UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-001/2012

ResDial – Descrição da Codificação (v.1.0)

Norton Trevisan Roman

Agosto - 2012

Série de Relatórios Técnicos

PPgSI-EACH-USP. Rua Arlindo Béttio, 1000 - Ermelino Matarazzo - 03828-000
São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

ResDial – Descrição da Codificação (v.1.0)

Norton Trevisan Roman[1]

[1]Escola de Artes, Ciências e Humanidades
Universidade de São Paulo (EACH-USP)
Av. Arlindo Béttio, 1000 – 03828-000 – Ermelino Matarazzo, São Paulo – SP – Brazil

norton@usp.br

Resumo. *Este documento descreve o conjunto de tags XML usadas na codificação do ResDial, um conjunto de corpora de resumos de diálogos produzidos por humanos.*

1. Introdução

Correspondendo a um conjunto de *corpora* de resumos de diálogos produzidos por humanos, o repositório ResDial conta, até o presente momento, com dois *corpora*. O primeiro deles produzido a partir de diálogos automaticamente gerados [Roman et al. 2006a], e o segundo a partir de diálogos encontrados em filmes [Roman et al. 2006b]. Dessa forma, muito embora os diálogos resumidos sejam artificiais, ambos *corpora* diferem em que, no primeiro, os diálogos foram gerados por um sistema computacional, enquanto que no segundo esses diálogos foram gerados por humanos.

Independentemente da fonte geradora, todos os diálogos usados como fonte para os resumos ilustram interações de compra e venda de mercadorias (ou seja, entre um vendedor e um possível comprador), estando os diálogos automaticamente gerados limitados à compra e venda de automóveis, enquanto que os demais tratam de relações mais gerais. Além disso, ambos os conjuntos de diálogos também compartilham a característica de ilustrarem situações em que ou um dos participantes do diálogo é impolido, ou ambos são impolidos, ou nenhum o é.

Para a produção dos resumos foram escolhidos quatro diálogos de cada fonte geradora, variando o grau de polidez de seus participantes. Assim, o primeiro deles ilustra uma situação em que o cliente foi rude, enquanto que o segundo apresenta uma situação em que o atendente foi rude, sendo os demais neutros (ou seja, apresentando situações em que nenhuma parte foi rude). Cada conjunto de diálogos foi então entregue a um grupo de voluntários, que deveriam resumi-los conforme um de três pontos de vista (vendedor, cliente, e um observador neutro).

De modo a reduzir o risco de viés nos dados, o conjunto de sumarizadores foi dividido aleatoriamente em uma das três supra-citadas categorias. Além disso, cada um desses conjuntos foi subdividido em outros dois subconjuntos: o de sumarizadores que não teriam limite no tamanho de seus resumos, e o de sumarizadores cujos resumos não poderiam ir além de 10% do número de palavras usadas no diálogo resumido. Cada voluntário deveria, então, resumir todos os quatro diálogos, conforme um determinado ponto de vista, não ultrapassando, se fosse o caso, 10% do número de palavras do diálogo fonte.

Uma vez produzidos os resumos, e com o intuito de padronizar a aplicação de esquemas de codificação aos *corpora* disponíveis, foi desenvolvido um conjunto de *tags* XML, a ser aplicado tanto nos *corpora* já existentes quanto em futuros *corpora* que venham a ser disponibilizados no repositório. Nesse relatório, esse conjunto é descrito em detalhes

(sendo resumido no Apêndice A), permitindo assim não somente a codificação de novos *corpora*, mas também o desenvolvimento de ferramentas para análise e marcação dos *corpora* existentes.

2. Organização dos Corpora

De modo a deixar claro o uso de *stand off annotation*, segundo a qual anotação e fonte são mantidas em arquivos separados, de alguma forma ligados entre si [Ide and Brew 2000], todo *corpus* dentro do ResDial é mantido em estado bruto: um arquivo texto, codificado com UTF-8, contendo apenas informação de identificação do documento (ver Seção 2.2.1 para detalhes). Além disso, as anotações dentro do ResDial buscam fazer uma separação clara entre as tarefas de anotação da estrutura do corpus, e as tarefas de anotação dos elementos dentro dessa estrutura.

Assim, *tags* ResDial podem ser divididas em *tags* para identificação do documento, *tags* para segmentação em unidades mínimas de anotação, e *tags* para classificação, segundo algum esquema de anotação definido pelo usuário, dessas unidades mínimas. O fato do usuário poder, em um primeiro momento, segmentar o corpus e, em seguida, aplicar algum esquema de anotação próprio possui a vantagem de permitir que ele possa usar as ferramentas disponíveis, tanto no repositório quanto externas, para cálculos estatísticos, como concordância entre anotadores por exemplo, obtendo valores específicos para a definição da unidade básica e a aplicação do esquema de anotação em si.

Essa prática, por sua vez, permite a redução da ambiguidade de resultados obtidos, em que não se sabe se um resultado negativo é devido, por exemplo, a falhas no esquema de anotação aplicado, ou na definição das unidades básicas. Além disso, a separação entre tarefas de estruturação e anotação abre a possibilidade para o desenvolvimento de ferramentas mais específicas à anotação em mãos e, consequentemente, mais simples de usar, quando comparadas a outras mais gerais, como as apresentadas em [Orăsan 2003, Ogren 2006, O'Donnell 2008, Verhagen 2010].

2.1. A Árvore de Diretórios

Dentro do ResDial, documentos são mantidos em arquivos separados, sendo cada *corpus* (em estado bruto) mantido em um diretório próprio (definido pelo usuário). A cada documento é associado um identificador único (dentro do *corpus* ao qual pertence), identificador este que deve ser também usado no nome do arquivo que contém o documento. Este arquivo, por sua vez, deve ter seu nome iniciado com "src", segundo o padrão "src_[identificador do arquivo]_[identificador do corpus]_[identificador do autor do documento].xml". Da mesma forma, também é possível registrar os textos-fonte para o corpus (como os diálogos usados pelos sumarizadores, por exemplo), como se fossem um segundo corpus. Sua forma de armazenamento é idêntica ao corpus de documentos em estado bruto.

A segmentação de um documento em estado bruto, por sua vez, é mantida em um diretório separado (definido pelo usuário), permitindo assim que diferentes segmentações possam ser aplicadas ao mesmo texto-fonte. Uma vez que cada esquema de segmentação possui um identificador, além de ter sido segmentado por um segmentador específico (que pode representar a opinião da maioria dentre um grupo de segmentadores, por exemplo), o nome de um arquivo contendo um documento segmentado deve seguir o padrão

"seg_[identificador do arquivo]_[identificador do esquema de segmentação]_[identificador do corpus-fonte]_ [identificador do documento-fonte no corpus]_[identificador do segmentador]. xml" (para arquivos resultantes de múltiplos segmentadores, a letra 'M' deve ser usada como identificador, conforme indicado na Seção 2.2.5).

Anotações em documentos são também mantidas em diretórios separados, definidos pelo usuário. Com relação a estas, o nome de cada arquivo usado para armazenar uma anotação deve seguir o padrão "ann_[identificador do arquivo]_[identificador do esquema de anotação utilizado]_[identificador do esquema de segmentação]_[identificador do documento anotado (no esquema de segmentação)]_[identificador do anotador].xml". Vale lembrar que anotações são aplicadas a um texto segmentado, e não em estado bruto.

Finalmente, uma vez que em qualquer uma das etapas descritas pode-se fazer uso de múltiplos anotadores ou participantes, cada um desses conjuntos admite um sub-diretório "participant", em que são armazenados arquivos com informação acerca dos anotadores que produziram o *corpus*. Nesse caso, cada anotador possui um arquivo separado, cujo nome segue o padrão "part_[identificador do participante]_[identificador do corpus ao qual contribuiu (fontes, de segmentação, ou de classificação)].xml".

Assim, e supondo que o diretório do *corpus* a ser anotado seja definido, pelo pesquisador, como "corpus2" (com "corpus1" contendo algum corpus adicional), que a segmentação de cada documento seja mantida em "segmentação1", e que o diretório que conterá as anotações seja definido como "esquema1", uma possível árvore de diretórios seria, por exemplo (vale notar que, nesse exemplo, tanto a segmentação quanto a classificação contaram com anotadores diferentes, conforme determinado por seus identificadores):

- corpus1
  - src_0001_c01_null.xml
  - src_0002_c01_null.xml
  - ...
- corpus2
  - src_0001_c02_null.xml
  - src_0002_c02_null.xml
  - ...
- segmentação1
  - seg_0001_S01_c02_0001_M.xml
  - seg_0002_S01_c02_0002_M.xml
  - ...
  - participant
    - part_an001_S01.xml
    - part_an002_S01.xml
- esquema1
  - ann_0001_E01_S01_0001_a01.xml
  - ann_0002_E01_S01_0001_a02.xml
  - ann_0003_E01_S01_0001_a03.xml
  - ann_0004_E01_S01_0002_a01.xml
  - ann_0005_E01_S01_0002_a02.xml
  - ann_0006_E01_S01_0002_a03.xml
  - ...

- participant
    - part_a01_E01.xml
    - part_a02_E01.xml
    - part_a03_E01.xml

em que "c02" é o identificador do corpus utilizado, "0001" o identificador do primeiro arquivo (documento) no *corpus*, "S01" o identificador do esquema de segmentação aplicado, em que foi usada a opinião de múltiplos anotadores (de fato, sua maioria), representada pelo identificador "M" (ver Seção 2.2.5), "E01" o identificador do esquema de anotação aplicado (e cujo resultado é armazenado em "esquema1"), e "a01" é o identificador do anotador que produziu o arquivo em questão. Um "null" no campo do autor dos arquivos em estado bruto indica que a informação sobre o autor do documento é irrelevante ou não pode ser determinada.

2.2. Especificação das Tags

A separação das tarefas executadas no *corpus* em tarefas de anotação da estrutura e de anotação dos elementos dentro dessa estrutura é feita por meio de três conjuntos de *tags*: *tags* de identificação, de segmentação e de classificação. Além disso, um quarto conjunto é usado para caracterização dos anotadores ou participantes de cada *corpus*. Qualquer que seja o conjunto, comentários podem ser incluídos segundo o padrão XML, como em "<!-- Isto é um comentário -->". No que segue, cada um desses conjuntos será visto em mais detalhes.

2.2.1. Tags para Identificação

Na raiz de cada documento em estado bruto, dentro do ResDial, está um tag <plainDocument> e seu correspondente </plainDocument> (vindo ao final do arquivo). Nesse estado, um documento deve conter, pelo menos, *tags* <info type="tipo" value="valor">, em que "tipo" é um identificador do tipo de informação sendo apresentada (definida pelo criador do *corpus*), enquanto que "valor" é a informação em si, além das *tags* <text> e </text>, que devem conter o texto do documento em estado bruto.

Como exemplo, considere os *corpora* hoje presentes no ResDial, descritos na Seção 1 deste documento. Dentro desses *corpora*, são relevantes para o documento seu identificador e o de seu diálogo fonte, bem como o identificador do *corpus* contendo esse diálogo-fonte, além do identificador do *corpus* que contém o documento em questão, resultando no arquivo ilustrado na Figura 1.

É importante ressaltar que a mesma codificação usada para o documento em estado bruto pode ser usada para qualquer outro documento (também em estado bruto). Assim, nesse exemplo, tanto o diálogo fonte, quanto seus resumos seriam postos em arquivos com extensão "src", codificados conforme o exemplo acima (ainda que estando em diretórios diferentes).

2.2.2. Tags para Segmentação

Antes que qualquer anotação possa ser aplicada ao documento em estado bruto, há que se definir a que porção do texto ela será aplicada, ou seja, há que se definir sua unidade

```
<?xml version="1.0" encoding="UTF-8"?>
<plainDocument>
    <info type="id" value="001">
    <info type="corpus" value="c01">
    <info type="source" value="d1">
    <info type="source-corpus" value="f01">
    <text>
        Texto do documento...
    </text>
</plainDocument>
```

Figura 1. Codificação de um documento em estado bruto.

básica de anotação [van der Vliet et al. 2011, Rodrigues et al. 2012]. Essas unidades, por sua vez, nem sempre podem ser definidas de maneira exclusiva, sendo bastante comum a superposição de unidades, em que um mesmo pedaço de texto pertence a duas unidades independentes, e o embutimento de unidades, ou seja, unidades definidas dentro de outras unidades, independentes de sua unidade anfitriã.

Além desses fatores, é desejável que as unidades básicas também permitam [Reidsma et al. 2005, Rodrigues et al. 2012]:

- Entrelaçamento, em que dois ou mais segmentos, embora independentes, tenham partes entrelaçadas; e
- Descontinuidade, em que duas porções de texto não contínuas pertencem à mesma unidade.

Seguindo essa linha, dentro do ResDial cada documento segmentado é mantido em um arquivo separado. Este documento, por sua vez, deve estar encapsulado por *tags* <document> e </document>. Também aqui são possíveis *tags* do tipo <info>, nos moldes definidos para o documento bruto (Seção 2.2.1), com a finalidade de registrar qualquer informação que se julgue necessária sobre o documento segmentado (normalmente, um identificador para o arquivo com a segmentação, para o texto em estado bruto, seu *corpus* de origem, para o esquema usado para segmentação, bem como o gerador desse arquivo, dentre outros possíveis).

Cada unidade definida é encapsulada por *tags* <unit> e </unit>, sendo que cada <unit> e correspondente </unit> devem possuir necessariamente um campo para identificação, tornando-se <unit id="identificador"> e </unit id="identificador">. Com isso, é possível definir unidades embutidas, como as ilustradas na Figura 2, de modo a que se saiba a qual <unit> um determinado </unit> pertence. Além disso, o campo de identificação pode ser manipulado para se permitir tanto unidades descontínuas quanto entrelaçadas. Para isso, basta que duas unidades possuam o mesmo identificador para que sejam interpretadas como sendo uma única unidade.

O embutimento de unidades, por sua vez, pode acontecer tanto entre unidades de alguma forma relacionadas (e a interpretação disso depende unica e exclusivamente do esquema adotado para segmentação), quanto totalmente independentes. Para o primeiro caso, basta que se use um *tag* <unit> dentro de outro. Já para o segundo, a independência é deixada clara adicionando-se a palavra "ind" à *tag*, criando então tags <unit ind id="identificador"> e </unit ind id="identificador">, como mostrado na Figura 2.

```
<document>
    <info type="id" value="01">
    <info type="scheme" value="S1">
    <info type="source" value="039">
    <info type="source-corpus" value="c01">
    <info type="annotator" value="M">
    <unit id="0001">Este é um texto</unit id="0001">
    <unit id="0002">que, <unit ind id="0003">embora curto,
    </unit ind id="0003"> permite a identificação de
    embutimentos</unit id="0002">
</document>
```

Figura 2. Separação do texto "Este é um texto que, embora curto, permite a identificação de embutimentos" em unidades básicas.

A necessidade de se ter embutimento, superposição e descontinuidade, por sua vez, faz com que o esquema de segmentação definido não siga realmente a definição do XML, que possui uma representação plana, formando uma hierarquia de *tags* em árvore [Krauthammer et al. 2002]. Na ausência de uma alternativa que possa satisfazer às necessidades desse projeto, optou-se por permitir esse "afrouxamento" da definição do XML. Vale lembrar, no entanto, que na ausência de tais fenômenos, o documento gerado é perfeitamente entendido como XML.

2.2.3. Tags para Classificação do Corpus

Um arquivo de anotação (ou classificação) deve começar com um *tag* <annotation> e terminar com </annotation>. A exemplo dos demais, arquivos de classificação também possuem *tags* <info>, responsáveis pela definição do identificador do arquivo, de seu anotador, do documento segmentado que esse arquivo classifica[1], além de qualquer outra informação que se julgue necessária (a depender do esquema de anotação empregado).

A classificação propriamente dita se dá por meio de *tags* <mark unit="unidade anotada"> e </mark>. Cada *tag* <mark>, por sua vez, pode conter em seu interior um número variável de *tags* <ann type="categoria da anotação" value="valor">. Assim, enquanto que <mark> faz a ligação entre a anotação feita e a unidade anotada (definida no documento segmentado apontado em <info>), <ann> corresponde à aplicação do esquema de anotação em si. Nesse caso, para cada categoria diferente do esquema de anotação há um *tag* <ann>. O nome dessa categoria é armazenado então em seu campo "type", enquanto que o valor escolhido pelo anotador vai em "value".

A Figura 3 ilustra o exemplo do arquivo apresentado na Figura 2, anotado com o esquema descrito em [Roman and Carvalho 2010]. Nela, pode-se ver, dentre outros, o identificador do arquivo de anotação, o identificador do anotador que produziu o arquivo, bem como o do arquivo contendo o texto-fonte (separado em unidades básicas de anotação). Em seguida, cada unidade desse arquivo é anotada (em *tags* <mark>) segundo três dimensões (descritas em [Roman and Carvalho 2010]), conforme ilustram as *tags* <ann>.

[1] Vale lembrar que classificações são aplicadas a unidades mínimas, ou seja, ao resultado da segmentação do texto bruto.

```
<?xml version="1.0" encoding="UTF-8"?>
<annotation>
    <info type="id" value="a01">
    <info type="scheme" value="A1">
    <info type="annotator" value="an1">
    <info type="source" value="01">
    <info type="source-corpus" value="S1">
    <mark unit="0001">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
    <mark unit="0002">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
    <mark unit="0003">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
</annotation>
```

Figura 3. Classificação das unidades da Figura 2.

Vale notar que essa codificação, a exemplo da usada no documento em estado bruto, constitui um XML bem formado. Assim, e por um requerimento do repositório ReDial, o único esquema de codificação que não segue exatamente o XML é o usado na segmentação do texto em unidades básicas de anotação.

2.2.4. Tags para Caracterização dos Participantes

Uma vez que, em geral, todas as etapas de desenvolvimento do *corpus* (ou seja, coleta, segmentação e anotação) são feitas por mais de uma pessoa, pode ser interessante ao pesquisador manter registro de algumas informações sobre cada participante, como sexo, idade, grau de escolaridade, dentre outros. Essas informações são mantidas em arquivos dentro de um sub-diretório do *corpus* para o qual o anotador contribuiu, sendo um arquivo diferente para cada participante. Cada arquivo, por sua vez, deve iniciar com a *tag* <participant> e finalizar com </participant>. Dentro dele, somente são permitidas *tags* <info>, em que são armazenadas as informações relevantes para o pesquisador, conforme ilustrado na Figura 4.

2.2.5. Múltiplos Identificadores

Algumas vezes, o arquivo resultante pode ser oriundo não de um único anotador, mas sim de vários (tomando-se a opinião da maioria, por exemplo). Casos assim são

```
<?xml version="1.0" encoding="UTF-8"?>
<participant>
    <info type="id" value="annot01">
    <info type="gender" value="m">
    <info type="age" value="20-30">
</participant>
```

Figura 4. Caracterização de um anotador.

caracterizadas pela existência de múltiplos *tags* do tipo "<info type="annotator"...>". Embora não afete a anotação em si, a existência de múltiplos anotadores requer uma mudança no padrão de nomenclatura do arquivo, com o uso do caractere "M" em vez do identificador do anotador. Esse tipo de situação, contudo, não se restringe ao anotador somente, podendo ser estendida a qualquer outro *tag* para o qual haja a necessidade de multiplicidade.

Por fim, uma vez que arquivos compostos a partir de vários anotadores representam, em geral, a opinião da maioria, cabe adicionar ao arquivo um *tag* extra: "<info type="multiple" value="majority">", ou "<info type="multiple" value="mode">", para os casos em que a maioria deve ser absoluta (ou seja, mais de 50% dos anotadores) ou apenas refletir a classificação mais comumente associada.

3. Exemplo de Uso

Como exemplo de uso, considere um resumo produzido para o *corpus* coletado conforme descrito em [Roman et al. 2006a], e anotado segundo o esquema descrito em [Roman and Carvalho 2010]. Dentro do SegDial, cada resumo desse *corpus* é armazenado em estado bruto em um arquivo isolado. Nesse caso, o resumo usado para exemplo é armazenado conforme ilustra a Figura 5, em um arquivo denominado "src_039_ c01_sum01 .xml" (o arquivo contendo o texto do diálogo foi omitido, para simplificar).

Vale notar que, além da informação normalmente necessária para gerenciamento do *corpus* (ou seja, identificadores do diálogo-fonte e seu *corpus*, bem como do arquivo com o resumo e seu *corpus*), há aqui também outras três informações, específicas do esquema de anotação e de criação do *corpus*: *viewpoint*, *constraint* e *summariser*. Nesse caso, *viewpoint* corresponde ao ponto de vista sob o qual o resumo foi gerado, enquanto que *constraint* determina se o sumarizador estava livre para produzir o texto do tamanho que melhor lhe conviesse, ou se deveria limitar-se a um máximo de 10% do número de palavras do diálogo-fonte, conforme mencionado na Seção 1. *Summariser*, por sua vez, refere-se à pessoa que produziu o resumo contido em <text> e </text>, cujos detalhes estão armazenados em um arquivo dentro de "participant" – subdiretório do diretório contendo o resumo da Figura 5.

A segmentação em unidades básicas pode ser vista na Figura 6, que apresenta o arquivo "seg_01_S1_c01_039_M.xml". Nela, a unidade básica escolhida é a oração. Vale notar que a unidade 412 possui uma unidade embutida (413), o que acaba gerando uma má formação no XML. A anotação do arquivo mostrado na Figura 6, por sua vez, pode ser vista nas Figuras 7 e 8, correspondendo à classificação desse resumo por um determinado anotador (ann01). Nesse caso, a anotação é mantida em um arquivo de nome "ann_01_A1_S1_01_ann01.xml".

```
<?xml version="1.0" encoding="UTF-8"?>
<plainDocument>
    <info type="id" value="039">
    <info type="corpus" value="c01">
    <info type="source" value="d2">
    <info type="source-corpus" value="f01">
    <info type="viewpoint" value="customer">
    <info type="constraint" value="free">
    <info type="summariser" value="sum01">
    <text>
        A espera foi um pouco longa, mas finalmente alguém
        me atendeu. Pedi que me falasse sobre o carro. O
        vendedor comentou sobre suas principais caracterís-
        ticas, mostrou-me o interior, o porta-malas e ou-
        tras coisas. Um carro aparentemente bonito e ele-
        gante, e seguro também, mas muito caro, o que, jun-
        to com a pouca vontade em me atender, me fizeram
        voltar para casa sem comprá-lo.
    </text>
</plainDocument>
```

Figura 5. Resumo 039 em estado bruto.

Por fim, o anotador responsável pelo resultado mostrado nas Figuras 7 e 8 está definido, dentro do subdiretório "participant", no diretório correspondente ao esquema A1, em um arquivo denominado "part_ann01_A1.xml", conforme ilustra a Figura 9. Note que seu identificador – ann01 – bate com o identificador usado no campo "annotator" da Figura 7. Da mesma forma, o sumarizador responsável pelo arquivo mostrado na Figura 5 também possui um registro semelhante, conforme mencionado, dentro do diretório do *corpus* em estado bruto (ou seja, "c01").

4. Conclusão

Esse relatório descreve o esquema de codificação XML utilizado no projeto ResDial. Nele, pode-se ver como cada *corpus* do projeto é armazenado dentro de diretórios, bem como o modo que a informação relevante a cada etapa de anotação deve ser incluída nos arquivos. Embora o projeto siga majoritariamente o modelo XML, algumas necessidades, como a existência de superposição e embutimento, por exemplo, fazem com que alguns arquivos não sigam exatamente esse padrão.

Projetado de maneira a permitir *stand-off annotation*, cada *corpus* dentro do projeto possui pelo menos três arquivos relacionados: um arquivo com o texto em estado bruto, um com o texto segmentado em unidades básicas de anotação, e um com a classificação de cada uma dessas unidades conforme algum esquema de anotação específico. Naturalmente, tanto a parte de segmentação quanto de classificação admitem a inclusão de múltiplos arquivos para o mesmo texto-fonte (produto de diferentes anotadores), permitindo assim que comparações e análises de concordância possam ser executadas.

Essa padronização da codificação dos *corpora* no ResDial, por sua vez, não serve somente ao propósito de tornar mais fácil a leitura e uso desses *corpora* por outras pessoas,

```
<document>
    <info type="id" value="01">
    <info type="scheme" value="S1">
    <info type="source" value="039">
    <info type="source-corpus" value="c01">
    <info type="annotator" value="M">
    <unit id="0400">A espera foi um pouco longa,</unit
    id="0400">
    <unit id="0401">mas finalmente alguém me atendeu.</unit
    id="0401">
    <unit id="0402">Pedi</unit id="0402">
    <unit id="0403">que me falasse sobre o carro.</unit
    id="0403">
    <unit id="0404">O vendedor comentou sobre suas
    principais características,</unit id="0404">
    <unit id="0405">mostrou-me o interior,</unit id="0405">
    <unit id="0406">o porta-malas</unit id="0406">
    <unit id="0407">e outras coisas.</unit id="0407">
    <unit id="0408">Um carro aparentemente bonito</unit
    id="0408">
    <unit id="0409">e elegante,</unit id="0409">
    <unit id="0410">e seguro também,</unit id="0410">
    <unit id="0411">mas muito caro,</unit id="0411">
    <unit id="0412">o que, <unit ind id="0413">junto com
    a pouca vontade em me atender,</unit ind id="0413"> me
    fizeram voltar para casa</unit id="0412">
    <unit id="0414">sem comprá-lo.</unit id="0414">
</document>
```

Figura 6. Separação do resumo 039 em orações.

mas também permitir o desenvolvimento das mais variadas ferramentas para o projeto. Nesse sentido, já foi desenvolvido o protótipo de um segmentador (ver [Rodrigues et al. 2012]), havendo também outras ferramentas em desenvolvimento no presente momento.

Referências

Ide, N. and Brew, C. (2000). Requirements, tools, and architectures for annotated corpora. In *Proceedings of Data Architectures and Software Support for Large Corpora*, pages 1–5, Paris, France. European Language Resources Association.

Krauthammer, M., Johnson, S. B., Hripcsak, G., Campbell, D. A., and Friedman, C. (2002). Representing nested semantic information in a linear string of text using xml. In *Proceedings of the AMIA 2002 Symposium*, pages 405–409, San Antonio, TX, USA. ISBN 1-56053-600-4.

O'Donnell, M. (2008). The uam corpustool: software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA*, Almeria, Spain.

Ogren, P. V. (2006). Knowtator: A plug-in for creating training and evaluation data sets for biomedical natural language systems. In *Proceedings of the 9th International Protégé*

```
<?xml version="1.0" encoding="UTF-8"?>
<annotation>
    <info type="id" value="01">
    <info type="scheme" value="A1">
    <info type="annotator" value="ann01">
    <info type="source" value="01">
    <info type="source-corpus" value="S1">
    <mark unit="0400">
        <ann type="CATE" value="Relato negativo sobre o
                                atendente">
        <ann type="INTE" value="Baixa">
        <ann type="CONS" value="-">
    </mark>
    <mark unit="0401">
        <ann type="CATE" value="Relato negativo sobre o
                                atendente">
        <ann type="INTE" value="Não-Baixa">
        <ann type="CONS" value="-">
    </mark>
    <mark unit="0402">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
    <mark unit="0403">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
    <mark unit="0404">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
    <mark unit="0405">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>
    <mark unit="0406">
        <ann type="CATE" value="Relato neutro">
        <ann type="INTE" value="">
        <ann type="CONS" value="">
    </mark>

    (continua ...)
```

Figura 7. Classificação das unidades da Figura 2.

```
(... continuação)

 <mark unit="0407">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
<mark unit="0408">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
<mark unit="0409">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
<mark unit="0410">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
<mark unit="0411">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
<mark unit="0412">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
<mark unit="0413">
    <ann type="CATE" value="Relato negativo sobre o
                            atendente">
    <ann type="INTE" value="Não-Baixa">
    <ann type="CONS" value="-">
</mark>
<mark unit="0414">
    <ann type="CATE" value="Relato neutro">
    <ann type="INTE" value="">
    <ann type="CONS" value="">
</mark>
</annotation>
```

Figura 8. Classificação das unidades da Figura 2 (cont.).

```
<?xml version="1.0" encoding="UTF-8"?>
<participant>
    <info type="id" value="ann01">
    <info type="target-corpus" value="A1">
    <info type="gender" value="m">
    <info type="age" value="20-30">
    <info type="study" value="mphil student">
    <info type="area" value="exact sciences">
</participant>
```

Figura 9. Caracterização do anotador da Figura 7.

Conference, Stanford, USA.

Orăsan, C. (2003). Palinka: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39–43, Sapporo, Japan.

Reidsma, D., sa Jovanović, N., and Hofs, D. (2005). Designing annotation tools based on properties of annotation problems. In *Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*.

Rodrigues, F., Semolini, R., Roman, N. T., and Monteiro, A. M. (2012). Tseg – a text segmenter for corpus annotation. In *Proceedings of the VIII Brazilian Symposium on Information Systems (SBSI 2012)*, São Paulo, SP, Brazil.

Roman, N. T. and Carvalho, A. M. B. R. (2010). A multi-dimensional annotation scheme for behaviour in dialogues. In Kuri-Morales, A. and Simari, G. R., editors, *Proceedings of the 12th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2010)*, volume 6433 of *Advances in Artificial Intelligence*, pages 386–395, Bahía Blanca, Argentina. Springer. ISBN: 978-3-642-16951-9.

Roman, N. T., Piwek, P., and Carvalho, A. M. B. R. (2006a). *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, chapter Politeness and Bias in Dialogue Summarization: Two Exploratory Studies, pages 171–185. Springer Netherlands, Dordrecht, The Netherlands. ISBN: 1-4020-4026-1.

Roman, N. T., Piwek, P., and Carvalho, A. M. B. R. (2006b). A web-experiment on dialogue classification. In Rezende, S. O. and da Silva Filho, A. C. R., editors, *Proceedings of the Fourth Workshop in Information and Human Language Technology (TIL'2006)*, Ribeir ao Preto, Brazil. ICMC-USP.

van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171. ISSN: 2190-0949.

Verhagen, M. (2010). The brandeis annotation tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3638–3643, Valletta, Malta.

A. Conjunto de Tags Usadas no ResDial

Tag	*Onde Usada*	*Uso*
ann	Texto anotado	Define a classificação de uma unidade de anotação, de acordo com alguma categoria pré-definida. Possui os parâmetros "type", indicando o nome da categoria, e "value", correspondendo a seu valor. Deve ser usado dentro de mark.
annotation	Texto anotado	Define os limites do texto anotado. Possui correspondente </annotation>
document	Texto segmentado	Define os limites do texto segmentado. Possui correspondente </document>
info	Todos os documentos	Apresenta informação adicional ao arquivo. Possui os parâmetros "type", usado para definir o nome da informação adicionada, e "value", que carrega o valor dessa informação
mark	Texto anotado	Define a aplicação do esquema de anotação a uma unidade básica de anotação. Possui o parâmetro "unit", contendo o identificador da unidade anotada. Possui correspondente </mark>
participant	Participante	Define os limites dos dados de caracterização de um participante. Possui correspondente </participant>
plainDocument	Texto bruto	Define os limites do texto em estado bruto. Possui correspondente </plainDocument>
text	Texto bruto	Define o texto do documento (em UTF-8). Possui correspondente </text>
unit	Texto segmentado	Define uma unidade mínima de anotação. Possui o parâmetro "id", que deve carregar o identificador único da unidade. Possui correspondente </unit>
unit ind	Texto segmentado	Define uma unidade independente. Possui o parâmetro "id", que deve carregar o identificador único da unidade. Possui correspondente </unit ind>
<!--	Todos os documentos	Define um comentário. Possui correspondente -->