Formats de représentation des alignements pour TransRead et Alibi

Charlotte Rudnik Yong Xu Guillaume Wisniewski
François Yvon
LIMSI-CNRS, rue John von Neumann, Orsay CEDEX, France
{Prénom.Nom}@limsi.fr

1 Introduction

Le format de représentation des alignements proposé ici reprend le format proposé dans le livrable 1.1 Format de représentation des alignements par Yong Xu [Xu et al., 2013], il l'enrichit en particulier pour préciser des notions de parsing et pour en faciliter l'utilisation lors de la recherche d'information.

2 Proposition v1.3 du format Transread

2.1 Le format d'annotation

Dans cette section nous décrivons le format d'annotation et le schéma XML correspondant. Le schéma XSD et un exemple de document annoté selon ce schéma sont donnés en annexe.

2.1.1 L'élément `trAnnot`

L'élément hiérarchiquement le plus haut est `<trAnnot>`. Cette balise marque le début et la fin d'un document d'annotation. Elle contient un sous-élément `<docList>`, un ou plusieurs sous-éléments `<linkList>`. Il dispose d'un seul attribut obligatoire "`version`", dont la valeur indique la version du schéma XML utilisée par le document.

2.1.2 L'élément `docList`

Un élément `<docList>` contient au moins un sous-élément `<docName>`.

2.1.3 L'élément `docName`

L'élément `<docName>` décrit un document original. Il a deux attributs :
— id : l'identifiant du document. Cet identifiant est utilisé dans les positions des unités.
— xml :lang : la langue du document.

2.1.4 L'élément `linkList`

Un élément `<linkList>` regroupe tous les liens d'un même niveau linguistique (phrase, mot, etc). Il contient un ou plusieurs sous-éléments `<linkGroup>`. Un `<linkList>` possède un attribut obligatoire "`level`", dont les valeurs possibles sont "`sentence`" (annotations au niveau de phrases), "`token`" (celles au niveau de mots) et "`chunk`" (au niveau de segments).

2.1.5 L'élément `linkGroup`

Un `<linkGroup>` est un groupe de liens, dont les contenus sont extraits d'un même fragment d'un document. Un `<linkGroup>` contient plusieurs sous-éléments `<docPart>` qui indiquent les fragments des documents, suivis par une liste de `<link>` ou une liste de `<annotation>`. Tous les `<link>` ou `<annotation>` d'un `<linkGroup>` ont le niveau linguistique indiqué par le parent `<linkList>`. Les unités textuelles dans les `<link>` ou `<annotation>` se trouvent strictement dans les fragments indiqués par les `<docPart>`. Un `<linkGroup>` a un attribut obligatoire "`type`", dont les valeurs possibles sont "`alignment`" et "`annotation`". Si la valeur de "`type`" est "`alignment`", ce `<linkGroup>` ne peut pas contenir des sous-éléments `<annotation>` ; si cette valeur est "`annotation`", il ne peut pas contenir des `<link>`.

2.1.6 L'élément `docPart`

Un `<docPart>` indique un fragment d'un document. Il a trois attributs :
- `doc` : attribut obligatoire, indique un document. La valeur doit être une référence d'un *id* d'un élément `<docName>`.
- `beginPos` : attribut facultatif, indique la position du début de ce fragment dans le document.
- `endPos` : attribut facultatif, indique la position de la fin de ce fragment dans le document.

Un `<docPart>` est un élément vide. La présence de cette balise a pour objectif d'accélérer la recherche des informations dans les documents d'annotations volumineux.

La position des débuts ou fin de fragments s'appuie sur l'adressage défini par Xu et al. [2013] qui en détaille le format. En résumé, chaque unité est identifiée par le chemin de la racine vers cette unité dans l'arbre DOM (Document Object Model [1]) du document. L'adressage est fait au niveau des caractères. Le chemin d'un caractère se compose de trois parties : l'identifiant du document, la position de l'élément contenant le caractère dans l'arbre DOM du document (donc toujours un nœud `<text>`), et la position relative du caractère dans cet élément `<text>`. Tous les indices positionnels commencent par la valeur 0, et sont séparés par des points.

Par exemple, le quatrième caractère du troisième fils du deuxième fils du `<document>` a pour position "`1.2.3`". Cette position et l'identifiant du document sont séparés par un espace. Donc pour l'exemple ci-dessus, si l'*id* du document est "`doc`", alors la position complète du caractère en question est "`doc 1.2.3`".

2.2 Le format d'annotation

Dans cette section nous décrivons le format d'annotation. Le schéma XML correspondant, ainsi qu'un exemple de document annoté selon ce schéma sont donnés en annexe.

1. `http://www.w3.org/DOM/`

2.2.1 L'élément `link`

Un `<link>` spécifie une unité textuelle d'un document et sa correspondance dans le document parallèle. Il inclut un sous-élément `<docSpan>` qui décrit une unité, et un autre sous-élément *facultatif* `<docSpan>` qui décrit l'unité correspondante dans l'autre document. L'absence du deuxième `<docSpan>` signifie un lien nul. Il faut remarquer que, dans un `<linkGroup>`, tous les sous-éléments `<docSpan>` doivent venir d'un fragment parmi les sous-éléments `<docPart>` de ce `<linkGroup>`. Un `<link>` possède deux attributs :

- `id` : attribut obligatoire, tel que `"align_tok_15"` ou `" 2.0_2-5_3-6 "` comme dans les exemples fournis en annexe.
- `certainty` : attribut facultatif, la valeur est un nombre entre 0 et 1, indiquant le niveau de confiance de l'annotateur sur ce `<link>`.
- `parentID` : attribut facultatif, la valeur doit être l'`id` d'un `<link>` parent existant. Par défaut si le lien est n'a pas de parent, alors on pourra introduire l'`id` `"ROOT"`.

A titre d'exemple :

```
<?xml version="1.0" encoding="utf-8"?>
<link certainty="1" id="2.1_0-18_0-18" parentID="2.0_0-38_0-35">
     <docSpan beginPos="doc_fr 1.2.5.9.7.1.15.1.1"
                endPos="doc_fr 1.2.5.9.7.1.15.1.83">
     Il etait une fois un homme qui avait de belles maisons a la ville
         et a la campagne ,</docSpan>
     <docSpan beginPos="doc_en 1.2.5.9.7.3.1.4.0.0.0"
                endPos="doc_en 1.2.5.9.7.3.1.4.1.74">
     Once upon a time there was a man who had fine houses , both in
         town and country ,</docSpan>
</link>
```

2.2.2 L'élément `annotation`

Un `<annotation>` encode des propriétés attachées aux unités. Par exemple, des analyses linguistiques peuvent permettre d'identifier son lemme, sa catégorie grammaticale, etc ; la relation entre un mot et des entrées de dictionnaires est un autre type d'informations. L'unité peut être un objet non textuel, par exemple une image ou une vidéo, où nous pouvons annoter la durée, la langue, etc. Un `<annotation>` contient un `<docSpan>`, et 0, 1 ou plusieurs `<mark>`.

Un `<annotation>` possède deux attributs :

- `id` : attribut obligatoire, l'identifiant de cette `<annotation>`.
- `type` : attribut obligatoire, qui encode le type de l'annotation. Les valeurs possibles sont `"gram"`, `"QE"`, `"URI"` . Le type `"gram"` est utilisé pour enregistrer les résultats obtenus par l'analyse syntaxique ; `"QE"` encode les résultats de l'estimation de qualité de traduction ; `"URI"` indique les liens vers les ressources externes, qui sont utilisés par exemple pour la désambiguïsation des mots ;

La liste des types d'information est évolutive et pourra être complétée au cours des évolutions du format.

2.2.3 L'élément `docSpan`

Un élément `<docSpan>` permet d'identifier une unité à l'intérieur du document. Il peut posséder jusqu'à sept attributs :

— `beginPos` : attribut facultatif, la valeur doit indiquer la position du caractère du début de l'unité.
— `endPos` : attribut facultatif, la valeur doit indiquer la position du caractère de la fin de l'unité.
— `tokenID` : attribut facultatif, la valeur doit indiquer l'identifiant du token lorsque l'unité est constituée d'un seul token. Cet attribut a pour but de simplifier les traitements, et n'a de sens que si le token est par ailleurs identifié dans un `<link>` de type `token`. Dans l'exemple fourni en annexe, le token est identifié par trois éléments : l'identifiant du document `doc`, le numéro de la phrase dans laquelle il se trouve et sa positon au sein de la phrase. Ainsi, le douzième token de la quatrième phrase du document `doc_en` aura pour identifiant "`doc_en 4.11`". Le compte des phrases commençant à 1, tandis que celui des tokens commence à 0.
— `beginTok` : attribut facultatif, la valeur doit indiquer le `tokenID` du token du début de l'unité.
— `endTok` : attribut facultatif, la valeur doit indiquer le `tokenID` du token de fin de l'unité.
— `sentID` : attribut facultatif, l'identifiant d'une phrase. Cet attribut a pour but de simplifier les traitements. Il n'a de sens que si la phrase est par ailleurs identifiée dans un `<link>` de type `sentence`.
— `context` : attribut facultatif, la valeur doit être une série d'*ids* de `<link>` ou `<annotation>`, dans lesquels se trouvent les contextes de l'unité.

Nous pouvons mettre ou non le contenu textuel de l'unité dans l'élément `<docSpan>`.

2.2.4 L'élément `mark`

L'élément `<mark>` contient des informations relativement riches, car cet élément stocke les informations attachées aux unités. Souvent un `<annotation>` contient plusieurs `<mark>`. Il se peut à l'inverse qu'aucun `<mark>` ne figure dans un `<annotation>`, au cas où aucune information n'a été trouvée.

Un `<mark>` dispose de nombreux attributs :

— `certainty` : attribut facultatif, qui indique le niveau de confiance de l'annotateur sur ce `<mark>`. La valeur doit être un nombre entre 0 et 1.
— `cat` : attribut facultatif, utilisé dans les `<annotation>` de type "`gram`". La valeur indique la catégorie du label linguistique. Les valeurs possibles sont "`POS`" (*Part Of Speech*), "`lemma`" (le lemme) et "`parse`" (arbre de paring), mais il sera possible d'étendre au besoin les possibilités des valeurs avec d'autres informations linguistiques. Nous pouvons se référer au ISOcat (ISO TC 37 Terminology and Other Language and Content Resources) [2] afin d'avoir une idée pour des développements possibles.
— `resource` : attribut facultatif, utilisé dans les `<annotation>` de type "`URI`". La valeur indique la ressource externe. Par exemple "`babelnet`", "`wordnet`", "`wiktionnary`" etc.
— xml :lang : attribut facultatif, la valeur doit être un code de langue existant dans la norme ISO 639-1 [3]. Cet attribut peut être utilisé pour, par exemple, indiquer la langue des explications trouvées dans les ressources externes.
— `entry` : attribut facultatif, utilisé dans les `<annotation>` de type "`URI`". La valeur est une entrée de la ressource externe.

2. `http://www.isocat.org/`
3. `http://www.loc.gov/standards/iso639-2/php/code_list.php`

— `qescore` : attribut facultatif, utilisé dans les `<annotation>` de type "`QE`". La valeur est le score de l'estimation de qualité sur l'unité.
— `method` : attribut facultatif, utilisé dans les `<annotation>` de type "`QE`". La valeur indique la méthode de l'estimation de qualité.

Pour les `<mark>` apparaissant dans les `<annotation>` de type "`gram`", le contenu de l'élément `<mark>` doit être le label linguistique ; pour ceux de type "`URI`", le contenu de l'élément `<mark>` doit être des informations associées au `<docSpan>`.

Dans l'exemple annoté disponible en annexe, les annotations sont de deux types : soit au niveau des tokens, soit au niveau des phrases. Voici un exemple d'annotation au niveau `token` de "`a rock`" :

```
<?xml version="1.0" encoding="utf-8"?>
<annotation certainty="1" id="doc_en annot_token_1371" type="gram">
   <docSpan beginPos="doc_en 1.2.5.9.7.3.1.52.0.157"
   endPos="doc_en 1.2.5.9.7.3.1.52.0.160"
   tokenID="doc_en 40.5">rock</docSpan>
   <mark cat="POS" certainty="1">NN</mark>
   <mark cat="parse" certainty="1">nsubj(she-13, rock-6)</mark>
</annotation>
<annotation certainty="1" id="doc_en annot_token_1372" type="gram">
   <docSpan beginPos="doc_en 1.2.5.9.7.3.1.52.0.155"
   endPos="doc_en 1.2.5.9.7.3.1.52.0.155"
   tokenID="doc_en 40.4">a</docSpan>
   <mark cat="POS" certainty="1">DT</mark>
   <mark cat="parse" certainty="1">det(rock-6, a-5)</mark>
</annotation>
```

Voici un exemple d'annotation au niveau `sentence` :

```
<?xml version="1.0" encoding="utf-8"?>
<annotation certainty="1" id="doc_en annot_sent_67" type="gram">
   <docSpan beginPos="doc_en 1.2.5.9.7.3.1.108.0.0"
   endPos="doc_en 1.2.5.9.7.3.1.108.0.74"
   sentID="doc_en 68" />
   <mark cat="parse" certainty="1">(S (S (NP (NNP Blue) (NNP Beard)) (
       VP (VBD had) (NP (DT no) (NNS heirs)))) (, ,) (CC and) (IN so) (S
        (NP (PRP$ his) (NN wife)) (VP (VBD became) (NP (NP (NN mistress)
       ) (PP (IN of) (NP (PDT all) (PRP$ his) (NN estate)))))) (. .))</
       mark>
</annotation>
```

Pour l'anglais, le label label linguistique est issu du MaxentTagger de Stanford et l'informartion de dépendences syntaxiques (dependency parsing) est issue du modèle englishPCFG de Stanford. Pour le français, les modèles utilisés sont le MaxentTagger pour les labels linguistiques, et le frenchFactored pour les informations syntaxiques.

Un exemple annoté extrait du livre de F. Cooper « The last of the Mohicans » est donné ci-dessous.

Références

Yong Xu, Guillaume Wisniewski, and François Yvon. Formats de représentation des alignements pour transread. https ://transread.limsi.fr/Deliverables/, 2013.

La schéma XSD

Listing 1 – Le schéma XSD complet

```
<?xml version="1.0" encoding="utf-8"?>

<!--                                                                -->
<!--                                                                -->
<!--              TRANSREAD annotation XML schema                   -->
<!--                                                                -->
<!--                         Version  1.3                           -->
<!--                                                                -->


<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  targetNamespace="http://transread.limsi.fr"
  xmlns="http://transread.limsi.fr"
  elementFormDefault="qualified">

  <xsd:import namespace="http://www.w3.org/XML/1998/namespace"
              schemaLocation="http://www.w3.org/2001/xml.xsd"/>

  <xsd:simpleType name="certaintytype">
    <xsd:restriction base="xsd:decimal">
      <xsd:minInclusive value="0.0"/>
      <xsd:maxInclusive value="1.0"/>
    </xsd:restriction>
  </xsd:simpleType>

  <xsd:simpleType name="qescoretype">
    <xsd:restriction base="xsd:decimal">
      <xsd:minInclusive value="0.0"/>
    </xsd:restriction>
  </xsd:simpleType>

  <xsd:simpleType name="shortpostype">
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="([0-9]+\.)+[0-9]+-[0-9]+"/>
    </xsd:restriction>
  </xsd:simpleType>

  <xsd:simpleType name="tmppostype">
    <xsd:union memberTypes="xsd:IDREF shortpostype"/>
  </xsd:simpleType>

  <xsd:simpleType name="postype">
    <xsd:restriction>
      <xsd:simpleType>
        <xsd:list itemType="tmppostype"/>
      </xsd:simpleType>
      <xsd:pattern value="[0-9a-zA-Z_-]+ ([0-9]+\.)+[0-9]+-[0-9]+"/>
    </xsd:restriction>
  </xsd:simpleType>
```

```
50
51  <xsd:simpleType name="toktype">
52    <xsd:restriction base="xsd:string">
53      <xsd:pattern value="[0-9a-zA-Z_-]+ [0-9]+\.[0-9]+"/>
54    </xsd:restriction>
55  </xsd:simpleType>
56
57  <xsd:simpleType name="cattype">
58    <xsd:restriction base="xsd:string">
59      <xsd:enumeration value="POS"/>
60      <xsd:enumeration value="lemma"/>
61      <xsd:enumeration value="parse"/>
62    </xsd:restriction>
63  </xsd:simpleType>
64
65  <xsd:simpleType name="annotoption">
66    <xsd:restriction base="xsd:string">
67      <xsd:enumeration value="gram"/>
68      <xsd:enumeration value="URI"/>
69      <xsd:enumeration value="QE"/>
70    </xsd:restriction>
71  </xsd:simpleType>
72
73    <xsd:simpleType name="linkgroupoption">
74      <xsd:restriction base="xsd:string">
75        <xsd:enumeration value="alignment"/>
76        <xsd:enumeration value="annotation"/>
77      </xsd:restriction>
78    </xsd:simpleType>
79
80    <xsd:simpleType name="methodtype">
81      <xsd:restriction base="xsd:string">
82        <xsd:enumeration value="method1"/>
83      </xsd:restriction>
84    </xsd:simpleType>
85
86    <xsd:simpleType name="leveltype">
87      <xsd:restriction base="xsd:string">
88        <xsd:enumeration value="sentence"/>
89        <xsd:enumeration value="token"/>
90        <xsd:enumeration value="chunk"/>
91      </xsd:restriction>
92    </xsd:simpleType>
93
94    <xsd:complexType name="marktype">
95      <xsd:simpleContent>
96        <xsd:extension base="xsd:string">
97          <xsd:attribute name="cat" type="cattype"/>
98          <xsd:attribute name="resource" type="xsd:string"/>
99          <xsd:attribute ref="xml:lang"/>
100         <xsd:attribute name="entry" type="xsd:string"/>
101         <xsd:attribute name="certainty" type="certaintytype"/>
```

```
102           <xsd:attribute name="qescore" type="qescoretype"/>
103           <xsd:attribute name="method" type="methodtype"/>
104         </xsd:extension>
105       </xsd:simpleContent>
106     </xsd:complexType>
107
108   <xsd:complexType name="docspantype">
109     <xsd:simpleContent>
110       <xsd:extension base="xsd:string">
111         <xsd:attribute name="beginPos" type="postype" use="required"
              />
112         <xsd:attribute name="endPos" type="postype" use="required"/>
113       <xsd:attribute name="tokenID" type="xsd:CDATA"/>
114         <xsd:attribute name="beginTok" type="toktype"/>
115         <xsd:attribute name="endTok" type="toktype"/>
116         <xsd:attribute name="tokenID" type="toktype"/>
117         <xsd:attribute name="sentID" type="xsd:CDATA"/>
118         <xsd:attribute name="context" type="xsd:IDREFS"/>
119       </xsd:extension>
120     </xsd:simpleContent>
121   </xsd:complexType>
122
123   <xsd:complexType name="annotationtype">
124     <xsd:sequence>
125       <xsd:element name="docSpan" type="docspantype"/>
126       <xsd:element name="mark" type="marktype" maxOccurs="unbounded"
              />
127     </xsd:sequence>
128     <xsd:attribute name="type" type="annotoption" use="required"/>
129     <xsd:attribute name="id" type="xsd:ID" use="required"/>
130   </xsd:complexType>
131
132   <xsd:complexType name="linktype">
133     <xsd:sequence>
134       <xsd:element name="docSpan" type="docspantype" minOccurs="1"
            maxOccurs="unbounded"/>
135     </xsd:sequence>
136     <xsd:attribute name="id" type="xsd:ID" use="required"/>
137     <xsd:attribute name="certainty" type="certaintytype"/>
138     <xsd:attribute name="parentID" type="xsd:IDREFS"/>
139   </xsd:complexType>
140
141   <xsd:complexType name="docparttype">
142     <xsd:attribute name="doc" type="xsd:IDREF" use="required"/>
143     <xsd:attribute name="beginPos" type="postype"/>
144     <xsd:attribute name="endPos" type="postype"/>
145   </xsd:complexType>
146
147   <xsd:complexType name="linkgrouptype">
148     <xsd:sequence>
149       <xsd:element name="docPart" type="docparttype" maxOccurs="2"/>
150       <xsd:choice>
```

```
151         <xsd:element name="link" type="linktype" minOccurs="0"
                maxOccurs="unbounded"/>
152         <xsd:element name="annotation" type="annotationtype"
                minOccurs="0" maxOccurs="unbounded"/>
153       </xsd:choice>
154     </xsd:sequence>
155     <xsd:attribute name="type" type="linkgroupoption" use="required"
            />
156   </xsd:complexType>
157
158   <xsd:complexType name="linklisttype">
159     <xsd:sequence>
160       <xsd:element name="linkGroup" type="linkgrouptype" maxOccurs="
              unbounded"/>
161     </xsd:sequence>
162     <xsd:attribute name="level" type="leveltype"/>
163   </xsd:complexType>
164
165   <xsd:complexType name="typenametype">
166     <xsd:simpleContent>
167       <xsd:extension base="xsd:string">
168         <xsd:attribute name="id" type="xsd:ID" use="required"/>
169       </xsd:extension>
170     </xsd:simpleContent>
171   </xsd:complexType>
172
173   <xsd:complexType name="linktypetype">
174     <xsd:sequence>
175       <xsd:element name="typeName" type="typenametype" maxOccurs="
              unbounded"/>
176     </xsd:sequence>
177   </xsd:complexType>
178
179
180   <xsd:complexType name="docnametype">
181     <xsd:simpleContent>
182       <xsd:extension base="xsd:string">
183         <xsd:attribute name="id" type="xsd:ID" use="required"/>
184         <xsd:attribute ref="xml:lang"/>
185       </xsd:extension>
186     </xsd:simpleContent>
187   </xsd:complexType>
188
189   <xsd:complexType name="doclisttype">
190     <xsd:sequence>
191       <xsd:element name="docName" type="docnametype" maxOccurs="
              unbounded"/>
192     </xsd:sequence>
193   </xsd:complexType>
194
195   <xsd:complexType name="trannottype">
196     <xsd:sequence>
```

```
197         <xsd:element name="docList" type="doclisttype"/>
198         <xsd:element name="linkType" type="linktypetype" minOccurs="0"
                maxOccurs="1" />
199         <xsd:element name="linkList" type="linklisttype" maxOccurs="
                unbounded"/>
200       </xsd:sequence>
201       <xsd:attribute name="version" type="xsd:decimal" use="required"/
              >
202     </xsd:complexType>
203
204     <xsd:element name="trAnnot" type="trannottype"/>
205 </xsd:schema>
```

Un exemple annoté

Listing 2 – Sample annotation

```
<?xml version="1.3" encoding="UTF-8"?>
<!-- $Id: -->
<trAnnot xmlns="http://transread.limsi.fr"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://transread.limsi.fr/Resources/
             transread.xsd"
         version="1.1">
  <docList>
    <docName id="doc_en" xml:lang="en">Mohicans_en.xhtml</docName>
    <docName id="doc_fr" xml:lang="fr">Mohicans_fr.xhtml</docName>
  </docList>
  <linkList level="sentence">
    <linkGroup type="alignment">
      <docPart doc="doc_en"/>
      <docPart doc="doc_fr"/>
      <link id="align_sent_1" certainty="1">
        <docSpan beginPos="doc_en 1.2.5.0.0-0" endPos="doc_en
            1.2.5.0.0-46" />
        <docSpan beginPos="doc_fr 1.2.5.0.0-0" endPos="doc_fr
            1.2.5.0.0-45" />
      </link>
      <link id="align_sent_2" certainty="1">
        <docSpan beginPos="doc_en 1.2.7.0.0-0" endPos="doc_en
            1.2.7.0.0-9" />
        <docSpan beginPos="doc_fr 1.2.7.0.0-0" endPos="doc_fr
            1.2.7.0.0-16" />
      </link>
      <link id="align_sent_5" certainty="1">
        <docSpan beginPos="doc_en 1.2.11.0-0" endPos="doc_en
            1.2.11.0-171" />
        <docSpan beginPos="doc_fr 1.2.11.0-0" endPos="doc_fr
            1.2.11.0-243" />
      </link>
      <link id="align_sent_6" certainty="1">
        <docSpan beginPos="doc_en 1.2.11.0-172" endPos="doc_en
            1.2.11.0-300" />
        <docSpan beginPos="doc_fr 1.2.11.0-244" endPos="doc_fr
            1.2.11.0-386" />
      </link>
      <link id="align_sent_7" certainty="1">
        <docSpan beginPos="doc_en 1.2.11.0-301" endPos="doc_en
            1.2.11.0-582" />
        <docSpan beginPos="doc_fr 1.2.11.0-387" endPos="doc_fr
            1.2.11.0-692" />
      </link>
    </linkGroup>
    <linkGroup type="annotation">
      <docPart doc="doc_en" beginPos="doc_en 1.2.5.0-0" endPos="
          doc_en 1.2.11.0-582"/>
```

```
38        <annotation id="annot_sent_1" type="map">
39          <docSpan beginPos="doc_en 1.2.5.0.0-0" endPos="doc_en
                1.2.5.0.0-46" sentID="s1"></docSpan>
40        </annotation>
41        <annotation id="annot_sent_2" type="map">
42          <docSpan beginPos="doc_en 1.2.7.0.0-0" endPos="doc_en
                1.2.7.0.0-9" sentID="s2"/></docSpan>
43        </annotation>
44    </linkList>
45    <linkList level="token">
46      <linkGroup type="alignment">
47        <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-0" endPos="
              doc_en 1.2.11.0-171"/>
48        <docPart doc="doc_fr" beginPos="doc_fr 1.2.11.0-0" endPos="
              doc_fr 1.2.11.0-243"/>
49        <link id="align_tok_40">
50          <docSpan beginPos="doc_en 1.2.11.0-0" endPos="doc_en
                1.2.11.0-2" tokenID="s1.t0">it</docSpan>
51          <docSpan beginPos="doc_fr 1.2.11.0-0" endPos="doc_fr
                1.2.11.0-2" tokenID="s1.t0">c'</docSpan>
52        </link>
53        <link id="align_tok_41">
54          <docSpan beginPos="doc_en 1.2.11.0-3" endPos="doc_en
                1.2.11.0-6" tokenID="s1.t1">was</docSpan>
55          <docSpan beginPos="doc_fr 1.2.11.0-3" endPos="doc_fr
                1.2.11.0-8" tokenID="s1.t1">était</docSpan>
56        </link>
57        <link id="align_tok_42">
58          <docSpan beginPos="doc_en 1.2.11.0-7" endPos="doc_en
                1.2.11.0-8" tokenID="s1.t2">a</docSpan>
59          <docSpan beginPos="doc_fr 1.2.11.0-9" endPos="doc_fr
                1.2.11.0-11" tokenID="s1.t2">un</docSpan>
60        </link>
61        <link id="align_tok_43">
62          <docSpan beginPos="doc_en 1.2.11.0-9" endPos="doc_en
                1.2.11.0-16" tokenID="s1.t3">feature</docSpan>
63          <docSpan beginPos="doc_fr 1.2.11.0-16" endPos="doc_fr
                1.2.11.0-26" tokenID="s1.t3">caractères</docSpan>
64        </link>
65        <link id="align_tok_44">
66          <docSpan beginPos="doc_en 1.2.11.0-17" endPos="doc_en
                1.2.11.0-25" tokenID="s1.t4">peculiar</docSpan>
67          <docSpan beginPos="doc_fr 1.2.11.0-27" endPos="doc_fr
                1.2.11.0-39" tokenID="s1.t4">particuliers</docSpan>
68        </link>
69        <link id="align_tok_66">
70          <docSpan beginPos="doc_en 1.2.11.0-122" endPos="doc_en
                1.2.11.0-133" tokenID="s1.t10">encountered</docSpan>
71          <docSpan beginPos="doc_fr 1.2.11.0-133" endPos="doc_fr
                1.2.11.0-139" tokenID="s1.t9">braver</docSpan>
72        </link>
73      </linkGroup>
```

```
74      <linkGroup type="annotation">
75        <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-0" endPos="
              doc_en 1.2.11.0-171"/>
76        <annotation id="annot_tok_1" type="gram">
77          <docSpan beginPos="doc_en 1.2.11.0-122" endPos="doc_en
                1.2.11.0-133" tokenID="s1.t10">encountered</docSpan>
78          <mark cat="lemma" certainty="1">encounter</mark>
79          <mark cat="POS" certainty="1">VBN</mark>
80    <mark cat="parse" certainty="1" method="basic-dependencies">
          dep:root:ROOT</mark>
81          <mark cat="parse" certainty="1" method="basic-dependencies">
                gov:nsubj:mohicans</mark>
82          <mark cat="parse" certainty="1" method="collapsed-
                dependencies">dep:root:ROOT</mark>
83        </annotation>
84        <annotation id="annot_tok_2" type="URI">
85          <docSpan beginPos="doc_en 1.2.11.0-122"
86                   endPos="doc_en 1.2.11.0-133" tokenID="s1.t10">
                       encountered</docSpan>
87          <mark certainty="0.8" resource="babelnet" xml:lang="en">run
                into;
88              be beset by;"The project ran into numerous financial
                    difficulties"</mark>
89        </annotation>
90      </linkGroup>
91      <linkGroup type="alignment">
92        <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-301" endPos="
              doc_en 1.2.11.0-582"/>
93        <docPart doc="doc_fr" beginPos="doc_fr 1.2.11.0-387" endPos="
              doc_fr 1.2.11.0-692"/>
94        <link id="align_tok_102">
95          <docSpan beginPos="doc_en 1.2.11.0-326" endPos="doc_en
                1.2.11.0-329">the</docSpan>
96          <docSpan beginPos="doc_fr 1.2.11.0-419" endPos="doc_fr
                1.2.11.0-421">l'</docSpan>
97        </link>
98        <link id="align_tok_103">
99          <docSpan beginPos="doc_en 1.2.11.0-330" endPos="doc_en
                1.2.11.0-337">trained</docSpan>
100         <docSpan beginPos="doc_fr 1.2.11.0-431" endPos="doc_fr
                1.2.11.0-441">disciplin é</docSpan>
101       </link>
102       <link id="align_tok_104">
103         <docSpan beginPos="doc_en 1.2.11.0-338" endPos="doc_en
                1.2.11.0-346">european</docSpan>
104         <docSpan beginPos="doc_fr 1.2.11.0-422" endPos="doc_fr
                1.2.11.0-430">européen</docSpan>
105       </link>
106       <link id="align_tok_105">
107         <docSpan beginPos="doc_en 1.2.11.0-347" endPos="doc_en
                1.2.11.0-350">who</docSpan>
108         <docSpan beginPos="doc_fr 1.2.11.0-442" endPos="doc_fr
```

```
            1.2.11.0-445">qui</docSpan>
109       </link>
110       <link id="align_tok_106">
111         <docSpan beginPos="doc_en 1.2.11.0-351" endPos="doc_en
              1.2.11.0-357">fought</docSpan>
112         <docSpan beginPos="doc_fr 1.2.11.0-446" endPos="doc_fr
              1.2.11.0-456">combattait</docSpan>
113       </link>
114       <link id="align_tok_107">
115         <docSpan beginPos="doc_en 1.2.11.0-358" endPos="doc_en
              1.2.11.0-360" context="align_chunk_1">at</docSpan>
116         <docSpan beginPos="doc_fr 1.2.11.0-457" endPos="doc_fr
              1.2.11.0-461" context="align_chunk_1">sous</docSpan>
117       </link>
118       <link id="align_tok_108">
119         <docSpan beginPos="doc_en 1.2.11.0-361" endPos="doc_en
              1.2.11.0-364" context="align_chunk_1">his</docSpan>
120         <docSpan beginPos="doc_fr 1.2.11.0-446" endPos="doc_fr
              1.2.11.0-456">combattait</docSpan>
121       </link>
122       <link id="align_tok_109">
123         <docSpan beginPos="doc_en 1.2.11.0-361" endPos="doc_en
              1.2.11.0-364" context="align_chunk_1">his</docSpan>
124         <docSpan beginPos="doc_fr 1.2.11.0-462" endPos="doc_fr
              1.2.11.0-464" context="align_chunk_1">la</docSpan>
125       </link>
126       <link id="align_tok_110">
127         <docSpan beginPos="doc_en 1.2.11.0-365" endPos="doc_en
              1.2.11.0-369" context="align_chunk_1">side</docSpan>
128         <docSpan beginPos="doc_fr 1.2.11.0-465" endPos="doc_fr
              1.2.11.0-469" context="align_chunk_1">même</docSpan>
129       </link>
130       <link id="align_tok_111">
131         <docSpan beginPos="doc_en 1.2.11.0-365" endPos="doc_en
              1.2.11.0-369" context="align_chunk_1">side</docSpan>
132         <docSpan beginPos="doc_fr 1.2.11.0-470" endPos="doc_fr
              1.2.11.0-478" context="align_chunk_1">bannière</docSpan>
133       </link>
134       <link id="align_tok_134">
135         <docSpan beginPos="doc_en 1.2.11.0-502" endPos="doc_en
              1.2.11.0-504" context="align_chunk_3">in</docSpan>
136         <docSpan beginPos="doc_fr 1.2.11.0-610" endPos="doc_fr
              1.2.11.0-612" context="align_chunk_3">en</docSpan>
137       </link>
138       <link id="align_tok_135">
139         <docSpan beginPos="doc_en 1.2.11.0-505" endPos="doc_en
              1.2.11.0-510" context="align_chunk_3">quest</docSpan>
140         <docSpan beginPos="doc_fr 1.2.11.0-613" endPos="doc_fr
              1.2.11.0-622" context="align_chunk_3">cherchant</docSpan
              >
141       </link>
142       <link id="align_tok_136">
```

```
143         <docSpan beginPos="doc_en 1.2.11.0-514" endPos="doc_en
                1.2.11.0-516">an</docSpan>
144         <docSpan beginPos="doc_fr 1.2.11.0-623" endPos="doc_fr
                1.2.11.0-625">l'</docSpan>
145       </link>
146       <link id="align_tok_137">
147         <docSpan beginPos="doc_en 1.2.11.0-517" endPos="doc_en
                1.2.11.0-528">opportunity</docSpan>
148         <docSpan beginPos="doc_fr 1.2.11.0-626" endPos="doc_fr
                1.2.11.0-634">occasion</docSpan>
149       </link>
150     </linkGroup>
151   </linkList>
152   <linkList level="chunk">
153     <linkGroup type="alignment">
154       <docPart doc="doc_en"/>
155       <docPart doc="doc_fr"/>
156       <link id="align_chunk_1" parentID="align_chunk_20">
157         <docSpan beginPos="doc_en 1.2.11.0-358"
158                  endPos="doc_en 1.2.11.0-369"> at his side</docSpan>
159         <docSpan beginPos="doc_fr 1.2.11.0-457"
160                  endPos="doc_fr 1.2.11.0-478">sous la même bannière
                      </docSpan>
161       </link>
162       <link id="align_chunk_3" parentID="align_chunk_20">
163         <docSpan beginPos="doc_en 1.2.11.0-502"
164                  endPos="doc_en 1.2.11.0-513">in quest of</docSpan>
165         <docSpan beginPos="doc_fr 1.2.11.0-610"
166                  endPos="doc_fr 1.2.11.0-622">en cherchant</docSpan>
167       </link>
168       <link id="align_chunk_10" certainty="1" parentID="ROOT">
169         <docSpan beginPos="doc_en 1.2.5.0.0-0" endPos="doc_en
                1.2.5.0.0-6" />
170         <docSpan beginPos="doc_fr 1.2.5.0.0-0" endPos="doc_fr
                1.2.5.0.0-8" />
171       </link>
172       <link id="align_chunk_12" certainty="1" parentID="ROOT">
173         <docSpan beginPos="doc_en 1.2.7.0.0-7" endPos="doc_en
                1.2.7.0.0-25" />
174         <docSpan beginPos="doc_fr 1.2.7.0.0-9" endPos="doc_fr
                1.2.7.0.0-39" />
175       </link>
176       <link id="align_chunk_15" certainty="1" parentID="ROOT">
177         <docSpan beginPos="doc_en 1.2.11.0-122" endPos="doc_en
                1.2.11.0-133" />
178         <docSpan beginPos="doc_fr 1.2.11.0-133" endPos="doc_fr
                1.2.11.0-139" />
179       </link>
180       <link id="align_chunk_40" parentID="align_chunk_10">
181         <docSpan beginPos="doc_en 1.2.11.0-0" endPos="doc_en
                1.2.11.0-2">it</docSpan>
182         <docSpan beginPos="doc_fr 1.2.11.0-0" endPos="doc_fr
```

```
                1.2.11.0-2">c'</docSpan>
183         </link>
184         <link id="align_chunk_41" parentID="align_chunk_10">
185           <docSpan beginPos="doc_en 1.2.11.0-3" endPos="doc_en
                1.2.11.0-6">was</docSpan>
186           <docSpan beginPos="doc_fr 1.2.11.0-3" endPos="doc_fr
                1.2.11.0-8">était</docSpan>
187         </link>
188       </linkGroup>
189       <linkGroup type="annotation">
190         <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-0" endPos="
              doc_en 1.2.11.0-171"/>
191         <annotation id="annot_chunk_2" type="gram">
192           <docSpan beginPos="doc_en 1.2.11.0-0" endPos="doc_en
                1.2.11.0-6" beginTok="s1.t1" endTok="s1.t3">it was</
                docSpan>
193           <mark cat="POS" certainty="1">VB</mark>
194     <mark cat="parse" certainty="1" method="basic-dependencies">
          dep:root:ROOT</mark>
195           <mark cat="parse" certainty="1" method="basic-dependencies">
                gov:nsubj:feature</mark>
196         </annotation>
197     </linkList>
198 </trAnnot>
```