRA. Do ponto de vista dos anotadores dessa categoria, essa decisão implica a perda semântica da referenciação adequada entre as informações veiculadas no texto. O sumário C40 abaixo ilustra esse caso: o agente do tópico principal (1ª. sentença) 'Renan Calheiros' é referido novamente no segmento secundário (2ª. sentença), o qual contempla um tópico (o que a revista Veja disse) distinto do anterior (o que Renan Calheiros disse). Como ambos os segmentos têm agente coincidente, ele deveria ser anotado também como WHO_AGENT. No entanto, seguindo a convenção, ele foi marcado como WHO_AGENT_EXTRA, pois todo o segmento refere-se a um WHAT_EXTRA.

> [O presidente do Senado, Renan Calheiros, disse que a decisão do procurador geral da República, Antonio Fernando de Souza, de investigar as denúncias contra ele, atende a um pedido que fez há cerca de um mês.]**WHO_AGENT/WHAT/DECLARATION**
> [Renan referia-se à última reportagem da revista Veja que afirma que Renan é dono oculto de duas emissoras de rádio de Alagoas.] **WHAT_EXTRA/WHO_AGENT_EXTRA/DECLARATION**

Além da distinção entre apectos referentes ao tópico principal e aspectos EXTRA, surgiram também dúvidas em relação à anotação de aspectos específicos, um dos problemas que se tornou generalizado para todas as categorias. Esses casos foram revistos e decidiu-se em conjunto por um padrão final. O uso da etiqueta SITUATION constitui um exemplo dessa natureza: mesmo que reconhecida a existência de uma situação em vários segmentos da categoria 'Política', esse aspecto só foi anotado quando os segmentos situacionais estavam claramente marcados como no primeiro segmento abaixo, do sumário C17. No segundo exemplo, do sumário C9, não houve consenso para a anotação de SITUATION. Considerou-se que 'encontro' marca claramente uma situação, mas 'operação', não. Logo, não se anotou esse aspecto no segundo exemplo.

> [Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.]**WHEN/SITUATION/WHO_AGENT/WHAT**

> [Em uma operação chamada "Operação Dominó", a Polícia Federal prendeu na manhã desta sexta-feira 23 pessoas suspeitas de envolvimento em esquema da Assembléia Legislativa do Estado de Rondônia para desvio de recursos públicos e influência indevida sobre o Poder Judiciário, Ministério Público, Tribunal de Contas e Poder Executivo do Estado.]**WHO_AGENT/WHAT/WHEN/WHO_AFFECTED/WHY**

A diversidade de opiniões sobre considerar micro ou macroetiquetas (caso iii) também ocorreu em 'Política'. Foi o caso de GOAL, que, apesar de classificada como macroetiqueta

(Quadro 3), em política ela indica segmentos bem marcados no nível microestrutural, como mostra o segmento do sumário C43.

> [A nova perícia da Polícia Federal permitirá comprovar a veracidade de documentos apresentados para justificar esses ganhos.]**PREDICTION/WHAT_EXTRA/ WHO_AGENT_EXTRA/GOAL_EXTRA**

É relevante notar que, em todos os sumários de política (4 ao todo), GOAL reflete um microaspecto, indicado por segmentos textuais menores que uma sentença.

Já a etiqueta WHO_AFFECTED, que na TAC 2010, e mesmo nas outras categorias do CSTNews, aparece com conotação expressivamente negativa sobre pacientes de eventos (em geral, vítimas, como no caso de 'Desastres Naturais'), em política ela também é associada a casos positivos. No sumário C9, por exemplo, ela indica beneficiários, opostamente a vítimas:

> [As fraudes envolviam, também, o superfaturamento de contratos com fornecedores da Assembléia, além de uma verdadeira 'farra' com passagens aéreas distribuídas fartamente entre os parlamentares, seus familiares e amigos.]**WHO_AFFECTED_EXTRA/WHAT_EXTRA**

Identificou-se também a existência de uma clara relação entre a macroetiqueta CONSEQUENCE e a microetiqueta WHY, o que levou a questionar se seria verdade que CONSEQUENCE sempre admitiria um WHY, ou que WHY sempre levaria a uma CONSEQUENCE. Em outras palavras, as implicações lógicas WHY ⇒ CONSEQUENCE e CONSEQUENCE ⇒ WHY poderiam se sustentar na categoria 'Política'?

Do ponto de vista semântico e mesmo de algumas teorias semânticas (p.ex., Fillmore, etc.) reconhece-se que a relação entre WHY e CONSEQUENCE assinala uma relação do tipo *causa-efeito*. A questão se coloca, portanto, na interação entre os níveis *informativo* (de *causa-efeito*) e *intencional* (de CONSEQUENCE ou, até, JUSTIFICATION ou EXPLANATION, como trata a RST, por exemplo). Como a marcação do córpus não adota explicitamente um modelo de discurso, procurou-se preservar sua independência de considerações entre esses níveis discursivos. Por essa razão, nem sempre aspectos que poderiam estar interrelacionados irão aparecer juntos ou serão considerados interdependentes. É importante frisar, porém, que, na anotação humana, bastante dependente de noções subjetivas, reconheceu-se que a apreensão e determinação dos aspectos semânticos se dá pelo processamento dos diversos níveis comunicativos sugeridos pelos textos.

No que tange à etiqueta DECLARATION, entendeu-se que qualquer declaração já admitiria o declarante e a coisa declarada, portanto, marcar também as microetiquetas WHO_AGENT e WHAT, para segmentos principais ou secundários envolvendo declarações, seria redundante. No entanto, respeitou-se a decisão consensual de manter, para toda DECLARATION, seu WHO_AGENT correspondente (caso iv).

Um problema adicional desse atrelamento da microetiqueta WHO_AGENT à macroetiqueta DECLARATION leva, segundo a ótica dos anotadores, a uma maior fragilidade de anotação para a categoria 'Política': muitos segmentos indicativos de

declarações trazem outros agentes diferentes do declarante, os quais também devem ser marcados. Como não se repetem etiquetas, esses casos perdem especificidade. Em outras palavras, não se sabe se a ocorrência de uma etiqueta WHO_AGENT refere-se ao declarante ou a outros agentes de outros eventos referidos na mesma declaração. O texto C17 exemplifica essa fragilidade, em que Geraldo Alckmin é o declarante, mas Lula é outro agente, casos indistinguíveis, pela única etiqueta WHO_AGENT_EXTRA:

> [Enquanto isso, adversário tucano Geraldo Alckmin disse que Lula deu as costas para o povo brasileiro, a Justiça e os bons costumes.]**WHAT_EXTRA/ WHO_AGENT_EXTRA/DECLARATION**

Por um lado, se a ocorrência de DECLARATION já abarcasse o declarante, sem necessidade de anotá-lo com WHO_AGENT, seria possível considerar que a ocorrência da etiqueta WHO_AGENT_EXTRA, nesse sumário, indicaria inequivocamente o agente 'Lula', e não qualquer declarante mencionado no segmento DECLARATION. Por outro lado, a necessidade de distinguir agentes múltiplos nem sempre ocorre. É o que mostra outro trecho do mesmo sumário, em que 'Lula' é, ao mesmo tempo, declarante e agente da ação de vencer a eleição. Nesse caso, é necessário pressupor que a única etiqueta WHO_AGENT pode referenciar mais de uma ocorrência de agentes em um mesmo segmento textual.

> [Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.]**WHEN/SITUATION/WHO_AGENT/DECLARATION**

Para efeito de modelagem artificial de escolhas para a SAM, a subsunção de WHO_AGENT a DECLARATION parece ser de fácil resolução. Bastaria procurar o declarante a partir de verbos que assinalam declarações (*disse que, afirmou que,* etc.). No entanto, se houver a suposição de que uma única etiqueta WHO_AGENT pode indicar várias vezes o mesmo aspecto em um mesmo segmento, seria preciso processos mais profundos (de *parsing*) para identificar os respectivos agentes.

Quanto à associação da etiqueta WHAT com DECLARATION (caso v), na categoria 'Política' essa correlação não se mostra obrigatória, pois, por indicarem macroaspectos, eles aparecem isolados. Como consequência, quando WHAT aparece com DECLARATION em um mesmo texto, é porque de fato ele indica informação diferente de uma declaração.

Uma última consideração acerca da categoria 'Política' diz respeito ao conhecimento específico de domínio que é necessário para a identificação (e anotação) dos aspectos. Muitas vezes foi necessário recorrer aos textos-fonte dos sumários e, até, a outros textos disponíveis na web, para confirmar a compreensão de alguns assuntos. Isso porque, como o sumário já é um resumo do conteúdo de vários textos-fonte, as informações necessárias para a compreensão não necessariamente estão contidas nele. Além disso, a interpretação dos textos políticos exige certo conhecimento de mundo, que está relacionado ao sistema eleitoral, aos processos jurídicos, aos procedimentos de votação das leis, à hierarquia de cargos, dentre outros. Se considerada a tarefa de análise semântica minuciosa a realizar, para a

anotação de aspectos de qualquer categoria, esse problema poderá se repetir, o que o torna interessante e, possivelmente, generalizável, para a modelagem de PLN.

6. POSSÍVEIS PADRÕES DE ORGANIZAÇÃO DE CONTEÚDO PARA A SA

Os padrões organizacionais de conteúdo, que se aplicam tanto à SA monodocumento quanto à SA multidocumento, foram derivados do córpus já consensual, que está disponível na página do projeto SUCINTO. Como já apontado antes, esses padrões são dependentes dos domínios de conhecimento variados do córpus, aqui referenciados pelas categorias. Por essa razão, primeiramente se apresentam os padrões por categoria, depois sugerem-se as generalizações que, por ora, se tornam evidentes a partir dos aspectos do córpus.

6.1. Cotidiano

Nessa categoria, observou-se que não há um grupo de aspectos comum para todos os sumários, provalmente devido à variedade de temas que eles envolvem. No entanto, há um grupo que pode caracterizar a maioria deles, que apresenta uma ordenação parcial clara, porém não necessariamente em sequência direta, ou seja, um imediatamente após o outro. No Quadro 8, apresentam-se, respectivamente, os aspectos mais frequentes para a maioria dos sumários, aqueles que ocorrem no primeiro r oparafo do sumário, e sua ordenação parcial ("X<Y" indica que o aspecto X ocorre antes do aspecto Y no texto).

Quadro 8. Análise dos aspectos nos sumários de 'Cotidiano'.

Em comum	WHAT, WHERE, WHEN, WHO_AGENT, DECLARATION
No 1º parágrafo	WHAT, WHERE, WHEN
Ordenação parcial	WHAT<WHERE<WHEN

No geral, a análise dos 14 sumários multidocumento da categoria 'Cotidiano' indica que:

- WHAT acontece em 13 dos 14 sumários.
- Em todos eles, COMPARISON, GOAL e PREDICTION têm uma única ocorrência.
- COMMENT aparece só em dois sumários.
- Apesar de HISTORY aparecer dez vezes, sete ocorrem em um único sumário. Portanto, não se pode concluir que sua ocorrência é representativa dessa categoria.
- DECLARATION aparece em dez sumários, o que certamente eleva a frequência de WHO_AGENT, já que toda declaração aparece vinculada a um agente (caso iv, Seção 5).
- Entre os dez sumários marcados com DECLARATION, três relatam discursos políticos. Esses têm, respectivamente, 5, 13 e 6 sentenças. Nos dois primeiros há 5 etiquetas DECLARATION e no terceiro, só 3. Os outros sumários não apresentam padrão quanto ao uso desse aspecto.
- WHO ocorre em todos os sumários, sempre em uma de suas duas formas: WHO_AFFECTED ou WHO_AGENT.

- CONSEQUENCE ocorre em seis sumários, sendo que quatro deles descrevem acidentes, tempestades e ataques criminosos.
- SCORE nunca aparece nessa categoria.

6.2. Esporte

Nessa categoria também se observa que alguns aspectos são mais comuns que outros e que há uma ordenação parcial entre eles que também não obedece uma sequência direta nos sumários. Porém, isso não ocorre para todos os casos. Por essa razão, apresentam-se separadamente no Quadro 9 os dados comuns a todos os sumários analisados e os dados comuns à maioria deles. Nota-se que os aspectos não são repetidos nessa segunda seção: claramente, os que se apresentam para todos os sumários também ocorrem na maioria![6] A marca "---" indica que não se encontrou um padrão nos sumários analisados.

Distinguem-se duas subcategorias de sumários aqui: a de *eventos esportivos*, que é efetivamente representativa das evidências textuais sobre esportes (7 sumários), pois envolve notícias sobre natação, vôlei, futebol e atletismo, e a de *outros*, que, apesar de enquadrados nessa categoria, não relatam eventos tipicamente relacionados a esportes (3 sumários) – um desses discorre sobre a saúde do ex-jogador Maradona, por exemplo. Nota-se que aspectos comuns a todos os sumários ou só a parte (maioria) deles sempre ocorrem no 1º. parágrafo, para os sumários sobre eventos esportivos. Casos múltiplos de ordenação parcial dos aspectos sobre esportes são possíveis, como indicaram os sumários. Eles são indicados pela variedade de sequências de aspectos, ressaltada pela lista itemizada, na seção referente à maioria dos sumários nesse quadro[7].

Quadro 9. Análise dos aspectos nos sumários de 'Esporte'.

	Eventos esportivos (7 sumários)	**Outros** (3 sumários)
Para todos os sumários		
Em comum	WHO_AGENT , WHAT	WHAT
No 1º parágrafo	WHO_AGENT , WHAT	WHAT
Ordenação parcial	WHO_AGENT<WHAT	---
Para a maioria dos sumários		
Em comum	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE	WHO_AFFECTED, WHAT, HOW, PREDICTION, DECLARATION
No 1º parágrafo	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE	WHO_AFFECTED, WHAT
Ordenação parcial	• WHO_AGENT<WHAT • WHO_AGENT, WHAT<SCORE • WHO_AGENT, WHAT<CONSEQUENCE • WHO_AGENT, WHAT<SITUATION • WHO_AGENT, WHAT<WHERE • WHO_AGENT, WHAT, SCORE<CONSEQUENCE	• WHO_AFFECTED<WHAT • WHAT<HOW

[6] Esse padrão de exibição é repetido para a descrição dos aspectos das demais categorias.

[7] Padrão também repetido para as demais categorias. Vírgulas, na indicação de ordens parciais, indicam que a ordem dos aspectos nos textos não é distinguível, embora eles sejam coocorrentes.

Os sumários da categoria ‘Esporte’ sempre apresentam eventos envolvendo WHO_AGENT e WHAT em primeiros parágrafos e eles sempre aparecem nessa ordem. Para a maioria dos sumários (segunda parte do quadro), nota-se que, quando ambos coocorrem com outros, eles antecedem aspectos específicos, a saber: os macroaspectos CONSEQUENCE e SITUATION e os microaspectos SCORE e WHERE. Já quando WHO_AGENT, WHAT e SCORE coocorrem, eles antecedem CONSEQUENCE.

Essa análise leva às seguintes conclusões preliminares, para a categoria ‘Esporte’:

- O único aspecto que ocorre em todos os sumários é WHAT.
- WHAT nunca se repete em um mesmo sumário.
- WHO ocorre em todos os sumários. Mais especificamente, as 5 ocorrências de WHO_AFFECTED foram identificadas nos sumários da segunda classe (*outros*).
- SCORE e CONSEQUENCE aparecem em 6 dos 7 sumários sobre notícias esportivas.
- COMMENT ocorre em 5 sumários distintos, dentre os 7 de notícias esportivas.
- WHY ocorre em um único sumário – no caso, um dos 3 que não são sobre esportes (subcategoria *outros*).
- GOAL ocorre apenas uma vez na categoria ‘Esporte’.
- HOW é muito frequente nos sumários que versam sobre futebol.
- DECLARATION aparece apenas em 2 dos 3 sumários da subcategoria *outros.*
- Apenas um sumário não apresenta material secundário, isto é, material que contém segmentos rotulados por EXTRA.
- COUNTERMEASURES, aspecto típico de outras categorias, não ocorre na categoria ‘Esporte’.

6.3. Mundo

Além do Quadro 10 exibir a distribuição de padrões de modo análogo ao anterior, ele ainda diferencia os vários assuntos (ou domínios de conhecimento) encontrados nos sumários de ‘Mundo’ em cada uma de suas colunas. São 2 sumários sobre acidentes, 4 sobre ataques, 3 sobre decisões legais e políticas e 5 sobre desastres naturais.

Quadro 10. Análise dos aspectos nos sumários de 'Mundo'.

	Acidentes	Ataques	Decisões Legais e Políticas	Desastres Naturais
Para todos os sumários				
Em comum	WHAT, WHERE, WHO_AFFECTED, WHY	WHAT, WHERE, WHO_AFFECTED, WHEN	WHAT, WHO_AGENT	WHAT, WHERE, WHO_AFFECTED, CONSEQUENCE
No 1º parágrafo	WHAT, WHERE, WHO_AFFECTED	WHAT, WHERE, WHO_AFFECTED, WHEN	WHAT, WHO_AGENT	WHAT, WHERE
Ordenação parcial	• WHAT<WHERE • WHO_AFFECTED, WHAT, WHERE<WHY	WHAT<WHERE	---	• WHAT<WHERE • CONSEQUENCE
Para a maioria dos sumários				
Em comum	---	WHO_AGENT, WHY, HISTORY	HISTORY,WHO_AFFECTED,WHERE, DECLARATION, GOAL	COUNTERMEASURES, PREDICTION
No 1º parágrafo	---	---	WHO_AFFECTED, WHERE	WHO_AFFECTED, CONSEQUENCE
Ordenação parcial	---	• WHO_AFFECTED<WHEN, WHERE • WHO_AGENT<HISTORY • WHEN<WHY • WHAT<WHO_AFFECTED	• WHO_AFFECTED, WHAT, WHO_AGENT< HISTORY • WHAT, WHO_AGENT<GOAL • WHO_AGENT, WHAT, WHERE< DECLARATION • WHO_AGENT<WHAT	• WHAT, WHERE< WHO_AFFECTED • WHAT, WHERE, CONSEQUENCE, WHO_AFFECTED< PREDICTION

Observa-se, por exemplo, que, para todos os sumários sobre *desastres naturais*, o aspecto WHAT sempre ocorre antes de WHERE, mas quando WHAT e WHERE coocorrem, eles sempre antecedem CONSEQUENCE. Sobre esse mesmo tema, a maioria dos sumários ressalta COUNTERMEASURES e PREDICTION, sendo que este não é um aspecto comum às outras categorias.

É interessante notar que alguns dos aspectos só aparecem em poucas subcategorias, caso de CONSEQUENCE, que só ocorre em 'Desastres naturais'.

A síntese da análise para a categoria 'Mundo', quer genérica, quer para cada tema específico, é a seguinte:

- O único aspecto que ocorre em todos os sumários é o WHAT.
- Não existe sumário que não contenha alguma forma de WHO (ou WHO_AFFECTED ou WHO_AGENT).
- PREDICTION aparece pelo menos em um sumário de cada tema, exceto em *Acidentes*.
- GOAL é característico de *Decisões Legais e Políticas*, pois só ocorre em sumários sobre esse tema.
- CONSEQUENCE ocorre apenas ligado ao tema '*Acidentes* e *Desastres Naturais'*.
- WHAT possui um parágrafo dedicado a ele: sempre o primeiro.
- Raramente WHAT se repete (1 vez em 14), talvez mesmo pelo motivo anterior.
- Dos 14 sumários, 13 apresentam material secundário, cujos aspectos são rotulados com o sufixo EXTRA.

6.4. Política

A categoria 'Política' abarca três diferentes estruturas textuais: a) textos sobre pesquisas eleitorais, em que se divulgam intenções de voto, usualmente por meio da

comparação entre os candidatos (cf. coleção C2) – e esta é a razão da alta frequência da etiqueta COMPARISON; b) textos que mencionam trocas de agressões verbais entre políticos ou candidatos (cf. coleção C17), os quais possuem uma estrutura retórica diferente daquela geral em textos jornalísticos, devido às trocas de turnos e, como consequência, à grande frequência da etiqueta DECLARATION; c) textos que noticiam fatos ou eventos mais gerais relacionados aos personagens da vida política (cf. coleções C16, C20 ou C42), que de fato apresentam a organização textual comum aos textos jornalísticos. Essas estruturas, por trazerem suas próprias cargas semânticas, indicam padrões distintos de ocorrência e ordenação de aspectos.

Há só um sumário multidocumento que apresenta a estrutura (a), o C2. Sua primeira sentença incorpora os aspectos DECLARATION, WHEN, WHAT, WHO_AFFECTED e COMPARISON. Em quase todas as suas outras sentenças seguem as etiquetas WHAT_EXTRA, WHO_AFFECTED_EXTRA e COMPARISON. Essa organização pode ser compreendida em seu contexto de pesquisa eleitoral, em que o jornalista apresenta primeiro o candidato com a maior porcentagem das intenções de voto, seguido de seus adversários, claramente lançando mão de uma comparação de intenção de votos (indicada pelo aspecto COMPARISON). As probabilidades de intenções é que levam aos aspectos WHAT ou WHAT_EXTRA; os candidatos envolvidos, por sua vez, levam a WHO_AFFECTED ou WHO_AFFECTED_EXTRA.

Também há só um sumário (C17) que apresenta a estrutura (b), de troca de agressões verbais. Nesse caso, mencionam-se dois candidatos à presidência, que estão em embate político e permanecem acusando-se mutuamente de forma verbal. Essa agressão verbal é relatada pelos autores dos textos jornalísticos correspondentes ora em transcrição literal do discurso dos envolvidos, ora na forma reescrita, como mostram as três primeiras sentenças do referido sumário. Se uma vaga ideia (em inglês, *gist*) sobre o tópico principal desse sumário fosse inferida (e expressa) pelos leitores, ela poderia ser denominada "O embate político entre candidatos", muito embora essa ideia de agressão verbal não esteja explicitada na superfície textual, como claramente pode se ver:

> ["Eu não moverei uma palha contra eles [oposição] porque vocês moverão um paiol inteiro", afirmou o presidente Luiz Inácio Lula da Silva, candidato à reeleição pelo PT, sobre os ataques de seus adversários.]**WHO_AGENT/DECLARATION/WHAT**
> [Enquanto isso, adversário tucano Geraldo Alckmin disse que Lula deu as costas para o povo brasileiro, a Justiça e os bons costumes.] **WHAT_EXTRA/WHO_AGENT_EXTRA/DECLARATION**
> ["[Lula] trabalhou do lado do Waldomiro (Diniz), do mensalão, dos sanguessugas, do valerioduto, desses escândalos todos. Isso é que é grave", reagiu o candidato à declaração de Lula de que o PSDB abandonou os pobres e que ele é vítima de preconceito das elites do País porque fez o contrário.]**WHO_AGENT_EXTRA/DECLARATION/WHAT_EXTRA**

Todas as sentenças desse sumário (ao todo, seis) trazem ainda os aspectos WHO_AGENT ou WHO_AGENT_EXTRA, e DECLARATION só não ocorre em uma delas, o que reforça as evidências do embate político pela menção aos agentes das agressões mútuas.

Os textos mais gerais da categoria 'Política', que enfatizam a estrutura (c), são 8 (do total de 10): são os sumários humanos das coleções C9, C16, C20, C40, C42, C43, C44 e C50. É desses casos mais representativos que se extraem os padrões apresentados no Quadro 11. Os aspectos provenientes dos outros dois sumários não são incluídos no quadro por não serem representativos.

Quadro 11. Análise dos aspectos nos sumários de 'Política'.

Para todos os sumários	
Em comum	WHAT, WHO_AGENT_EXTRA, WHAT_EXTRA
No 1º parágrafo	WHAT
Ordenação parcial	---
Para a maioria dos sumários	
Em comum	WHO_AGENT, WHAT, WHEN, WHY
No 1º parágrafo	WHO_AGENT, WHAT, WHO_AFFECTED, WHEN
Ordenação parcial	• WHO_AGENT, WHAT, WHEN<PREDICTION • WHO_AGENT, WHAT, WHEN<COUNTERMEASURES • WHO_AGENT, WHAT, WHEN<DECLARATION • WHO_AGENT, WHAT, WHEN<HISTORY • PREDICTION<CONSEQUENCE

Os oito sumários apresentam segmentos WHAT, que assinalam o tópico principal sobre o qual se versa. Segue ainda que a maioria deles apresenta WHO_AGENT, WHAT, WHO_AFFECTED e WHEN no primeiro parágrafo, indicando, como parte do tópico principal, menções a *quem fez, o quê fez, a quem fez* e *quando o fez*. E, claro, já que WHAT ocorre em primeiros parágrafos – e raramente aparece mais de uma vez, reforçam-se as evidências de que é a primeira a sentença *lead* no gênero jornalístico. As sentenças que seguem a *lead* apresentam, com frequência ainda alta, WHAT_EXTRA, DECLARATION e COMPARISON. Todos os sumários também têm em comum as etiquetas WHAT_EXTRA e WHO_AGENT_EXTRA em segmentos indicativos de eventos secundários. Pela ordenação parcial das etiquetas para a maioria dos sumários, depreende-se ainda que, em geral, as microetiquetas ocorrem antes das macro (PREDICTION, COUNTERMEASURE, DECLARATION e HISTORY).

A distribuição dos aspectos na categoria 'Política' leva às seguintes conclusões:

- O único aspecto que ocorre em todos os sumários é o WHAT, em ambas as suas formas, principal ou secundária (com sufixo EXTRA).
- O aspecto WHO, ocorre em todos os sumários, ora como WHO_AFFECTED, ora como WHO_AGENT, e também em sua forma EXTRA.
- O aspecto COMPARISON ocorre com frequência muito alta no texto que indica pesquisa eleitoral, mas só aparece nesse sumário (C2).
- O aspecto DECLARATION aparece pelo menos uma vez na maioria dos sumários.
- Os aspectos HOW e CONSEQUENCE ocorrem 3 vezes cada, mas somente em um sumário.
- O aspecto WHAT sempre aparece na sentença *lead* (primeira).
- Raramente WHAT se repete (1 vez em 10).
- Os 10 sumários apresentam material secundário, caso de aspectos rotulados pela etiqueta EXTRA.

7. DESDOBRAMENTOS PARA A SA MULTIDOCUMENTO

Em geral, para as quatro categorias relatadas neste relatório técnico, observa-se que:

- Exceto para um único sumário ('Cotidiano'), WHAT ocorre em todos os sumários do córpus CSTNews, sempre em primeiros parágrafos.
- O aspecto WHO em ambas as suas formas – WHO_AGENT e WHO_AFFECTED – também aparece em todos eles.
- Todo sumário apresenta detalhes adicionais, isto é, informações que não são diretamente relativas àquelas apresentadas como tópico principal. Neste caso, são denotadas pelo sufixo EXTRA.

A frequência dos aspectos varia significativamente entre as categorias, ou seja, alguns ocorrem muito em algumas delas, outros nunca ocorrem ou ocorrem poucas vezes em outras. É o caso dos aspectos CONSEQUENCE, DECLARATION, HOW, SCORE, COMMENT, WHY e COUNTERMEASURES. Por exemplo, HOW, SCORE e COMMENT são muito frequentes em 'Esporte', mas são raros (ou totalmente ausentes) nas demais categorias. Há também alguns aspectos que se manifestam sempre muito pouco, como PREDICTION, GOAL e COMPARISON.

Também alguma ordem parcial é observada no córpus todo de sumários multidocumento. Frequentemente WHAT precede WHERE e WHO_AGENT precede HISTORY, quando esses aspectos aparecem em um mesmo sumário.

Este relatório, por fim, deixa evidentes as contribuições práticas e teóricas para a SAM de textos em português. Ele evidencia um estudo aprofundado sobre a natureza das escolhas manuais (humanas), para a produção dos sumários multidocumento manuais, de notícias jornalísticas. Especialmente, traz evidências sobre as escolhas dos aspectos textuais para compô-los. Assim, pode-se propor uma relação de aspectos para a SAM de textos jornalísticos a partir da análise aqui apresentada, mas ressalta-se que ela também pode se aplicar ao uso genérico, embora esse uso não tenha sido explorado ainda.

Os padrões extraídos da tarefa de anotação sistemática de aspectos textuais em sumários do córpus CSTNews demonstram o valor de se considerar as diversas categorias (e subcategorias, quando aplicáveis), com suas ocorrências particulares ou genéricas em função das posições e da ordenação parcial dos aspectos. Não se tem notícia de que haja padrões desse tipo para algum córpus, em análise similar à aqui relatada. Neste caso, o elenco de padrões consiste na primeira tentativa de se minerarem estruturas para a geração automática de sumários e, ao final, enriquecer os métodos de SA, em geral.

Embora não se tenha obtido uma estimativa da concordância entre os anotadores nessa tarefa de anotação de aspectos, e tampouco se saiba se a concordância será uma medida relevante para a definição e o uso computacional, a busca de consenso, seguida da clara convergência dos anotadores, permite assegurar que as definições reproduzem bastante fielmente os vários tipos de informação que acabam assinalados tanto pelas macro, quanto pelas microetiquetas.

Também é notável a correspondência de alguns aspectos com papeis semânticos, como se propôs na FrameNet (Baker et al., 1998). Por exemplo, WHEN e WHERE claramente correspondem aos aspectos usadas quando se determinam complementos verbais. Outros aspectos têm uma natureza diferente, como WHAT, PREDICTION e COMMENT, juntamente com suas versões EXTRA. Esses aspectos estão

mais relacionados ao contexto global dos sumários e seus fatos e eventos narrados, do que aos componentes sentenciais em si. Essa distinção é que levou à classificação dos aspectos em micro e macroetiquetas: as microetiquetas se manifestam no contexto sentencial, como se manifestam os papeis semânticos; as macro requerirão o contexto de vizinhança para sua definição.

Também é interessante que alguns aspectos possam se desdobrar ou em micro, ou em macroetiquetas. Por exemplo, não é dificil imaginar que GOAL pode ser parte de uma estrutura argumental de um verbo (e, assim, poderia ser classificado como microaspecto) ou pode ser usado para categorizar uma sentença inteira, se se referir à sua função no sumário (caso em que se torna macroaspecto).

Os macroaspectos, ainda, podem se enquadrar em algum modelo de discurso, como a *Rhetorical Structure Theory*, ou RST (Mann e Thompson, 1987), em que as unidades discursivas referentes a GOAL poderiam estar em uma relação RST PURPOSE com outras sentenças.

Os próprios padrões relatados na Seção 6, afinal, podem ser usados para se derivar os *templates* para geração ou interpretação de estruturas representativas de sumários de textos jornalísticos. Assim, os sistemas de sumarização automática poderão ser treinados ou desenvolvidos tanto com base nos aspectos, quanto com base nos *templates*, resultando em sistemas mais linguisticamente motivados e, assim, produtores de sumários mais informativos e mais direcionados ao tema ou categoria textual em foco. Um classificador de aspectos pode ainda ser desenvolvido com base em sumários anotados com aspectos, a fim de se automatizar por completo o processo. É interessante notar que um classificador desse tipo não precisaria necessariamente identificar todos os aspectos em um texto, mas somente aqueles mais úteis para a produção dos sumários, conforme indicações dos padrões ou *templates* adotados.

REFERÊNCIAS

Afantenos, S.D.; Karkaletsis, V.; Stamatopoulos, P.; Halatsis, C. (2008). Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. *Journal of Intelligent Information Systems*, Vol. 30, N. 3, pp. 183-226

Aleixo, P.; Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-documentStructureTheory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 326. 12p.

Baker, C.F.; Fillmore, C.J.; Lowe, J.B. (1998).The Berkeley FrameNet project. In the *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 86-90.

Barrera, A.; Verma, R.M.; Vicent, R. (2011).SemQuest: University of Houston's Semantics-based Question Answering System. In the *Proceedings of the4th Text Analysis Conference* (TAC 2011).

Boguraev, B.; Kennedy, C. (1997). Salience-Based Content Characterisation of Text Documents. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop*, pp. 2-9. ACL/EACL'97 Joint Conference. Madrid, Spain.

Camargo, R.T.; Maziero, E.G.; Pardo T.A.S. (2012). Corpus analysis of aspects in multi-document summaries - the case of news texts from 'world' section. In the *OnlineProceedings of the 11th Corpus Linguistics Symposium* (ELC 2012).

Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di-Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011a). 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.

Cardoso, P.C.F.; Pardo, T.A.S.; Nunes, M.G.V. (2011b). Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 59-74. October 26, Cuiabá/MT, Brazil.

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

Edmundson, H.P. (1969). New Methods in Automatic Extracting. *Journal of the ACM*, 16, pp. 264-285.

Endres-Niggemeyer, B. (1998). *Summarization Information*. Berlin: Springer.

Fillmore, C. (1968). The Case for Case. In E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*. Holt, Rinehart & Winston. New York.

Genest, P.; Lapalme, G.; Yousfi-Monod, M. (2009). HexTac: the Creation of a Manual Extractive Run. In the *Proceedings of the 4th Text Analysis Conference* (TAC 2011), November 14-15, Maryland, USA.

Jackendoff, R. (1983). *Semantics and Cognition*. MIT Press. Cambridge, MA.

Jorge, M.L.C.; Di Felippo, A.; Nóbrega, F.A.A.; Souza, J.W.C. (2012). Analysis of informational aspects in a corpus of manual multi-document summaries of 'sport' news. In the *Online Proceedings of the 11th Corpus Linguistics Symposium* (ELC 2012).

Jorge, M.L.C.; Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proc. of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82. Uppsala/Sweden.

Jorge, M.L.C.; Agostini, V.; Pardo, T.A.S. (2011). Multidocument Summarization Using Complex and Rich Features. In *Anais do VIII Encontro Nacional de Inteligência Artificial*. Natal-RN.

Li, P.; Wang, Y.; Gao, W.; Jiang, J. (2011). Generating Aspect-oriented Multi-Document Summarization with event-aspect model. In the *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 1137–1146.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.

Makino, T.; Takamura, H.; Okumura, M. (2011).Balanced Coverage of Aspects for Text Summarization. In the *Proceedings of the4th Text Analysis Conference* (TAC 2011), November 14-15, Maryland, USA.

Mani, I. (2001). *Automatic Summarization*. John Benjamin's Publishing Company. Amsterdam.

Mann, W.C.; Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Marcu, D. (1997a). From Discourse Structures to Text Summaries. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop*, pp. 82-88. ACL/EACL'97 Joint Conference. Madrid, Spain.

Marcu, D. (1997b). The Rhetorical Parsing of Natural Language Texts. In the *Proc. of the ACL/EACL'97 Joint Conference*, pp. 96-103. Madrid, Spain.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Nenkova, A.; Louis, A. (2008). Can you summarize this? Identifying correlates of input difficulty for multi-document summarization. In the *Proc. of the ACL: HLT,* pp. 825–833.

Nenkova, A.; McKeown, K. (2011). Automatic Summarization.*Foundations and Trends in Information Retrieval*, Vol. 5, N. 2 & 3.

Owczarzak, K.; Dang, H.T. (2011). Who wrote What Where: Analyzing the content of human and automatic summaries. *Proc. of the Workshop on Automatic Summarization for Different Genres, Media e Languages*, pp. 25–32. Portland, Oregon, June 23. Association for Computational Linguistics.

Pollock, J.J.; Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Compute Sciences* 15(4): 226-232.[8]

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

Rassi, A.P.; Rino, L.H.M.; Dias, M.S. (2012). Preliminary Aspects Distribution in Political Texts. In the *Online Proc. of the 11th Corpus Linguistics Symposium* (ELC 2012).

Steinberger, J.; Tanev, H.; Kabadjov, M.; Steinberger, R. (2010).JCR's Participation in the Guided Summarization Task at TAC 2010. In the *Proceedings of the 3rd Text Analysis Conference* (TAC 2010).

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2010). A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, Vol. 6, N. 4, pp. 1-20.

White, M.; Korelsky, T.; Cardie, C.; Ng, V.; Pierce, D.; Wagstaff, K. (2001). Multidocument summarization via information extraction. In the *Proceedings of the 1st International Conference on Human Language Technology Research*, pp. 1-7.

Zacarias, A.C.I.; Agostini, V.; Cardoso, P.C.F.; Seno, E.M.R. (2012). Análise de Aspectos de Sumários Multidocumento e sua Correlação com a Informatividade. In the *Online Proc. of the 11th Corpus Linguistics Symposium* (ELC 2012).

Zhang, Z.; Blair-Goldensohn, S.; Radev, D.R. (2002). Towards CST-Enhanced Summarization. In the *Proceedings of the 18thNational Conference on Artificial Intelligence*, pp. 439-445.

Zhou, L.; Ticrea, M.; Hovy, E. (2005). Multi-document Biography Summarization. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1-8.

[8] Obra reeditada em 1999.