CN RNIL N 6
2003-11-25

Secrétariat CN RNIL

Votre correspondant: **Tony HITTEMA**

Ligne directe: : 01 41 62 83 95
Télécopie : 01 41 62 90 33

E-mail : tony.hittema@afnor.fr

CN RNIL
Ressources Normalisées en Ingénierie de la Langue

Association

Française de

Normalisation

11 avenue Francis de Pressensé

93571 Saint-Denis La Plaine Cedex

France

Tél. : +33 (0)1 41 62 80 00

Fax : +33 (0)1 49 17 90 00

http://www.afnor.fr

TITRE : Contribution au projet **ISO/AWI 24611 « Language resource management -- Morpho syntactic annotation framework »**

SOURCE : Lionel CLEMENT, INRIA

PROJET : AWI 24611

STATUT : Contribution française

ACTION : Pour discussion à la réunion du 2 décembre 2003

DIFFUSION : Membres de la commission

Association reconnue

d'utilité publique

Comité membre français

du CEN et de l'ISO

Siret 775 724 818 00015

Code NAF 751 ER

Langage Resource Management
Morpho-Syntactic Annotation Framework
Draft French version

Lionel Clément
INRIA Rocquencourt – Atoll project

French version 0.2

1 General characteristics

L'annotation morpho-syntaxique consiste à segmenter un texte puis à assigner à chaque segment une étiquette marquant la partie du discours (nom, adjectif, verbe, etc.), les traits morphologiques, les catégories grammaticales (nombre, genre, personne, mode et temps verbal, etc.) et quelques propriétés linguistiques de langue et de parole.

Cette annotation est utilisée par la communauté des linguistes utilisant l'outil informatique et plus particulièrement par celle du Traitement Automatique des Langues comme résultat d'une opération de marquage linguistique précis qui s'inscrit relativement à d'autres opérations sans les recouvrir totalement. Les annotations de référence, de discours, de prosodie, de construction syntaxique, etc. sont autant d'annotations linguistiques qui complètent l'annotation morpho-syntaxique tout en la recoupant en partie seulement.

La sémantique et la syntaxe interviennent invariablement dans la définition des parties du discours et des catégories grammaticales. Ainsi, les pronoms et substantifs sont porteurs en propre d'une référence, le temps et l'aspect des verbes marquent la deixis temporelle, la personne, la modalité et d'autres catégories grammaticales marquent la situation d'énonciation, etc.

Il n'est donc pas aisé de décrire le champ exact de l'annotation morpho-syntaxique, car celle-ci est fortement corrélée en langue ou en parole aux autres propriétés linguistiques des textes. Nous délimiterons cependant les séquences minimales et maximales de texte pouvant être identifiées comme *unités morpho-syntaxiques* dans un rapport syntagmatique, et nous catégoriserons les particularités linguistiques propres à marquer ces unités, dans

un rapport associatif. Les unités minimales ne se décomposent pas en sous-unités qui pourraient être identifiées selon les mêmes critères, elles sont donc *atomiques* de ce point de vue, mais elles peuvent s'analyser en morphologie ou en phonologie. Symétriquement, les unités maximales ne sont pas constituantes d'*unités morpho-syntaxiques*, elle sont en revanche analysables en syntaxe.

2 Segmentation

L'annotation *morpho-syntaxique* préserve le caractère linéaire du signifiant, qui s'inscrit dans l'espace ou dans le temps, suivant que l'objet de l'annotation est une image acoustique ou un document écrit. Les documents annotés sont des séquences linéaires de texte de natures très différentes (enregistrements sonores, retranscriptions, écritures diverses, etc.).

Les éléments identifiés comme séquences propres à recevoir une annotation *morpho-syntaxique* peuvent, par conséquent, ne pas être seulement des chaînes de caractères, mais plus généralement des références de pointeurs dans des documents de diverses natures. Dans ce cas, les annotations qui n'intéressent pas directement la *morpho-syntaxe* pourront être apportées au document original. Elles pourront présenter des segmentations plus fines ou plus grossières du document sans produire d'incohérence avec l'annotation *morpho-syntaxique*.

La *Text Encoding Initiative* (TEI) offre différents mécanismes pour identifier des séquences de documents et les relier entre eux. La section 14.3 (*Blocks, Segments and Anchors*) les présente :

« *<**anchor**> attaches an identifier to a point within a text, whether or not it corresponds with a textual element.*

» *The <**anchor**> element may be thought of as an empty <**seg**>, or as an artifice enabling an identifier to be attached to any position in a text. Like the <**mileStone**> element discussed in section 6.9 Reference Systems, it is useful where multiple views of a document are to be combined, for exemple, when a logical view based on paragraphs or verse lines is to be mapped on to a physical view based on manuscript lines. However, it differs from the milestone and related elements in that the <**anchor**> element should not be used to mark the start or end of an arbitrary zone within a text, but only to mark an arbitrary point used for alignment, or as the target of a spanning element such as those discussed in section 18.1.4 Additions and Deletions.*

» *<**seg**> contains any arbitrary phrase-level unit of text (including other*

seg elements). subtype provides a sub-categorization of the segment marked.

» *The <**seg**> element may be used at the encoder's discretion to mark almost any segment of the text of interest for processing. One use of the element is to mark text features for which no appropriate markup is otherwise defined, i.e. as a simple extension mechanism. Another use is to provide an identifier for some segment which is to be pointed at by some other element, i.e. to provide a target, or a part of a target, for a <**ptr**> or other similar element.* »

La section 15.1 *Linguistic Segment Categories* de la *TEI* offre quant à elle, un mécanisme pour annoter les documents avec des informations linguistiques :

« *In this section we introduce specialized linguistic segment category elements which may be used to represent the segmentation of a text into the traditional linguistic categories of sentence, clause, phrase, word, morpheme, and characters.*

» *<**s**> contains a sentence-like division of a text. No attributes other than those globally available (see definition for a.global)*

» *<**cl**> represents a grammatical clause. No attributes other than those globally available (see definition for a.global)*

» *<**phr**> represents a grammatical phrase. No attributes other than those globally available (see definition for a.global)*

» *<**w**> represents a grammatical (not necessarily orthographic) word. lemma identifies the word's lemma (dictionary entry form).*

» *<**m**> represents a grammatical morpheme. baseform identifies the morpheme's base form.*

» *<**c**> represents a character. No attributes other than those globally available (see definition for a.global)* »

Ces catégories linguistiques ne sont pas assez précises pour une normalisation en *morpho-syntaxe*. En particulier, nous proposons une architecture de segmentation de l'unité morpho-syntaxique qui permet une description plus fine que les éléments <**w**> et <**m**> de la TEI. En revanche, nous voudrions que l'usage de la norme en *morpho-syntaxe* ne soit pas rédhibitoire pour un usage de la TEI ou de tout autre type de segmentation.

La segmentation de document que nous proposons doit donc pouvoir s'appuyer sur des **positions** dans le document existant et susceptible de modifications, selon un mécanisme de pointage unique. La *TEI* (*14.1.1 Pointers and Links*) propose un tel mécanisme pour identifier les liens hypertextes et autres références croisées d'un document :

« *<**ptr**> defines a pointer to another location in the current document*

in terms of one or more identifiable elements. target specifies the destination of the pointer by supplying the values used on the id attribute of one or more other elements in the current document »

Nous utiliserons ce mécanisme simple pour faire référence à une séquence linéaire de texte dans des documents variés porteurs d'annotations *morpho-syntaxiques.*

2.1 Token

Si les unités d'annotation morpho-syntaxiques correspondent à des segments *in praesentia* du flux textuel, ceci ne suppose nullement que le texte annoté soit immédiatement une séquence connexe de segments partitionnant le document. Il est en effet important de distinguer les unités *morpho-syntaxiques* de leurs réalisations effectives. Certaines parties pourront ne pas être annotées (signes de mise en page, didascalies, signes de balisage du document) d'autres pourront ne pas correspondre exactement à la forme segmentée (abréviations, brachygraphies, erreurs typographiques, variations typographiques, contractions typographiques ou morphologiques, etc.). Nous devrons être en mesure d'annoter ces flux contigus de texte, sans marque repérable *a priori* de balisage dans le texte original (écriture du sanscrit sans séparation entre les mots, composition nominale en allemand, retranscription de parole, etc.).

L'élément <**token**> marque ces unités morphologiques qui s'articulent entre elles pour fournir le matériau de la construction morpho-phonologique. En effet, elles représentent assez naturellement les éléments de la structure phonologique des morphèmes, mais peuvent être adaptées à d'autres analyses de la constitution des *mots*. Ainsi, un ***mot-forme*** (**word-form** dorénavant), seul élément propre à recevoir une annotation morpho-syntaxique, est une composition, une agglutination ou toute autre construction issue d'un ou plusieurs **tokens**. Nous ne fixons pas la nature linguistique de l'élément <**token**>. Tantôt il s'agit d'une simple séquence typographique, tantôt de l'analyse morphologique d'un terme (racine, affixe, morphème, etc.). Dans tous les cas, l'analyse de la **constitution** morphologique, phonologique voire lexicologique des termes échappe à l'annotation morpho-syntaxique en propre et ne figure donc pas dans cette norme.

Les marques typographiques de mise en page, de séparation des mots et des paragraphes, ainsi que tous les encodages associés à une annotation linguistique du texte qui échappe à la morpho-syntaxe, pourront être conservés dans le document auquel font référence les éléments <**token**>. Une séquence de texte doit donc pouvoir faire référence à une occurrence d'un intervalle

dans un document pointé par l'élément <**ptr**> de la *TEI* (*14.1.1 Pointers and Links*).

L'élément <**token**> peut ainsi faire référence à un couple de pointeurs sur un document :

```
<s><ptr target="p1"/>The <ptr target="p2"/>victim
<ptr target="p3"/>'<ptr target="p4"/>s
<ptr target="p5"/>friends <ptr target="p6"/>told
<ptr target="p7"/>police <ptr target="p8"/>that
<ptr target="p9"/>Krueger <ptr target="p10"/>drove
<ptr target="p11"/>into <ptr target="p12"/>the
<ptr target="p13"/>quarry <ptr target="p14"/>and
<ptr target="p15"/>never <ptr target="p16"/>surfaced
<ptr target="p17"/>.<ptr target="p18"/></s>
```

```
<token id="t1" from="p1" to="p2"/>
<token id="t2" from="p2" to="p4"/>
<token id="t3" from="p4" to="p5"/>
<token id="t4" from="p5" to="p6"/>
...
```

Pour des applications plus immédiates et une représentation sans pointeurs, nous proposons, en alternative, que la réalisation du texte annoté soit directement présent comme *contenu* de l'élément <**token**>. Dans ce cas, le document lui-même doit ne pas être imcompatible avec l'annotation en séquences minimales (comme ce serait le cas avec l'usage de signes de mise en page qui embrassent plusieurs **tokens**)[1].

```
<token id="t1">The</token>
<token id="t2">victim</token>
<token id="t3">'s</token>
<token id="t4">friends</token>
<token id="t5">told</token>
<token id="t6">police</token>
<token id="t7">that</token>
<token id="t8">Krueger</token>
<token id="t9">drove</token>
<token id="t10">into</token>
<token id="t11">the</token>
<token id="t12">quarry</token>
<token id="t13">and</token>
<token id="t14">never</token>
```

[1] Autrement dit, et pour une définition de *DTD*, le contenu de l'élément XML correspondant à <**token**> doit être du PCDATA.

```
<token id="t15">surfaced</token>
<token id="t16">.</token>
```

Les <**token**> peuvent se chevaucher, appartenir aux mêmes intervalles séquentiels (par example sur des documents multi-locuteurs avec recouvrement de parole), et éventuellement correspondre à des séquences nulles. Dans ces cas, ils ne correspondent pas immédiatement à la réalisation d'une séquence graphique ou sonore, mais sont les représentations d'unités linguistiques propres à segmenter un texte.

Ainsi, les variations graphiques ou phoniques des mêmes séquences recevront une valeur unique définitoire de l'élément <**token**>. Il peut s'agir de l'extension d'une écriture abrégée, d'une forme corrigée du texte ou d'une retranscription d'un texte[2].

La séquence *etc.* en français ou en anglais, en fin de phrase, peut-elle correspondre à deux ***tokens*** marquant respectivement l'abréviation et la ponctuation.

Voici deux façons de représenter cette séquence :

a

```
<token value="et caetera" id="t1">etc.</token>
<token value="#dot#" id="t2"/>
```

b

```
<token value="et caetera" id="t1" from="p1" to="p3"/>
<token value="#dot#" id="t2" from="p2" to="p3"/>
```

```
<ptr target="p1"/>etc<ptr target="p2"/>.<ptr target="p3"/>
```

En **a**, la séquence "***etc.***" est scindée en deux segments ("***etc.***" et "$\emptyset$"). En **b**, la même séquence correspond à deux segments qui se chevauchent ("***etc.***" et "**.**"). Dans les deux cas, deux **tokens** distincts correspondent à l'agglutination graphique de deux éléments.

L'attribut **value** de l'élément <**token**> contient la valeur de l'interprétation linguistique de la séquence. Elle permet de représenter non la réalisation morpho-phonologique du flux textuel lui-même, mais le matériau linguistique pertinent du point de vue de la morpho-syntaxe.

En grec moderne, par exemple, l'expression idiomatique "$\kappa\alpha\lambda\acute{o}\kappa\alpha\gamma\alpha\theta o\varsigma$" (*bon et brave*) peut se segmenter en trois termes agglutinés : "$\kappa\alpha\lambda\acute{o}\varsigma$", "$\kappa\alpha\iota$", et "$\alpha\gamma\alpha\theta o\varsigma$".

Ce que nous noterons :

[2] Nous pourrions avoir besoin d'une notation de l'alternative, comme celle qui est utilisée pour représenter les différentes interprétations d'une même séquence sonore lors de la retranscription. Est-ce l'objet d'un standard sur l'annotation morpho-syntaxique que de proposer une telle possibilité ?

```
<token value="καλός" id="t0">καλο</token>
<token value="και" id="t1">κ</token>
<token value="αγαθός" id="t2">αγαθος</token>
```

Récapitulatif : élément <token>

- <**token**> est une unité morpho-phonologique (elle peut être définie selon des analyses différentes de la constitution des unité *morpho-syntaxiques*. Elle correspond à une séquence connexe d'un document
- Cette séquence est définie :
  - Soit par le contenu de l'élément (il ne contient alors que du texte et éventuellement des annotations non enchâssantes)
  - Soit par les attributs **from** et **to** qui pointent vers des marques uniques (IDREF) d'un document pour en définir une séquence (i.e. attribut **target** de l'élément <**ptr**> de la TEI sur un autre document)
- **from** et **to** sont des identifiants de pointeurs sur un document source
- **value** fournit le contenu linguistique du token (morphème réalisé ou non, extention d'abréviation, etc.)
- **id** introduit un identifiant unique pour l'élément <**token**>

2.2 *Word-form*

L'agglutination morphologique *auquel* en français peut donner lieu à plusieurs analyses :

1. La séquence *auquel* n'est pas décomposée et correspond à un seul **token**

```
<token form="auquel" id="t0">auquel</token>
```

   (a) Le mot fait référence à ce seul token et porte une étiquette qui renseigne sa nature polycatégorielle

```
<wordForm entry="auquel" tag="Prép.+Pro._Relatif"
    tokens="t0"/>
```

   (b) Le mot est décomposée en deux parties qui portent sur ce même **token**

```
<wordForm entry="à" tag="Prép." tokens="t0"/>
<wordForm entry="lequel" tag="Pro._Relatif" tokens=
    "t0"/>
```

2. La séquence est décomposée en deux **tokens** *à*, *lequel*

```
<token form="à" id="t0">auquel</token>
<token form="lequel" id="t1"/>
```

(a) Le mot fait référence à cette séquence de tokens et porte une étiquette qui renseigne sa nature polycatégorielle

```
<wordForm entry="auquel" tag="Prép.+Pro. Relatif"
    tokens="t0 t1"/>
```

(b) Le mot est décomposé en deux parties qui portent respectivement sur les deux **tokens**

```
<wordForm entry="à" tag="Prép." tokens="t0"/>
<wordForm entry="lequel" tag="Pro. Relatif" tokens=
    "t1"/>
```

Ces différentes analyses peuvent toutes être motivées par les usages ou en fonction des outils de traitement de la langue utilisés. Elles pourront toutes être correctement annotées en suivant les recommandations que nous décrivons. La norme ne doit donc rien dire sur la nature de la décomposition morpho-phonologiques des éléments textuels. Le **token** peut être l'objet d'une analyse morphologique ou être une séquence reconnue automatiquement comme appartenant à un langage régulier par exemple.

Cela n'en fait pas un élément linguistique défini dans un rapport syntagmatique. Le **word-form** est cet élément. Il se rapporte directement au **token** ou à la séquence de **tokens** qui segmentent le texte, et a le statut d'une unité linguistique sur laquelle porte l'étiquetage *morpho-syntaxique*.

Nous ne discutons pas des choix théoriques qui légitiment ce marquage. Il peut être donné selon les propriétés lexicales ou morphologiques en contexte ou en langue (selon la *nature* et la *fonction* des mots, pour reprendre une terminologie répandue). Ici encore, les spécifications ne fournissent pas une réponse à ces questions, mais elles offrent le moyen d'annoter un élément linguistique.

Un **word-form** pointe vers un ou plusieurs **tokens** selon un mécanisme d'identification unique (IDREFS).

Il peut correspondre à une séquence non connexe de **tokens** :

```
<token value="afin" id="t1">afin</token>
<token value="justement" id="t2">justement</token>
```

```
<token value="de" id="t3">de</token>

<wordForm entry="Afin de" tokens="t1 t3"/>
<wordForm entry="justement" tokens="t2"/>
```

Il peut correspondre à une réalisation vide d'un morphème (dans ce cas, l'attribut ***tokens*** a une valeur vide) :

```
<token value="Jean" id="t1">Jean</token>
<token value="propose" id="t2">propose</token>
<token value="de" id="t3">de</token>
<token value="partir" id="t4">partir</token>

<wordForm entry="Jean" tokens="t1"/>
<wordForm entry="propose" tokens="t2"/>
<wordForm entry="de" tokens="t3"/>
<wordForm entry="PRO" tokens=""/>
<wordForm entry="partir" tokens="t4"/>
```

Enfin plusieurs **word-forms** peuvent se rapporter au même **token** :

```
<token value="damelo" id="t1">Damelo</token>
      <!-- (Donne-le moi) -->

<wordForm entry="da" tokens="t1"/> <!-- (Donne) -->
<wordForm entry="me" tokens="t1"/> <!-- (le) -->
<wordForm entry="lo" tokens="t1"/> <!-- (moi) -->
```

Prenons le cas du subtantif allemand *Geburtstagsgeschenkpapier* (papier cadeau pour anniversaire) :

Dans le document original, on peut identifier des pointeurs sur les intervalles de textes en suivant les conventions de la *TEI*, afin d'identifier trois *tokens* :

```
<seg>
<ptr target="p1"/>Geburtstags<ptr target="p2"/>geschenk<ptr
    target="p3"/>papier<ptr target="p4"/>
</seg>
```

```
<token value="Geburtstag" id="t1" from="p1" to="p2"/>
<token value="Geschenk" id="t2" from="p2" to="p3"/>
<token value="Papier" id="t3" from="p3" to="p4"/>
```

```
<wordForm entry="Geburtstagsgeschenkpapier" tokens="t1 t2 t3
    "/>
```

La séquence textuelle ne comporte aucun élément séparateur. Mais une analyse morphologique peut décomposer ce substantif en trois **tokens** : *Geburtstag*, *Geschenk* et *Papier*. Cette composition nominale de l'allemand permet d'identifier la séquence comme une unité *morpho-syntaxique* qui s'articule dans le texte avec les autres unités *morpho-syntaxiques*.

Il est important de remarquer que l'absence de séparateur dans le mot n'est pas cruciale pour identifier l'unité. En revanche, le "s" entre *Geburtstag* et *Geschenk* est un "Fugelement", un element introduit pour marquer la composition et qui ne marque pas en soi le cas genitif. Dans ce sens le "s" est un element separateur et non pas une partie du mot *Geburtstag*. Les compositions nominales en français écrites avec des espaces, les phrases sanskrites écrites sans séparateurs, les agglutinations des pronoms clitiques des langues romanes, etc., recevront une analyse en **tokens** et **word-forms** pareillement. Le **word-form** est un élément linguistique identifié pour ses propriétés morpho-syntaxiques. Ici *Geburtstagsgeschenkpapier* est un terme, tel qu'il a été analysé comme une unité lexicographique (la décomposition en plusieurs lexèmes est évidemment possible), ou comme une unité ayant une fonction grammaticale dans la phrase. Cette identification est notée grâce à l'attribut **entry**.

L'attribut **entry** identifie le contenu linguistique du **word-form** comme unité sur laquelle porte l'annotation *morpho-syntaxique*.

```
  <token value="prime" id="t1">Prime</token> <token value="
      minister"
  id="t2">minister</token>
```

```
<wordForm entry="prime minister" tokens="t1 t2"/>
```

Jusqu'à présent, les unité linguistiques ont été définies dans un rapport syntagmatique comme séquences textuelles plus ou moins complexes. L'étiquetage morpho-syntaxique permet de caractériser les unités non dans ce rapport, mais *in absentia*, relativement à la nature linguistique qui les caractérise, et relativement aux fonctions grammaticales qu'elles ont dans le texte.

Le **word-form** se distingue dans un rapport associatif comme porteur d'une catégorie. Cette catégorie peut être aussi complexe que possible :

– Associant à une catégorie grammaticale ou à une autre propriété linguistique (i.e. partie du discours, lemme, lexème, etc.) une ou plusieurs valeurs.
– Associant à un type une valeur complexe (i.e. ensemble de traits morphologiques)

Il peut également contenir des informations intéressant la morpho-syntaxe mais qui ne caractérisent pas l'élément comme unité morpho-syntaxique :

– Mémoire de correction et/ou de traitements automatiques pour l'annotation morpho-syntaxique.
– Probabilité d'une étiquette choisie par un étiqueteur stochastique.
– Référence à un lexique d'exceptions ou de spécialité
– etc.

Cette catégorie sera alors contenue dans le corps de l'élément sous forme d'une structure de traits comme nous allons le voir. Ce marquage *morpho-syntaxique* peut être représenté de façon plus économique par une simple étiquette sous l'attribut *tag* : Dans les usages les plus courants, où il s'agit d'assortir une étiquette unique et simple à chaque mot d'un texte, ce seul attribut suffira comme étiquetage morpho-syntaxique.

Résumé : élément <word-form>

– <**word-form**> correspond à une unité morpho-syntaxique
– **tokens** (IDREFS) pointe vers un ou plusieurs identifieurs de **tokens** connexes ou non. Si cet attribut est vide, le **word-form** n'est pas réalisé sous forme de **tokens** (il n'a pas de réalistation dans le texte) [3]
– **entry** correspond à une identité linguistique du **word-form**. Ce peut être l'identifieur d'une ou plusieurs entrées lexicales, une catégorie grammaticale, le lemme d'un mot, son lexème, etc.

3 Étiquetage

Le rapport associatif suppose une analyse linguistique que la norme ne prévoie pas de fixer. La même norme pourra être utilisée par un ensemble de linguistes qui donnent des définitions aux étiquettes morpho-syntaxiques selon des critères très différents. Cependant, l'usage systématique des registres de catégories de données permettra d'indiquer au lecteur quels sens sont donnés ici aux éléments et attributs utilisés. Les parties du discours

[3] La notation d'une réalisation vide pose un problème de représentation. Doit-on réserver un identifiant pour le morphème vide et ses variantes (∅, PRO, etc.) ?

pourront se rapporter à l'analyse distributionnelle ou morphologique, des étiquettes pourront être données selon des analyses morphologiques, syntaxiques, sémantiques ou pragmatiques sans que le modèle ne soit jamais remis en cause.

L'usage en morpho-syntaxe est d'assortir chaque unité morpho-syntaxique d'une étiquette plus ou moins complexe qui dénote un ensemble de propriétés linguistiques :

- Une ou plusieurs parties du discours parmi une liste plus ou moins classique (nom, adjectif, verbe, adverbe, interjection, ...).
  Nous notons que ces *parties du discours* peuvent aussi être des catégories distributionnelles selon certaines analyses. Elles comprennent généralement des types très hétérogènes (nombres, ponctuations, mots étrangers, abréviations, mots résiduels, classe à membre unique, ...). L'usage le plus courant est cependant que ces *parties du discours* constituent une partition de la nature des mots.
- Sous-types très variés. En voici quelques exemples courants : catégories distributionnelles (pronoms conjoints), propriété sémantique (mots négatifs), propriété syntaxique (verbes auxiliaires), propriété discursive (déictiques spatio-temporelles) ...
  Nous pouvons naturellement augmenter cette liste selon les propriétés intrinsèques ou extrinsèques des unités morpho-syntaxiques qu'offrent de multiples analyses.
- Des marques morphologiques, soit issues de l'analyse contextuelle (marques d'accord, formes casuelles, ...), soit définies selon les propriétés lexicales du *mot* (genre du nom en français).
- La présence inégale du **lemme** ou du **lexème**
  Le **lemme** est une abstraction sur l'ensemble des formes fléchies d'un même *mot* qui ne diffèrent que par la morphologie flexionnelle. Le **lexème** (pas nécessairement unique par *mot*) désigne le morphème lexical contenu dans ce *mot*. Le lemme n'est donc pas défini en fonction du sens contrairement au lexème.

Nous ré-exploitons la norme sur les structures de traits en cours de rédaction dans le *TC37/SC4* (ISO TC37/SC4 N033) pour l'annotation de l'étiquette morpho-syntaxique. La généricité des structures de traits permet de représenter le marquage morpho-syntaxique pour les cas les plus complexes. Une étiquette morpho-syntaxique correspond à un ensemble de propriétés linguistiques qui peut se représenter par une liste de couples $(attribut, valeurs)$. [4]

[4] La référence doit être plus claire dans le texte, par exemple au sujet des bibliothèques

```
<wordForm entry="prime minister" tokens="t1 t2">
<fs>
      <f name="part of speech">
         <sym value="noun"/>
      </f>
      <f name="gender">
         <sym value="masculine"/>
      </f>
      <f name="number">
         <sym value="singular"/>
      </f>
</fs>
</wordForm>
```

L'usage est cependant de proposer une étiquette compacte et mnémonique. La norme *FS* offre la possibilité de noter une telle forme compacte pour une structure de traits complexe :

```
<fvLib>
      <sym id="noun" value="noun"/>
      <sym id="sing" value="singular"/>
      <sym id="masc" value="masculine"/>
</fvLib>
<fLib>
   <f id="pos@n" name="pos" fVal="noun"/>
   <f id="num@sing" name="num" fVal="sing"/>
   <f id="gen@masc" name="gen" fVal="masc"/>
</fLib>
<wordForm entry="prime minister" tokens="t1">
       <fs feats ="pos@n num@sing gen@masc"/>
</wordForm>
```

Le marquage linguistique d'une unité morpho-syntaxique peut alors se résumer en une *étiquette* simple donnée par l'attribut **tag** :

```
  <wordForm entry="prime minister" tokens="t1" tag="pos@n
      num@sing
  gen@masc"/>
```

Les propriétés linguistiques des traits morpho-syntaxiques seront définies dans la norme sur les catégories de données (réf.).

La possibilité est offerte d'utiliser le typage et les traits définis en suivant les recommandations de la norme (*Data Categories Register*).

etc.

Alternativement, il est possible d'utiliser le mécanisme de déclaration de bibliothèques et de définir, pour un usage spécifique, ses propres types, traits, et valeurs pour un trait donné.

Nous proposons un mécanisme de lien entre ces deux approches. L'usager de la norme peut suivre les recommandations générales sur les catégories de données pour l'ensemble de l'annotation morpho-syntaxique, et définir des bibliothèques spécifiques à un usage *privé* :

```
<fsmap>
<!-- private structure -->
  <fs>
  <f name="pos"><sym value="noun"/></f>
  <f name="kind"><sym value="numeral"/></f>
  </fs>
<!-- registered structure -->
  <fs><f name="pos"><sym value="numeral"/></f></fs>
</fsmap>
```

Compléter pour l'implémentation du mapping entre DCR et "private libraries" : le mécanisme générique que l'on souhaite appliquer.

3.1 Marque de la composition

"L'étiquette" morpho-syntaxique d'un terme composé peut contenir une partie de la description de l'analyse de la composition. Les étiquettes polycatégorielles peuvent par exemple remplir ce rôle :

```
<token value="pomme" id="t1"/>
<token value="de" id="t2"/>
<token value="terre" id="t3"/>

<wordForm entry="pomme de terre" tag="Nom + Préposition + Nom"
tokens="t1 t2 t3"/>
```

Mais nous pouvons envisager que la composition morphologique d'un terme s'analyse en termes d'***unités morpho-syntaxiques***. Il est alors naturel de décrire cette analyse compositionnelle comme contenu de l'élément <**wordform**> :

```
<token value="pomme" id="t1"/>
<token value="de" id="t2"/>
```

```
<token value="terre" id="t3"/>

<wordForm entry="pomme de terre">
    <wordForm entry="pomme" tag="Nom" tokens="t1"/>
    <wordForm entry="de" tag="Préposition" tokens="t2"/>
    <wordForm entry="terre" tag="Nom" tokens="t3"/>
</wordForm>
```

Résumé : élément <word-form>, suite et fin

- l'attribut **tag** a comme valeur une étiquette qui définit le marquage morpho-syntaxique de l'élément.
- Le contenu de <**word-form**> est une structure de traits qui fait référence à un unique registre de catégories de données (par mécanisme de *Name Space*) et/ou à une bibliothèque définie par l'usager.
- L'élément <**word-form**> d'une forme composée peut contenir récursivement une définition des composants par enchâssements successifs.

4 Annotation de l'ambiguïté

4.1 Ambiguïté de segmentation

Une séquence de ***tokens*** correspond à un ensemble de ***word-forms*** qui se définissent comme unités syntagmatiques. Il est indispensable de pouvoir annoter les ambiguïtés entre mots simples et mots composés ou entre différentes compositions possibles de ces unités pour être complet.

Le graphe connexe non-cyclique est la représentation minimale qui permet d'encoder toutes les alternatives possibles de façon non déterministe. Il permet de représenter les ambiguïtés entre formes simples et formes composées d'une séquence linéaire de mots et les sommets du graphe représentent les transitions entre les mots. Cette représentation pourra être augmentée pour encoder d'autres types d'informations : probabilité d'une composition relativement à un autre, description de l'information linguistique qui sépare deux mots, etc.

Nous proposons cependant une notation minimale de la composition qui n'est qu'une possibilité offerte pour les applications réclamant une représentation de l'ambiguïté (analyse automatique non déterministe, représentation d'un état non corrigé de l'annotation, etc.)

Il est toujours possible de se passer de cette représentation pour les cas simples, et de noter les *word-forms* selon leur succession naturelle sans que

les alternatives entre mot simples et composés soient jamais notées. Dans ce cas, les mots se suivent dans le document sans aucune autre marque d'ordre linéaire :

```
<token form="fer" id="t1">fer</token>
<token form="à" id="t2">à</token>
<token form="cheval" id="t3">cheval</token>
<wordForm entry="fer" tokens="t1"/>
<wordForm entry="à" tokens="t2"/>
<wordForm entry="cheval" tokens="t3"/>
```

Les séquences sont alternativement représentés par des transitions dans un graphe connexe non-cyclique (*DAG*). Les sommets du graphe représentent les séparations entre les **word-forms** (rappelons qu'il ne s'agit pas nécessairement de marques typographiques et que les séquences de **word-forms** ne sont pas toujours linéaires). Chaque transition contient un ou plusieurs ***word-forms*** reconnus comme une unité suffisante pour recevoir une marque morpho-syntaxique.

Pour la séquence "*fer à cheval*", dont le caractère figé de la composition n'est pas exprimé de façon définitive ici, nous notons le DAG suivant :

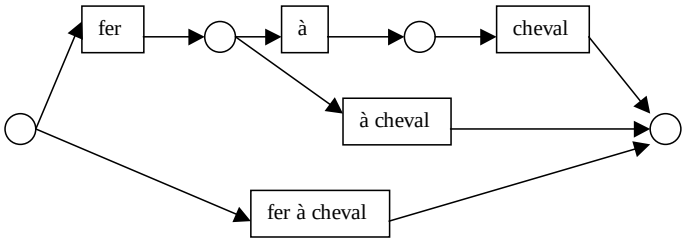

FIG. 1 – DAG de fer à cheval

Ce *DAG* représente à la fois la séquence de mots simples "fer", "à", "cheval", et de mots composés "fer à cheval", "à cheval". Il peut être représenté en *XML* de la façon suivante :

```
<token value="fer" id="t1">fer</token>
<token value="à" id="t2">à</token>
<token value="cheval" id="t3">cheval</token>
<state id="S0" type="initial"/>
<state id="S2"/>
<state id="S3" type="final"/>
<transition source="S0" target ="S3">
   <wordForm entry="fer␣à␣cheval" tokens="t1␣t2␣t3"/>
</transition >
```

```
<transition source="S0" target ="S1">
    <wordForm entry="fer" tokens="t1"/>
</transition >
<transition source="S1" target ="S2">
    <wordForm entry="à" tokens="t2"/>
</transition >
<transition source="S2" target ="S3">
    <wordForm entry="cheval" tokens="t3"/>
</transition >
<transition source="S1" target ="S3">
    <wordForm entry="à cheval" tokens="t2 t3"/>
</transition >
```

Les unités linguistiques (*entry*) “fer à cheval”, “fer”, “à”, “cheval” et “à cheval” correspondent bien à des unités syntagmatiques minimales sur lesquelles porte l'annotation.

Compléter pour l'implémentation des alternatives

Résumé : élément <fsm>

- L'élément <**fsm**> (Finite State Machine) implémente un graphe connexe acyclique pour une séquence de <**word-form**>.
- L'élément <**state**> implémente un sommet du graphe.
  L'attribut **type** permet de spécialiser un état **initial** et un état **final**.
- L'élément <**transition**> implémente une transition du graphe entre deux sommets (attributs **from**, **to**.
  Le contenu d'une transition correspond à un ou plusieurs <**word-form**>