Projet ANR 2012 CORD 015

TRANSREAD

Lecture et interaction bilingues

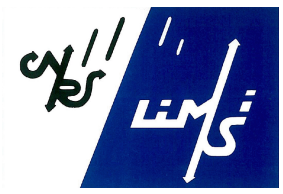
enrichies par les données d'alignement

LIVRABLE 1.1

FORMATS DE REPRÉSENTATION DES ALIGNEMENTS

Octobre 2013

Yong Xu - Guillaume Wisniewski - François Yvon



Résumé

Dans le cadre du projet *Lecture et interaction bilingues enrichies par les données d'alignement* (TransRead, ANR-12-CORD-0015), il est important de concevoir un formalisme de représentation et d'échange pour les documents parallèles alignés qui pourra être utilisé par tous les participants. Dans ce but, nous avons fait une étude des formats pour ces documents.

Dans ce rapport, nous montrons d'abord la nécessité et les exigences d'un format commun pour les participants de TransRead ; ensuite nous étudions certains formats standards utilisés dans le monde académique et industriel ; enfin nous proposons un format conçu à partir du format **cesAlign** [1], qui remplit toutes ces exigences.

1. `http://www.cs.vassar.edu/CES/CES1-5.html`

Formats de représentation des alignements pour TransRead

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

1 Introduction

En linguistique, un **corpus** est un ensemble structuré de textes. Les corpus sont souvent annotés afin qu'ils soient plus utilisables. Un **treebank** est un corpus de textes analysés syntaxiquement et/ou sémantiquement avec les annotations correspondantes. Un corpus peut être monolingue ou multilingue. Un **bitexte** est composé d'un côté d'un texte dans une langue et d'autre côté un texte d'une autre langue, et ces deux textes sont mutuellement en relation de traduction. Un **lien** d'alignement met en relation un groupe d'unités textuelles (par exemple des paragraphes, des phrases ou des mots) d'un côté du bitexte avec un groupe de l'autre côté (souvent on distingue le côté *source* du côté *cible*). Il est possible qu'une unité d'un côté ne soit alignée à rien de l'autre, dans ce cas le lien ne contient qu'un côté, ce qui donne le nom *lien nul*. Un **alignement** est l'ensemble des liens entre les deux textes. On se reportera par exemple à [Véronis, 2000, Melamed, 2001, Tiedemann, 2011] pour une présentation des méthodes pour construire et utiliser des alignements.

L'objectif du projet TransRead est d'étudier de nouvelles applications multilingues destinées à faciliter la consultation de documents dans plusieurs langues par des utilisateurs imparfaitement bilingues, pour qui des bitextes et, lorsqu'ils sont disponibles, des alignements sont des ressources de valeur. À l'inverse des approches « boîte noire » en traduction, qui ciblent un public monolingue, TransRead s'intéresse donc en premier lieu à la visualisation de textes bilingues et des alignements qui les lient.

Le domaine du projet étant principalement le traitement de documents, il est nécessaire que tous les participants emploient le même format d'annotation. D'une part, le système de visualisation des documents et celui qui calcule les alignements doivent avoir exactement la même notion de position sur chaque unité textuelle ; d'autre part, un format commun prédéfini facilite les développements individuels des programmes et logiciels.

Du point de vue de l'alignement, l'exigence pour ce format est que nous puissions représenter les alignements de tous les niveaux : non seulement entre des unités classiques comme les phrases ou les mots, mais aussi des segments des mots, des unités grammaticales, etc. Par ailleurs, d'autres types des données utiles à aider la compréhension des textes, comme la désambiguïsation du sens des mots, doivent aussi être représentées. Au final, il s'avère que nous devons avoir un mécanisme pour identifier uniquement chaque entité lexicale (une entité lexicale étant une chaine de caractères qui correspond à un symbole, qui est généralement un mot) des textes alignés. Avec ce mécanisme, la représentation des liens entre les entités et ses informations devient facile.

Du point de vue de la visualisation, le problème du format est plus compliqué. Dans la mesure où le calcul d'alignements traite principalement des textes bruts (sans indication de format ou de mise en page), les formats standards du domaine ne prennent généralement pas en compte les informations de présentation (les fontes, la mise en page, etc). Il est donc difficile (voire impossible) de développer des applications de lecture bilingue (sur des applications Web ou sur des terminaux mobiles comme des liseuses) à partir d'eux. En conséquence, nous avons décidé d'encoder les documents originaux dans le format du EPUB (Electronic PUBlication) [1], afin de faciliter les développements des modules de visualisation.

Enfin, du point de vue de la valorisation du projet, les applications de TransRead doivent être capables de fournir une bonne expérience utilisateur. Les plus décisifs sont la précision des résultats et la vitesse de réaction du système. Le format des fichiers est un facteur important pour la vitesse de réaction. Donc il faut que le format permette d'effectuer les requêtes efficaces sur des annotations disponibles dans les fichiers d'alignement. Compte-tenu de la capacité limitée de calcul des terminaux mobiles, la représentation des alignements doit être claire, complète, bien structurée, et permettre de les récupérer sans calculs lourds. Ces exigences nous orientent vers un format basé sur le standard XML.

2 Des formats standards de la traduction

Dans cette partie, nous étudions des formats du domaine de la traduction, en nous focalisant sur l'alignement de phrases et l'alignement de mots, qui sont les unités d'alignement les plus étudiées. À partir de cette étude, il sera facile de représenter les liens aux autres niveaux. La traduction étant un domaine très dynamique à la fois dans le monde académique et dans l'industrie, plusieurs formats sont largement diffusés dans la communauté. Nous introduisons ceux qui sont considérés comme les standards et les analysons, dans le but de mettre en évidence leurs qualités et défauts par rapport aux exigences de TransRead.

On remarque qu'il existe plusieurs formats pour représenter les treebanks : Penn Treebank [2], Susanne [3], TIGER-XML [4], etc. Ces formats, n'étant pas conçus pour l'alignement, peuvent nous aider à développer la partie des annotations linguistiques pour le format de TransRead.

2.1 Des formats pour l'alignement de phrases

La phrase est probablement l'unité textuelle la plus souvent traitée dans l'industrie de la localisation et de l'internationalisation. Elle est aussi beaucoup étudiée dans les applications de traitement de bitextes. En traduction statistique, l'alignement de phrases est le point de départ de la chaine de traitements. Il est donc normal qu'un riche ensemble de formats existe pour représenter les traitements de phrases.

2.1.1 Le format de Uplug

Uplug [5] [Tiedemann, 2003] est un logiciel dédié à la recherche de l'alignement de textes. Le format qu'il propose est basé sur le *Corpus Encoding Standard for XML* (XCES) [6]. Es-

1. http ://www.idpf.org/epub/30/spec/epub30-overview.html
2. http ://www.cis.upenn.edu/ treebank/
3. http ://www.grsampson.net/SueDoc.html
4. http ://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html
5. `http://stp.lingfil.uu.se/~joerg/Uplug/home.html`
6. `http://www.xces.org/`

sentiellement quatre balises, `<text>`, `<p>`, `<s>` et `<w>` sont employées pour repérer respectivement un document, un paragraphe, une phrase et un mot dans les documents constituant le bitexte. Les éléments `<p>`, `<s>` et `<w>` contiennent l'attribut d'identifiant (*id*) en respectant la hiérarchie linguistique classique. Par exemple, le premier paragraphe possède l'*id* "`1`", la première phrase du paragraphe "`1.1`", et le premier mot de la phrase "`1.1.1`". Donc tous les mots possèdent un *id* (unique au sein d'un document source ou cible) qui indique leur position dans le document. En ce qui concerne la représentation des alignements de phrases, un fichier tiers est créé dans lequel la balise `<cesAlign>` est utilisée pour décrire un alignement et la balise `<link>` pour décrire les liens le constituant. La balise `<cesAlign>` inclut les attributs "`fromDoc`" et "`toDoc`" indiquant les documents respectivement source et cible. Chaque `<link>` possède un attribut "`xtargets`" désignant les identifiants des unités mises en relation par ce lien, un attribut "`id`" qui est l'identifiant unique du lien et un attribut "`certainty`" une valeur donnant la confiance selon une certaine mesure. Il est possible d'ajouter d'autres attributs au besoin.

Les figures 1 et 2 représentent deux documents originaux et 3 est le fichier d'alignement correspondant. Dans le listing 3, l'attribut "`type`" de l'élément `<cesAlign>` a pour valeur "`sent`", indiquant que ce groupe de liens décrit des relations entre phrases. Dans chaque élément `<link>`, la valeur de l'attribut "`xtargets`" est composée d'une série d'identifants sources et une série d'identifiants cibles. Les deux séries sont séparées par un point-virgule, et les identifiants dans une série sont séparés par un espace. Ainsi, un "`xtargets`" dont la valeur ne contient pas de point-virgule représente un lien nul.

Listing 1 – Le document source

```
<?xml version="1.0" encoding="utf-8"?>
<text>
  <p id="1">
    <s id="1.1">
      <w id="1.1.1">DE</w>
      <w id="1.1.2">LA</w>
      <w id="1.1.3">TERRE</w>
      <w id="1.1.4">A</w>
      <w id="1.1.5">LA</w>
      <w id="1.1.6">LUNE</w>
    </s>
    <s id="1.2">
      <w id="1.2.1">Trajet</w>
      <w id="1.2.2">Direct</w>
      <w id="1.2.3">en</w>
      <w id="1.2.4">97</w>
      <w id="1.2.5">Heures</w>
      <w id="1.2.6">20</w>
      <w id="1.2.7">Minutes</w>
    </s>
    <s id="1.3">
      <w id="1.3.1">par</w>
      <w id="1.3.2">Jules</w>
      <w id="1.3.3">Verne</w>
```

```
    </s>
    <s id="1.4">
      <w id="1.4.1">I</w>
    </s>
  </p>
</text>
```

Listing 2 – Le document cible

```
<?xml version="1.0" encoding="utf-8"?>
<text>
  <p id="1">
    <s id="1.1">
      <w id="1.1.1">FROM</w>
      <w id="1.1.2">THE</w>
      <w id="1.1.3">EARTH</w>
      <w id="1.1.4">TO</w>
      <w id="1.1.5">THE</w>
      <w id="1.1.6">MOON</w>
    </s>
    <s id="1.2">
      <w id="1.2.1">CHAPTER</w>
      <w id="1.2.2">I</w>
    </s>
  </p>
</text>
```

Listing 3 – Le fichier d'alignement

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Doctype cesAlign  PUBLIC "-//CES//DTD cesAlign//EN" -->
<cesAlign fromDoc="xml/fr.xml" toDoc="xml/en.xml" type="sent">
  <linkList>
    <linkGrp>
      <link certainty="1" id="SL2" xtargets="1.1;1.1" />
      <link certainty="1" id="SL3" xtargets="1.2 1.3;" />
      <link certainty="1" id="SL4" xtargets="1.4;1.2" />
    </linkGrp>
  </linkList>
</cesAlign>
```

Une représentation également dérivée des propositions de XCES a été adopté pour le projet Européen PANACEA [7] ; elle utilise une version « modernisée » du format de UPLUG [8].

7. http://www.panacea-lr.eu/
8. Voir en particulier http://www.panacea-lr.eu/system/xcesXSD/T01-documentation-v1.pdf

Dans la mesure où notre proposition, développée plus bas, repose sur des principes relativement proches, la conversion depuis et vers le format de PANACEA ne semble pas poser de difficultés majeures. Notons enfin qu'une variante de ce format est également utilisée dans le projet OpenCorp, qui inclut un outil pour visualiser les alignements de phrases [9].

2.1.2 Le format XLIFF

XLIFF [10] (XML Localization Interchange File Format) est un format standard dans l'industrie de localisation Il est proposé par l'OASIS (*Organization for the Advancement of Structured Information Standards*) [11] afin de standardiser les échanges des données de localisation entre les outils. Même si cela signifie que les documents XLIFF sont souvent des intermédiaires dans le processus de localisation, leur structure peut être une source d'inspiration pour la tâche d'alignement.

Dans le processus de localisation, la première étape consiste à séparer les parties visibles d'un document des données de mise en page représentées, par exemples, par des balises. Les données de mise en page sont stockées dans un fichier squelette, dans lequel une marque spéciale est attribuée à chaque partie visible. Le document représenté Figure 4 est un document original, et 5 le squelette correspondant.

Listing 4 – Exemple.html, un document HTML contennant 2 phrases visibles

```
<html>
  <head>
    <title>Un titre</title>
  </head>
  <body>
    <p>Un paragraphe</p>
  </body>
</html>
```

Listing 5 – Exemple.skl, le fichier squelette correspondant

```
<html>
   <head>
      <title>%%%1%%%</title>
   </head>
   <body>
      <p>%%%2%%%</p>
   </body>
 </html>
```

9. http://wanthalf.saga.cz/doc_intertext
10. La dernière version de XLIFF (1.2) est décrite dans le document : http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html
11. https://www.oasis-open.org

Les textes visibles sont segmentés avant d'être mis dans un document XML au format XLIFF, qui est essentiellement une liste d'éléments `<trans-unit>`, chacun contenant une unité de traduction. En pratique, une unité est souvent une phrase. Un élément `<trans-unit>` est composé d'un sous-élément `<source>` qui contient une phrase source, et d'un sous-élément `<target>` qui contient la traduction proposée. En plus de ces deux éléments, il peut y avoir un nombre non limité de sous-éléments `<alt-trans>` qui contiendront chacun une alternative de traduction, par exemple des versions anciennes ou les traductions dans d'autres langues (même si XLIFF est principalement conçu pour une seule paire de langues), sachant que tous ces éléments peuvent spécifier la langue de leur contenu par l'attribut "`xml:lang`".

Chaque élément `<trans-unit>` doit avoir un `id` correspondant à la marque spéciale dans le fichier squelette afin d'établir le resultat final de la localisation. Selon le même principe, le fichier squelette doit être indiqué dans le document XLIFF. Le listing 6 représente un document XLIFF correspondant aux exemples 4 et 5.

Listing 6 – Le document XLIFF

```
<? xml version="1.0" ?>
<xliff version="1.0">
  <file original="Exemple.html"
        source-language="fr"
        datatype="HTML Page">
    <header>
      <skl>
        <external-file href="Exemple.skl"/>
      </skl>
    </header>
    <body>
      <trans-unit id="%%%1%%%">
        <source xml:lang="fr">Un titre</source>
        <target xml:lang="en">A title</target>
      </trans-unit>
      <trans-unit id="%%%2%%%">
        <source xml:lang="fr">Un paragraphe</source>
        <target xml:lang="en">A paragraph</target>
        <alt-trans xml:lang="en">One paragraph</alt-trans>
      </trans-unit>
    </body>
  </file>
</xliff>
```

2.1.3 Le format TMX

TMX (Translation Memory eXchange) [12] est un autre format standard dans la localisation. Il est développé par le LISA (*Localisation Industry Standards Association*) comme un

12. `http://www.gala-global.org/oscarStandards/tmx/tmx14b.html`

format commun pour les bases de données des mémoires de traduction (TM, Translation Memory).

En général, les spécifications de TMX ressemblent beaucoup à celles de XLIFF. Il s'agit d'une représentation XML dans laquelle les unités de traduction sont entourées par la balise `<tu>`. Cette balise contient plusieurs sous-éléments `<tuv>`, chacun contenant un sous-élément `<seg>`, dont le contenu est un segment du texte. Nous pouvons choisir, pour chaque `<tu>`, le type de segment en utilisant l'attribut "`segtype`", dont quatre valeurs sont possibles : "`block`", "`paragraph`", "`sentence`", et "`phrase`". En pratique, un segment est souvent une phrase. Les différences entre TMX et XLIFF viennent principalement de leurs différentes utilités : dans le format TMX, l'ordre des unités n'est pas pertinente.

Le listing 7 représente un document au format TMX.

Listing 7 – Un document TMX

```
<?xml version="1.0" ?>
<tmx version="1.4">
 <header datatype="PlainText" segtype="sentence"
  adminlang="en-us" srclang="EN" o-tmf="ABCTransMem">
 </header>
 <body>
  <tu>
   <tuv xml:lang="EN">
    <seg>Text <bpt i="1">&lt;B&gt;</bpt>bold<ept i="1">&lt;/B&gt;</ept></seg>
   </tuv>
   <tuv xml:lang="FR">
    <seg>Texte <bpt i="1">&lt;B&gt;</bpt>gras<ept i="1">&lt;/B&gt;</ept></seg>
   </tuv>
  </tu>
 </body>
</tmx>
```

2.1.4 L'analyse des formats standards

Nous avons brièvement présenté, dans les sections précédentes, trois formats standardisés d'alignement. XCES propose une annotation à la fois pour la représentation des documents originaux et l'alignement, tandis que XLIFF et TMX représentent principalement des alignements. Du point de vue de TransRead, XCES et TMX ne sont pas directement utilisables, puisque :

1. L'encodage des documents originaux de XCES (donc celui de Uplug) n'est pas suffisamment fin. Bien que le mécanisme d'annotation de XCES soit théoriquement capable d'encoder tous les objets dans les documents (les textes, les images, les tableaux, etc), il ne fournit pas de moyens pour gérer les informations de présentation, concernant tant la mise en forme (la police, la disposition, etc.), que la mise en page des textes. Cela pose un problème pour la visualisation et l'interaction humaine-machine avec les bitextes. L'objectif de TransRead étant de faciliter la consultation de documents existant en plusieurs langues, ce défaut devient rédhibitoire.

2. TMX ne permet pas d'établir des correspondances entre les documents originaux et le fichier d'alignement. Si nous pouvons parfaitement stocker tous les liens trouvés dans un fichier TMX, il ne fournit aucun moyen pour repositionner les segments extraits dans les document originaux, ce qui rend leur visualisation impossible.

En revanche, XLIFF et la partie représentant les alignements de XCES peuvent tous les deux servir de point de départ pour la représentation d'alignements adaptée aux besoins du projet TransRead. XCES propose un jeu de balise très complet et adapté à l'alignement dans sa DTD **cesAlign** [13]. XLIFF dispose d'une remarquable extensibilité pour les balises et les attributs, qui rend l'encodage des autres niveaux possible. Le format proposé sera donc basé sur ces deux formats.

2.2 Des formats pour l'alignement de mots

L'alignement de mots est un sujet rarement traité dans les applications industrielles (sinon pour des systèmes de traduction automatique). Il est, au contraire, très important dans les travaux de recherche en traduction automatique. En conséquence, la plupart des formats d'alignement de mots ont été initialement proposés dans la communauté scientifique. La figure 1 est une illustration en matrice d'un alignement de mots d'une paire de phrases simples.

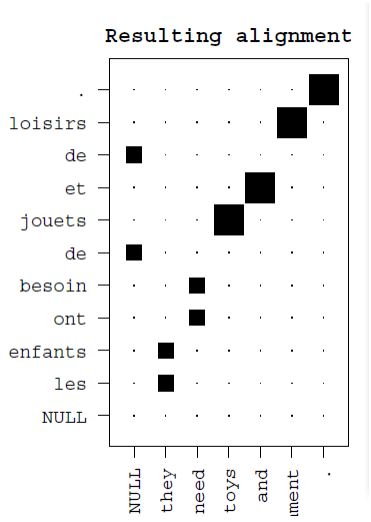


Figure 1 – Un exemple de l'alignement de mots, représenté en matrice

L'alignement de mots étant un problème de recherche en pleine évolution, il n'existe pas, à notre connaissance, de format standardisé. Nous allons néanmoins introduire plusieurs formats populaires dans la communauté, afin de décrire quelques problèmes et conventions de l'alignement de mots.

13. `http://www.cs.vassar.edu/CES/sgml/cesAlign.dtd`

2.2.1 L'évaluation de 2003

En 2003, a eu lieu un atelier d'alignement de mots lors de la conférence NAACL (North American Chapter of the Association for Computational Linguistics) [Mihalcea and Pedersen, 2003]. Les participants étaient invités à aligner les mots dans des paires de phrases parallèles. Chaque paire de phrases parallèles est identifiée par un `id`. L'alignement des tokens d'une paire (L_e, L_f) est représenté par plusieurs lignes dans un fichiers selon le format :

`id_paire_phrase`	`pos_dans_`L_e	`pos_dans_`L_f	`[S` *or* `P]`	`[score]`

où `id_paire_phrase` est l'identifiant de la paire de phrases, `pos_dans_`L_e (resp. L_f) est la position d'un mot de L_e, (resp. L_f). `[S` *or* `P]` indique si ce lien est *sûr* ou seulement *possible*, `[score]` un score de confiance de ce lien. Les deux derniers champs sont facultatifs.

Dans ce format, tous les symboles séparés par un espace dans les textes parallèles sont considérés comme des mots et sont donc alignables ; y compris les ponctuations. Tous les mots doivent être alignés, ceux qui n'ont pas de correspondances doivent être liés au mot supplémentaire `NULL` ajouté au début de chaque phrase, à la position conventionnelle 0. Un mot d'un coté peut être aligné à un ou plusieurs mots de l'autre coté. Cela permet de représenter les liens entre les groupes de mots.

Ce format est inspiré de celui d'alignement de mots utilisé dans le corpus Blinker [14] dont les principes d'annotation sont expliqués dans [Melamed, 1998]. Il a été utilisé dans plusieurs projets de recherche, par exemple le corpus décrit dans [de Almeida Varelas Graça et al., 2008], décrivant des alignements dans six langues [15]. Un trait caractéristique de ce format, qui a été repris dans de nombreuses expériences ultérieures, est la distinction entre les liens surs et possibles. Cette distinction s'est progressivement imposée pour distinguer les configurations dans lesquels l'alignement est sans ambiguïté des cas où la traduction est non littérale ou bien non compositionnelle, ou bien encore les cas où des mots ne sont pas traduits [16].

Le listing 8 est l'alignement de mots d'une paire de phrases ayant l'*id* 18.

Listing 8 – Un alignement de mots

```
#<s snum=18> They had gone . </s>
#<s snum=18> Ils étaient allés . </s>
18 1 1 1
18 2 2 P 0.7
18 3 3 S
18 4 4 S 1
```

On note sur cet exemple le lien de type `[P]` entre *have* et *sont*, qui, bien que l'un ne soit pas la traduction de l'autre, marque ici le fait qu'ils occupent la même fonction d'auxiliaire.

14. `http://nlp.cs.nyu.edu/blinker/index.html`
15. `https://www.l2f.inesc-id.pt/wiki/index.php/Word_Alignments`
16. Cette distinction est introduite dans [Och and Ney, 2003] avec la définition suivante (p. 33) :

    (...) an S (sure) alignment, for alignments that are un- ambiguous, and a P (possible) alignment, for ambiguous alignments. The P label is used especially to align words within idiomatic expressions and free translations and missing function words (...)

2.2.2 Le format de Moses

Moses [Koehn et al., 2007] est un logiciel très utilisé dans la communauté de la traduction automatique. Il inclut un paquet GIZA++, décrit dans [Och and Ney, 2003] dédié à la prédiction des alignement de mots, qui est le point de départ de la plupart des activités de la recherche du domaine. L'alignement de mots d'une phrase parallèle est représenté par trois lignes dans le fichier de résultat de Moses : la première pour indiquer l'*id* de la paire de phrases ; la deuxième est la phrase source ; la troisième ligne est la plus critique : un mot supplémentaire "NULL" est ajouté au début de la phrase cible, chaque mot cible est suivi par la liste des mots sources correspondants. Le listing 9 est un alignement de mots représenté dans ce format .

Listing 9 – Un alignement de mots dans Moses

```
# Paire de phrase 18
He meets her with pleasure .
NULL ({}) Avec ({4}) plaisir ({5}) ,({}) il ({1}) la ({3}) voit ({2}) . ({6})
```

2.2.3 Le format du *Alignment Set Toolkit*

Alignment Set Toolkit [17] est un logiciel destiné à traiter les alignements manuels établis en respectant les principes décrits dans [Lambert et al., 2005]. Trois fichiers séparés sont utilisés pour représenter une tâche d'alignement : un pour les phrases sources, un pour les phrases cibles, et un troisième pour les alignements. Dans chaque fichier, chaque ligne correspond à une paire de phrases parallèles, et l'ordre des paires de phrases est le même dans les trois fichiers. Ainsi, pour une paire de phrases parallèles, la phrase source, la cible, et les liens d'alignement relatifs à la paire de phrase ont le même indice de ligne. Un lien est représenté par l'indice d'un mot source et l'indice du mot cible aligné, separé par un "s" ou un tiret pour les liens sûrs, ou un "p" pour les liens non sûrs. Les indices commencent par 1, car 0 est utilisé pour "NULL". Le listing 10 est un alignement de mots représenté dans ce format.

Listing 10 – Un alignement de mots dans le format de Alignment Set Tookkit

```
Le document source
    I can not say anything at this stage .
Le document cible
    En ce moment , je ne peux rien dire .
Le fichier d'alignement source-cible
    0-4 1-5 2-7 3-6 4-9 5p8 6-1 7-2 8-3 9-10
```

2.2.4 Le format du projet SMULTRON

SMULTRON (The Stockholm MULtilingual parallel TReebank) est un projet d'annotation linguistique réalisé à l'Université de Stockholm et l'Université de Zurich [Volk

17. http ://www-lium.univ-lemans.fr/ lambert/software/AlignmentSet.html

et al., 2010], qui vise à produire des alignements sous-phrastiques entre structures linguistiques. La banque d'arbres représentant d'un document original est stockée dans un fichier XML respectant le format TIGER-XML. En général, une phrase est décrite par un élément `<s>` ayant l'attribut "`id`" et deux sous-éléments `<termninals>` et `<nonterminals>`. Le `<nonterminals>` contient plusieurs sous-éléments `<nt>` encodant les différents constituants. Le `<termninals>` contient une liste ordonnée de sous-éléments `<t>`, chacun correspondant à un mot de la phrase. Les attributs d'un `<t>` ne sont pas spécifiés. Normalement le groupe d'attributs contient "`id`", "`word`" dont la valeur est le mot, "`pos`" qui indique la catégorie grammaticale du mot, "`lemma`" pour le lemme, etc.

Le format d'alignement proposé par SMULTRON est également basé sur XML. L'alignement est représenté dans un fichier tiers. Deux éléments `<treebank>` sont présents pour indiquer les deux *treebanks* qui sont alignées. Ils possèdent trois attributs : "`id`", "`language`" et "`filename`". L'élément `<alignments>` regroupe tous les `<align>`. Un `<align>` possède comme attribut "`type`", dont les valeurs possibles sont "`good`" et "`fuzzy`" pour indiquer respectivement les liens sûrs et les approximatifs. Il a deux sous-éléments `<node>`, chacun encodant un mot d'un *treebank*. Un `<node>` a deux attributs : "`treebank_id`" qui doit se référer à l'`id` d'un *treebank*, et "`node_id`" dont la valeur doit être une référence à l'`id` d'un mot dans un fichier de *treebank*.

Les listings 11, 12 et 13, extraits du corpus « Sophie's world » délivré par SMULTRON, illustrent l'utilisation de ce format. Ci-dessous, le fichier *smultron_en_sophie.xml* est le *treebank* de la version anglaise, et *smultron_sv_sophie.xml* celui de la version suédoise.

Listing 11 – Un extrait de smultron_en_sophie.xml

```
<?xml version="1.0"?>
<corpus>
  <head>...</head>
  <body>
    ...
    <s id="s2">
      <graph root="s2_508">
        <terminals>
          <t id="s2_1" word="..." pos=":" morph="--"/>
          <t id="s2_2" word="at" pos="IN" morph="--"/>
          <t id="s2_3" word="some" pos="DT" morph="--"/>
          <t id="s2_4" word="point" pos="NN" morph="--"/>
          <t id="s2_5" word="something" pos="NN" morph="--"/>
          <t id="s2_6" word="must" pos="MD" morph="--"/>
          <t id="s2_7" word="have" pos="VB" morph="--"/>
          <t id="s2_8" word="come" pos="VBN" morph="--"/>
          <t id="s2_9" word="from" pos="IN" morph="--"/>
          <t id="s2_10" word="nothing" pos="NN" morph="--"/>
          <t id="s2_11" word="..." pos=":" morph="--"/>
        </terminals>
        <nonterminals>
          <nt id="s2_500" cat="NP">
            <edge label="--" idref="s2_3"/>
            <edge label="--" idref="s2_4"/>
```

```
            </nt>
            <nt id="s2_501" cat="NP">
              <edge label="--" idref="s2_5"/>
            </nt>
            <nt id="s2_502" cat="NP">
              <edge label="--" idref="s2_10"/>
            </nt>
            <nt id="s2_503" cat="PP">
              <edge label="--" idref="s2_2"/>
              <edge label="--" idref="s2_500"/>
            </nt>
            <nt id="s2_504" cat="PP">
              <edge label="--" idref="s2_9"/>
              <edge label="--" idref="s2_502"/>
            </nt>
            <nt id="s2_505" cat="VP">
              <edge label="--" idref="s2_8"/>
              <edge label="CLR" idref="s2_504"/>
            </nt>
            <nt id="s2_506" cat="VP">
              <edge label="--" idref="s2_7"/>
              <edge label="--" idref="s2_505"/>
            </nt>
            <nt id="s2_507" cat="VP">
              <edge label="--" idref="s2_6"/>
              <edge label="--" idref="s2_506"/>
            </nt>
            <nt id="s2_508" cat="S">
              <edge label="--" idref="s2_1"/>
              <edge label="--" idref="s2_11"/>
              <edge label="SBJ" idref="s2_501"/>
              <edge label="TMP" idref="s2_503"/>
              <edge label="--" idref="s2_507"/>
            </nt>
          </nonterminals>
        </graph>
      </s>
      ...
    </body>
  </corpus>
```

Listing 12 – Un extrait de smultron_sv_sophie.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <head>...</head>
```

```
<body>
  ...
  <s id="s2">
    <graph root="s2_506">
      <terminals>
        <t id="s2_1" word="..." pos="DL" morph="--" lemma="--" type="--"/>
        <t id="s2_2" word="en" pos="DT" morph="--" lemma="en" type="--"/>
        <t id="s2_3" word="gång" pos="NN" morph="UTR" lemma="gång" type="--"/>
        <t id="s2_4" word="i" pos="PR" morph="--" lemma="i" type="--"/>
        <t id="s2_5" word="tiden" pos="NN" morph="UTR" lemma="tid" type="--"/>
        <t id="s2_6" word="måste" pos="VBFIN" morph="--" lemma="måste" type="--"/>
        <t id="s2_7" word="ändå" pos="AB" morph="--" lemma="ändå" type="--"/>
        <t id="s2_8" word="allting" pos="PN" morph="--" lemma="allting" type="--"/>
        <t id="s2_9" word="ha" pos="VBINF" morph="--" lemma="ha" type="--"/>
        <t id="s2_10" word="blivit" pos="VBSUP" morph="--" lemma="bliva" type="--"/>
        <t id="s2_11" word="till" pos="PL" morph="--" lemma="till" type="--"/>
        <t id="s2_12" word="av" pos="PR" morph="--" lemma="av" type="--"/>
        <t id="s2_13" word="noll" pos="RG" morph="--" lemma="noll" type="--"/>
        <t id="s2_14" word="och" pos="KN" morph="--" lemma="och" type="--"/>
        <t id="s2_15" word="ingenting" pos="PN" morph="--" lemma="ingenting" type="--"/>
        <t id="s2_16" word="." pos="DL" morph="--" lemma="--" type="--"/>
      </terminals>
      <nonterminals>
        <nt id="s2_500" cat="PP">
          <edge label="HD" idref="s2_4"/>
          <edge label="NK" idref="s2_510"/>
        </nt>
        <nt id="s2_501" cat="CNP">
          <edge label="CJ" idref="s2_13"/>
          <edge label="CD" idref="s2_14"/>
          <edge label="CJ" idref="s2_507"/>
        </nt>
        <nt id="s2_502" cat="NP">
          <edge label="NK" idref="s2_2"/>
          <edge label="HD" idref="s2_3"/>
          <edge label="MNR" idref="s2_500"/>
        </nt>
        <nt id="s2_503" cat="PP">
          <edge label="HD" idref="s2_12"/>
          <edge label="NK" idref="s2_501"/>
        </nt>
        <nt id="s2_504" cat="VP">
          <edge label="HD" idref="s2_10"/>
          <edge label="SVP" idref="s2_11"/>
          <edge label="MO" idref="s2_503"/>
        </nt>
        <nt id="s2_505" cat="VP">
          <edge label="HD" idref="s2_9"/>
```

```
            <edge label="OC" idref="s2_504"/>
          </nt>
          <nt id="s2_506" cat="S">
            <edge label="HD" idref="s2_6"/>
            <edge label="MO" idref="s2_508"/>
            <edge label="SB" idref="s2_509"/>
            <edge label="MO" idref="s2_502"/>
            <edge label="OC" idref="s2_505"/>
          </nt>
          <nt id="s2_507" cat="NP">
            <edge label="HD" idref="s2_15"/>
          </nt>
          <nt id="s2_508" cat="AVP">
            <edge label="HD" idref="s2_7"/>
          </nt>
          <nt id="s2_509" cat="NP">
            <edge label="HD" idref="s2_8"/>
          </nt>
          <nt id="s2_510" cat="NP">
            <edge label="HD" idref="s2_5"/>
          </nt>
        </nonterminals>
      </graph>
    </s>
    ...
  </body>
</corpus>
```

Listing 13 – Un extrait de l'alignement pour Sophie's world

```
<?xml version="1.0" encoding="UTF-8"?>
<treealign subversion="3" version="2">
  <head>
    ...
    <treebanks>
      <treebank id="en" language="en_US" filename="smultron\_en\_sophie.xml"/>
      <treebank id="sv" language="sv_SE" filename="smultron\_sv\_sophie.xml"/>
    </treebanks>
    ...
  </head>
  <alignments>
    ...
    <align type="good">
      <node treebank_id="en" node_id="s2_6"/>
      <node treebank_id="sv" node_id="s2_6"/>
    </align>
```

```
    <align type="good">
      <node treebank_id="en" node_id="s2_7"/>
      <node treebank_id="sv" node_id="s2_9"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_8"/>
      <node treebank_id="sv" node_id="s2_10"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_8"/>
      <node treebank_id="sv" node_id="s2_11"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_9"/>
      <node treebank_id="sv" node_id="s2_12"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_10"/>
      <node treebank_id="sv" node_id="s2_15"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_502"/>
      <node treebank_id="sv" node_id="s2_507"/>
    </align>
    <align type="fuzzy">
      <node treebank_id="en" node_id="s2_503"/>
      <node treebank_id="sv" node_id="s2_502"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_504"/>
      <node treebank_id="sv" node_id="s2_503"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_505"/>
      <node treebank_id="sv" node_id="s2_504"/>
    </align>
    <align type="good">
      <node treebank_id="en" node_id="s2_506"/>
      <node treebank_id="sv" node_id="s2_505"/>
    </align>
    <align type="fuzzy">
      <node treebank_id="en" node_id="s2_508"/>
      <node treebank_id="sv" node_id="s2_506"/>
    </align>
    ...
  </alignments>
</treealign>
```

2.2.5 L'analyse des formats

Ces formats sont actuellement très utilisés dans la communauté de la traduction automatique. Le format de l'atelier 2003 permet essentiellement de représenter des éléments désirés (identifiant des mots, scores de confiance, etc). Le format de SMULTRON fournit une spécification plus complète qui permet de représenter les informations linguistiques et des alignements entre syntagmes ou plus généralement segments de mots. Il est d'ailleurs plus standardisé et extensible. Nous allons concevoir un format basé aussi sur XML pour l'alignement, en nous appuyant principalement sur ces deux propositions.

3 Le format proposé

En nous inspirant de la DTD d'annotation d'alignements **cesAlign**, nous proposons un format basé sur le standard XML. Afin de faciliter la visualisation des textes, nous avons adopté la stratégie d'enregistrer les annotations dans un document tiers, au lieu de les fusionner avec les documents originaux. La seule contrainte pour ces derniers est qu'ils soient représentés au format XHTML, sans autre restriction particulière. Les correspondances entre les documents originaux et le document d'annotation sont établies à l'aide d'un mécanisme d'adressage qui généralise celui de **cesAlign**. Dans cette section, nous allons d'abord présenter le mécanisme d'adressage pour les textes des documents originaux, ensuite décrire progressivement la proposition de format.

La DTD précise et le schéma XML sont présentés dans l'annexe 5.1 et 5.2.

3.1 L'adressage

Dans la section 1, nous avons noté la nécessité d'associer chaque unité textuelle à un identifiant unique. Idéalement, cette identification peut être faite en attribuant un attribut `"ID"` à chacun des éléments du texte. Cependant, cette méthode n'est pas toujours possible parce que cet attribut est le plus souvent absent dans des documents, et qu'il n'est souvent pas désirable (voire parfois impossible [18]) de les rajouter. Nous avons donc décidé d'identifier chaque unité par le chemin de la racine vers cette unité dans l'arbre DOM (Document Object Model [19]) du document. Le DOM représente un document XHTML comme un arbre généalogique dont la racine supérieure est toujours un nœud `<document>`, qui a pour fils le nœud `<html>`. Toutes les balises comprises dans le document XHTML sont considérées comme des nœuds. Il faut en particulier remarquer que le contenu texte de chaque balise est également considéré comme un nœud de type texte nommé `<text>`.

Le mécanisme d'adressage pour les documents XHTML ne peut être conçu au niveau de mots comme dans le format de Moses, parce que la tokenisation est plus compliquée. Dans les fichiers XHTML, il est possible qu'un mot soit décomposé dans plusieurs éléments. Le tokeniseur ne peut donc considérer les balises ou les espaces comme les frontières des mots. En conséquence, nous ne pouvons pas utiliser la notion d'indice pour des mots dans des éléments. L'adressage est donc fait au niveau des caractères.

Nous définissons le chemin d'un caractère comme étant composé par trois parties : un identifiant du document, la position de l'élément le contenant dans le document (donc toujours un nœud `<text>`), et la position relative du caractère dans cet élément `<text>`. Tous

18. En toute généralité, les mots des documents sources et cibles peuvent contenir un nombre arbitraire de balises de mise en forme.

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

les indices positionnels commencent par la valeur 0. La position de l'élément contenant ce caractère est déterminée dans l'arbre DOM du document de façon descendante : à partir du `<document>`, on traverse le chemin vers le nœud en question en prennant l'indice de chaque nœud parmi ses frères, ce qui donne une série de nombres. Cette série est séparée par le point. La position de l'élément et la position relative du caractère dans cet élément sont combinées par un tiret. 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 est séparés par une espace. Donc pour l'exemple ci-dessus, si l'*id* du document est `"doc"`, alors la position totale du caractère en question est `"doc 1.2-3"`. Avec cette définition, nous pouvons calculer le chemin d'une entité en indiquant la position de son premier et sont dernier caractère dans le document. En conséquence, un chemin comprend toujours deux positions. Pour les unités non-textuelles, nous mettons la valeur 0 pour la troisième partie des deux positions. Ci-dessous, nous donnons un exemple pour un vrai document XHTML.

Listing 14 – Un fichier XHTML

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE html PUBLIC '-//W3C//DTD XHTML 1.1//EN'
'http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd'>

<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>le titre</title>
    <meta content="text/css" http-equiv="Content-Style-Type" />
    <link href="pgepub.css" rel="stylesheet" type="text/css" />
  </head>

  <body>
    <p>L'exemple est fait par M<sup>me.</sup> XXX.</p>
  </body>
</html>
```

Donc, pour la page très simple représentée Figure 14 ayant pour identifiant `"ex_doc"`, nous allons montrer l'adressage réalisé pour le mot "exemple". Il faut d'abord déterminer la position du premier caractère "e". La racine `<document>` a deux nœuds fils `<documenttype>` et `<html>`. Le nœud `<html>` possède cinq fils : un `<text>` pour l'espace entre la balise `<html>` et la balise `<head>`, `<head>`, `<text>` pour l'espace entre `<head>` et `<body>`, `<body>`, `<text>` pour l'espace entre `<body>` et `<html>`. Ainsi nous pouvons déduire que la première partie du chemin du premier caractère (qui est entourné par un nœud `<text>`) est : `"1.3.1.0"` puisque `<html>` est le deuxième fils du nœud `<document>` ; `<body>` le quatrième fils du `<html>`, `<p>` le deuxième fils de `<body>` ; `<text>` le premier fils du `<p>`. La deuxième partie de la position est l'indice de ce caractère dans l'élément `<text>` (ici, 2, car c'est le troisième caractère dans cet élément), ce qui nous donne la position complète du caractère : `"ex_doc 1.3.1.0-2"`. De même façon, nous obtenons la position du dernier caractère du mot « exemple » comme étant égal à `"ex_doc 1.3.1.0-8"`. Ce mot est donc identifié par les deux adresses `"ex_doc 1.3.1.0-2"` et `"ex_doc 1.3.1.0-8"`.

Un point important de ce mécanisme est que nous supposons que les chemins définissent toujours des unités *connexes*. C'est-à-dire que tous les caractères entre les deux positions

appartiennent à cette unité. Dans l'exemple 14, le mot « Mme. » étant divisé dans deux éléments différents, son chemin est "ex_doc 1.3.1.0-23 - ex_doc 1.3.1.1.0-2", où les premières parties de l'adresse renvoient à des éléments différents. Si, d'un autre point de vue, nous considérons que « Mme. XXX » est une unité unique, alors notre méthode donne le chemin "ex_doc 1.3.1.0-27 - ex_doc 1.3.1.2-3". La règle de continuité s'applique ici : l'élément <sup> ("1.3.1.1") étant entre les deux positions du chemin, son contenu fait partie de l'unité. Grâce à la règle de contiguité, un algorithme simple permet de facilement retrouver le contenu d'une unité à partir de son chemin.

3.2 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.

3.2.1 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.

3.2.2 L'élément docList

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

3.2.3 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.

3.2.4 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).

3.2.5 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>.

3.2.6 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.

3.2.7 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, la valeur doit indiquer le niveau de l'unité alignée, par exemple "align_tok_15".
- certainty : attribut facultatif, la valeur est un nombre entre 0 et 1, indiquant le niveau de confiance de l'annotateur sur ce <link>.

3.2.8 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 ce <annotation>.
- type : attribut obligatoire, qui encode le type de l'annotation. Les valeurs possibles sont "gram", "QE", "URI". Le type "gram" est 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 fur de l'avancement du projet.

3.2.9 L'élément docSpan

Un élément <docSpan> permet d'identifier une unité à l'intérieur du document. Il possède trois attributs :
- beginPos : attribut facultatif, la valeur doit indiquer la position du début de l'unité.
- endPos : attribut facultatif, la valeur doit indiquer la position de la fin de l'unité.

– `context` : attribut facultatif, la valeur doit être une série des *ids* des `<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>`.

3.2.10 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*) et `"lemma"` (le lemme), 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) [20] 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 [21]. 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.
– 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>`.

Un exemple complètement annoté correspondant aux deux premiers chapitres du livre de F. Cooper « The last of the Mohicans » est disponible sur le site du projet [22].

4 Conclusion

Nous avons dans ce document analysé les exigences du projet TransRead pour définir un formalisme permettant de stocker les document originaux et de représenter les annotations associées, en particulier les liens d'alignement. Notre solution contient deux aspects : pour les documents originaux, nous allons tous convertir au format du EPUB ; pour les annotations, nous avons proposé un format dans la section 3 qui est, à notre avis, conforme aux toutes les exigences.

20. `http://www.isocat.org/`
21. `http://www.loc.gov/standards/iso639-2/php/code_list.php`
22. `http://transread.limsi.fr/Resources/`

Certainement, toutes les situations des traitements des documents ne peuvent être prévues. Des évolutions sont attendus au fur et à mesure des développements du projet. Néanmoins, ce format possède d'une grande flexibilité et extensibilité, qui rendra les modifications faciles et efficaces.

Références

João de Almeida Varelas Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino António Caseiro. Building a golden collection of parallel multi-language word alignment. In *6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses : Open source toolkit for statistical machine translation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech Republic, 2007.

Patrik Lambert, Adrià Gispert, Rafael Banchs, and José B. Mariño. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39 (4) :267–285, 2005. ISSN 1574-020X. doi : 10.1007/s10579-005-4822-5. URL `http://dx.doi.org/10.1007/s10579-005-4822-5`.

Dan Melamed. *Empirical methods for exploiting parallel texts*. The MIT Press, Cambridge, 2001. ISBN 0262133806.

I. Dan Melamed. Annotation style guide for the Blinker project. Technical Report IRCS-98-06, University of Pennsylvania Institute for Research in Cognitive Science, 1998.

Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts : data driven machine translation and beyond - Volume 3*, HLT-NAACL-PARALLEL '03, pages 1–10, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi : 10.3115/1118905.1118906. URL `http://dx.doi.org/10.3115/1118905.1118906`.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1) :19–51, 2003.

Jörg Tiedemann. *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. PhD thesis, Uppsala University, Uppsala, Sweden, 2003. URL `http://uu.diva-portal.org/smash/record.jsf?pid=diva2:163715`. Anna Sågvall Hein, Åke Viberg (eds) : Studia Linguistica Upsaliensia.

Jörg Tiedemann. *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed). Morgan & Claypool Publishers, 2011. URL `http://dx.doi.org/10.2200/S00367ED1V01Y201106HLT014`.

Jean Véronis, editor. *Parallel Text Processing*. Text, Speech and Language Technology. Kluwer Academic Publishers, 2000.

Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank. http ://www.cl.uzh.ch/research/paralleltreebanks_en.html, 2010.

5 Annexes

5.1 La DTD d'annotation

Listing 15 – La DTD d'annotation

```
<!--                                                                -->
<!--                                                                -->
<!--            La DTD d'annotation pour TransRead                  -->
<!--                                                                -->
<!--                        Version  1.1                            -->
<!--                                                                -->

<!ELEMENT trAnnot     (docList, linkList+)                           >
<!ATTLIST trAnnot
          version     CDATA                              #REQUIRED >

<!ELEMENT docList     (docName+)                                     >
<!ELEMENT docName     (#PCDATA)                                      >
<!ATTLIST docName
          id          ID                                 #REQUIRED
          xml:lang    CDATA                              #IMPLIED  >

<!ELEMENT linkList    (linkGroup+)                                   >
<!ATTLIST linkList
          level       (sentence | token | chunk)         #REQUIRED >

<!ELEMENT linkGroup   (docPart+, (link+ | annotation+))              >
<!ATTLIST linkGroup
          type        (alignment | annotation)           #REQUIRED >

<!ELEMENT docPart     EMPTY                                          >
<!ATTLIST docPart
          doc         IDREF                              #REQUIRED
          beginPos    CDATA                              #IMPLIED
          endPos      CDATA                              #IMPLIED  >

<!ELEMENT link        (docSpan, docSpan?)                            >
<!ATTLIST link
          certainty   CDATA                              #IMPLIED
          id          ID                                 #REQUIRED >

<!ELEMENT annotation  (docSpan, mark*)                               >
<!ATTLIST annotation
          type        (gram | URI | QE)                  #REQUIRED
          id          ID                                 #REQUIRED >

<!ELEMENT docSpan     (#PCDATA)                                      >
```

```
<!ATTLIST docSpan
          beginPos    CDATA                          #REQUIRED
          endPos      CDATA                          #REQUIRED
          context     IDREFS                         #IMPLIED  >

<!ELEMENT mark        (#PCDATA)                                 >
<!ATTLIST mark
          cat         (POS | lemma)                  #IMPLIED
          resource    CDATA                          #IMPLIED
          xml:lang    CDATA                          #IMPLIED
          entry       CDATA                          #IMPLIED
          certainty   CDATA                          #IMPLIED
          qescore     CDATA                          #IMPLIED
          method      CDATA                          #IMPLIED  >
```

5.2 Le schéma XML

Listing 16 – Le schéma XML d'annotation

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

<!--                                                              -->
<!--                                                              -->
<!--        Le schéma XML d'annotation pour TransRead             -->
<!--                                                              -->
<!--                   Version  1.1                               -->
<!--                                                              -->


<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>

<xsd:simpleType name="cattype">
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="POS"/>
    <xsd:enumeration value="lemma"/>
  </xsd:restriction>
</xsd:simpleType>

<xsd:simpleType name="annotoption">
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="gram"/>
    <xsd:enumeration value="URI"/>
    <xsd:enumeration value="QE"/>
  </xsd:restriction>
</xsd:simpleType>

  <xsd:simpleType name="linkgroupoption">
    <xsd:restriction base="xsd:string">
      <xsd:enumeration value="alignment"/>
      <xsd:enumeration value="annotation"/>
    </xsd:restriction>
  </xsd:simpleType>

  <xsd:simpleType name="methodtype">
    <xsd:restriction base="xsd:string">
      <xsd:enumeration value="method1"/>
    </xsd:restriction>
  </xsd:simpleType>

  <xsd:simpleType name="leveltype">
    <xsd:restriction base="xsd:string">
      <xsd:enumeration value="sentence"/>
      <xsd:enumeration value="token"/>
      <xsd:enumeration value="chunk"/>
    </xsd:restriction>
  </xsd:simpleType>

  <xsd:complexType name="marktype">
    <xsd:simpleContent>
      <xsd:extension base="xsd:string">
        <xsd:attribute name="cat" type="cattype"/>
        <xsd:attribute name="resource" type="xsd:string"/>
        <xsd:attribute ref="xml:lang"/>
```

```
        <xsd:attribute name="entry" type="xsd:string"/>
        <xsd:attribute name="certainty" type="certaintytype"/>
        <xsd:attribute name="qescore" type="qescoretype"/>
        <xsd:attribute name="method" type="methodtype"/>
      </xsd:extension>
    </xsd:simpleContent>
  </xsd:complexType>

<xsd:complexType name="docspantype">
  <xsd:simpleContent>
    <xsd:extension base="xsd:string">
      <xsd:attribute name="beginPos" type="postype" use="required"/>
      <xsd:attribute name="endPos" type="postype" use="required"/>
      <xsd:attribute name="context" type="xsd:IDREFS"/>
    </xsd:extension>
  </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="annotationtype">
  <xsd:sequence>
    <xsd:element name="docSpan" type="docspantype"/>
    <xsd:element name="mark" type="marktype" maxOccurs="unbounded"/>
  </xsd:sequence>
  <xsd:attribute name="type" type="annotoption" use="required"/>
  <xsd:attribute name="id" type="xsd:ID" use="required"/>
</xsd:complexType>

<xsd:complexType name="linktype">
  <xsd:sequence>
    <xsd:element name="docSpan" type="docspantype" minOccurs="1" maxOccurs="2"/>
  </xsd:sequence>
  <xsd:attribute name="id" type="xsd:ID" use="required"/>
  <xsd:attribute name="certainty" type="certaintytype"/>
</xsd:complexType>

<xsd:complexType name="docparttype">
  <xsd:attribute name="doc" type="xsd:IDREF" use="required"/>
  <xsd:attribute name="beginPos" type="postype"/>
  <xsd:attribute name="endPos" type="postype"/>
</xsd:complexType>

<xsd:complexType name="linkgrouptype">
  <xsd:sequence>
    <xsd:element name="docPart" type="docparttype" maxOccurs="2"/>
    <xsd:choice>
      <xsd:element name="link" type="linktype" maxOccurs="unbounded"/>
      <xsd:element name="annotation" type="annotationtype" maxOccurs="unbounded"/>
    </xsd:choice>
```

```
    </xsd:sequence>
    <xsd:attribute name="type" type="linkgroupoption" use="required"/>
  </xsd:complexType>

  <xsd:complexType name="linklisttype">
    <xsd:sequence>
      <xsd:element name="linkGroup" type="linkgrouptype" maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="level" type="leveltype"/>
  </xsd:complexType>

  <xsd:complexType name="docnametype">
    <xsd:simpleContent>
      <xsd:extension base="xsd:string">
        <xsd:attribute name="id" type="xsd:ID" use="required"/>
        <xsd:attribute ref="xml:lang"/>
      </xsd:extension>
    </xsd:simpleContent>
  </xsd:complexType>

  <xsd:complexType name="doclisttype">
    <xsd:sequence>
      <xsd:element name="docName" type="docnametype" maxOccurs="unbounded"/>
    </xsd:sequence>
  </xsd:complexType>

  <xsd:complexType name="trannottype">
    <xsd:sequence>
      <xsd:element name="docList" type="doclisttype"/>
      <xsd:element name="linkList" type="linklisttype" maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="version" type="xsd:decimal" use="required"/>
  </xsd:complexType>

  <xsd:element name="trAnnot" type="trannottype"/>
</xsd:schema>
```

5.3 Un exemple d'annotation

Ce qui suit est un exemple d'annotation. Les listings 17 et 18 sont deux documents originaux *Mohicans_en.xhtml* et *Mohicans_fr.xhtml*. Le listing 19 est une partie représentative de l'annotation pour ces deux documents.

Listing 17 – Mohicans_en.xhtml

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE html PUBLIC '-//W3C//DTD XHTML 1.1//EN'
        'http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd'>
<html xmlns="http://www.w3.org/1999/xhtml"><head>
<title>The last of the Mohicans</title>
</head>

<body>
<!--p>Le texte commence ici</p-->
<!--p>transread_begin</p-->

<p><h1>The last of the Mohicans James Fenimore Cooper</h1></p>

<p><h2>CHAPTER I</h2></p>

<p><pre>" Mine ear is open , and my heart prepared : The worst
is worldly loss thou canst unfold : Say , is my kingdom lost ? "
Shakespeare .</pre></p>

<p>It was a feature peculiar to the colonial wars of North America ,
that the toils and dangers of the wilderness were to be encountered
before the adverse hosts could meet .
A wide and apparently an impervious boundary of forests severed the
possessions of the hostile provinces of France and England .
The hardy colonist , and the trained European who fought at his side ,
frequently expended months in struggling against the rapids of the streams ,
or in effecting the rugged passes of the mountains , in quest of an
opportunity to exhibit their courage in a more martial conflict .</p>

</body>
</html>
```

Listing 18 – Mohicans_fr.xhtml

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE html PUBLIC '-//W3C//DTD XHTML 1.1//EN'
          'http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd'>
<html xml:lang="fr"  xmlns="http://www.w3.org/1999/xhtml"><head>
<title>Le dernier des Mohicans</title>
</head>

<body>
<!--p>Le texte commence ici</p-->
<!--p>transread_begin</p-->

<p><h1>Le dernier des Mohicans James Fenimore Cooper</h1></p>

<p><h2>Chapitre premier</h2></p>

<p><pre>Mon oreille est ouverte . Mon coeur est préparé ; quelque
perte que tu puisses me révéler , c' est une perte mondaine ;
parle , mon royaume est -il perdu ?
Shakespeare .</pre></p>

<p>C' était un des caractères particuliers des guerres qui ont eu
lieu dans les colonies de l' Amérique septentrionale , qu' il fallait
braver les fatigues et les dangers des déserts avant de pouvoir livrer
bataille à l' ennemi qu' on cherchait . Une large ceinture de forêts ,
en apparence impénétrables , séparait les possessions des provinces
hostiles de la France et de l' Angleterre . Le colon endurci aux travaux
et l' Européen discipliné qui combattait sous la même bannière , passaient
quelquefois des mois entiers à lutter contre les torrents , et à se frayer
un passage entre les gorges des montagnes , en cherchant l' occasion de
donner des preuves plus directes de leur intrépidité .</p>

</body>
</html>
```

Listing 19 – Une partie de l'annotation

```
<?xml version="1.0" 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>
  </linkList>
  <linkList level="token">
    <linkGroup type="alignment">
      <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-0" endPos="doc_en 1.2.11.0-171"/>
      <docPart doc="doc_fr" beginPos="doc_fr 1.2.11.0-0" endPos="doc_fr 1.2.11.0-243"/>
      <link id="align_tok_40">
        <docSpan beginPos="doc_en 1.2.11.0-0"
                 endPos="doc_en 1.2.11.0-2">it</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-0"
                 endPos="doc_fr 1.2.11.0-2">c'</docSpan>
      </link>
```

```
    <link id="align_tok_41">
      <docSpan beginPos="doc_en 1.2.11.0-3"
               endPos="doc_en 1.2.11.0-6">was</docSpan>
      <docSpan beginPos="doc_fr 1.2.11.0-3"
               endPos="doc_fr 1.2.11.0-8">était</docSpan>
    </link>
    <link id="align_tok_42">
      <docSpan beginPos="doc_en 1.2.11.0-7"
               endPos="doc_en 1.2.11.0-8">a</docSpan>
      <docSpan beginPos="doc_fr 1.2.11.0-9"
               endPos="doc_fr 1.2.11.0-11">un</docSpan>
    </link>
    <link id="align_tok_43">
      <docSpan beginPos="doc_en 1.2.11.0-9"
               endPos="doc_en 1.2.11.0-16">feature</docSpan>
      <docSpan beginPos="doc_fr 1.2.11.0-16"
               endPos="doc_fr 1.2.11.0-26">caractères</docSpan>
    </link>
    <link id="align_tok_44">
      <docSpan beginPos="doc_en 1.2.11.0-17"
               endPos="doc_en 1.2.11.0-25">peculiar</docSpan>
      <docSpan beginPos="doc_fr 1.2.11.0-27"
               endPos="doc_fr 1.2.11.0-39">particuliers</docSpan>
    </link>
    <link id="align_tok_66">
      <docSpan beginPos="doc_en 1.2.11.0-122"
               endPos="doc_en 1.2.11.0-133">encountered</docSpan>
      <docSpan beginPos="doc_fr 1.2.11.0-133"
               endPos="doc_fr 1.2.11.0-139">braver</docSpan>
    </link>
  </linkGroup>
  <linkGroup type="annotation">
    <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-0" endPos="doc_en 1.2.11.0-171"/>
    <annotation id="annot_tok_1" type="gram">
      <docSpan beginPos="doc_en 1.2.11.0-122"
               endPos="doc_en 1.2.11.0-133">encountered</docSpan>
      <mark cat="lemma" certainty="1">encounter</mark>
      <mark cat="POS" certainty="1">VBN</mark>
    </annotation>
    <annotation id="annot_tok_2" type="URI">
      <docSpan beginPos="doc_en 1.2.11.0-122"
               endPos="doc_en 1.2.11.0-133">encountered</docSpan>
      <mark certainty="0.8" resource="babelnet" xml:lang="en">run into;
            be beset by;"The project ran into numerous financial difficulties"</mark>
    </annotation>
  </linkGroup>
  <linkGroup type="alignment">
    <docPart doc="doc_en" beginPos="doc_en 1.2.11.0-301" endPos="doc_en 1.2.11.0-582"/>
```

```
<docPart doc="doc_fr" beginPos="doc_fr 1.2.11.0-387" endPos="doc_fr 1.2.11.0-692"/>
<link id="align_tok_102">
  <docSpan beginPos="doc_en 1.2.11.0-326"
           endPos="doc_en 1.2.11.0-329">the</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-419"
           endPos="doc_fr 1.2.11.0-421">l'</docSpan>
</link>
<link id="align_tok_103">
  <docSpan beginPos="doc_en 1.2.11.0-330"
           endPos="doc_en 1.2.11.0-337">trained</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-431"
           endPos="doc_fr 1.2.11.0-441">discipliné</docSpan>
</link>
<link id="align_tok_104">
  <docSpan beginPos="doc_en 1.2.11.0-338"
           endPos="doc_en 1.2.11.0-346">european</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-422"
           endPos="doc_fr 1.2.11.0-430">européen</docSpan>
</link>
<link id="align_tok_105">
  <docSpan beginPos="doc_en 1.2.11.0-347"
           endPos="doc_en 1.2.11.0-350">who</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-442"
           endPos="doc_fr 1.2.11.0-445">qui</docSpan>
</link>
<link id="align_tok_106">
  <docSpan beginPos="doc_en 1.2.11.0-351"
           endPos="doc_en 1.2.11.0-357">fought</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-446"
           endPos="doc_fr 1.2.11.0-456">combattait</docSpan>
</link>
<link id="align_tok_107">
  <docSpan beginPos="doc_en 1.2.11.0-358" endPos="doc_en 1.2.11.0-360"
        context="align_seg_1">at</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-457" endPos="doc_fr 1.2.11.0-461"
          context="align_seg_1">sous</docSpan>
</link>
<link id="align_tok_108">
  <docSpan beginPos="doc_en 1.2.11.0-361" endPos="doc_en 1.2.11.0-364"
        context="align_seg_1">his</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-446"
           endPos="doc_fr 1.2.11.0-456">combattait</docSpan>
</link>
<link id="align_tok_109">
  <docSpan beginPos="doc_en 1.2.11.0-361" endPos="doc_en 1.2.11.0-364"
        context="align_seg_1">his</docSpan>
  <docSpan beginPos="doc_fr 1.2.11.0-462" endPos="doc_fr 1.2.11.0-464"
        context="align_seg_1">la</docSpan>
```

```
      </link>
      <link id="align_tok_110">
        <docSpan beginPos="doc_en 1.2.11.0-365" endPos="doc_en 1.2.11.0-369"
              context="align_seg_1">side</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-465" endPos="doc_fr 1.2.11.0-469"
              context="align_seg_1">même</docSpan>
      </link>
      <link id="align_tok_111">
        <docSpan beginPos="doc_en 1.2.11.0-365" endPos="doc_en 1.2.11.0-369"
              context="align_seg_1">side</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-470" endPos="doc_fr 1.2.11.0-478"
              context="align_seg_1">bannière</docSpan>
      </link>
      <link id="align_tok_134">
        <docSpan beginPos="doc_en 1.2.11.0-502" endPos="doc_en 1.2.11.0-504"
              context="align_seg_3">in</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-610" endPos="doc_fr 1.2.11.0-612"
                context="align_seg_3">en</docSpan>
      </link>
      <link id="align_tok_135">
        <docSpan beginPos="doc_en 1.2.11.0-505" endPos="doc_en 1.2.11.0-510"
              context="align_seg_3">quest</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-613" endPos="doc_fr 1.2.11.0-622"
              context="align_seg_3">cherchant</docSpan>
      </link>
      <link id="align_tok_136">
        <docSpan beginPos="doc_en 1.2.11.0-514"
                 endPos="doc_en 1.2.11.0-516">an</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-623"
                 endPos="doc_fr 1.2.11.0-625">l'</docSpan>
      </link>
      <link id="align_tok_137">
        <docSpan beginPos="doc_en 1.2.11.0-517"
                 endPos="doc_en 1.2.11.0-528">opportdocSpany</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-626"
                 endPos="doc_fr 1.2.11.0-634">occasion</docSpan>
      </link>
    </linkGroup>
  </linkList>
  <linkList level="chunk">
    <linkGroup type="alignment">
      <docPart doc="doc_en"/>
      <docPart doc="doc_fr"/>
      <link id="align_seg_1">
        <docSpan beginPos="doc_en 1.2.11.0-358"
                 endPos="doc_en 1.2.11.0-369"> at his side</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-457"
                 endPos="doc_fr 1.2.11.0-478">sous la même bannière</docSpan>
```

```
      </link>
      <link id="align_seg_3">
        <docSpan beginPos="doc_en 1.2.11.0-502"
                 endPos="doc_en 1.2.11.0-513">in quest of</docSpan>
        <docSpan beginPos="doc_fr 1.2.11.0-610"
                 endPos="doc_fr 1.2.11.0-622">en cherchant</docSpan>
      </link>
    </linkGroup>
  </linkList>
</trAnnot>
```