RNIL – ISO.TC37.SC4.WG1 morpho-syntaxe

Lionel Clément

Projet Atoll – INRIA Rocquencourt

Lionel.Clement@inria.fr

mardi 10 juin 2003

Appel à contribution en morpho-syntaxe

- Orientations et choix généraux

  (nous partons de la présentation de Éric de la Clergerie du 7 avril)

- Quelques contributions et documents exploitables
- Convergence ?
- Susciter la discussion autour de cette proposition

Annotation morpho-syntaxique

Articulation **segmentation/étiquetage**
(syntagmatique ⇔ associatif)

- Définir des unités morpho-syntaxiques
  Pas de prise de position a priori : ≠ typographie,
  ≠ morphologie, ...
- Leur assigner une marque :
  * Un type (catégorie)
  * Des propriétés propres à ce type (distributionnelles, morphologiques, ...)
  * Un rapport de cette unité au sens ou au lexique (lemme, lexème, ...)

Segmentation - 1

Intervalles

- Le document est un **flux** (documents sonores, *multimédia*, textes indexés, . . .)
  ⇒**Réutilisation de normes existantes**
- ◇ Segmentation = partitionnement du document ? (parties non annotées, chevauchements possibles)

Segmentation - 2

Token

intervalle $\neq$ *token* (ancrage morpho-syntaxique)

- ◇ abréviations, signes brachygraphiques, erreurs typographiques, variations typographiques, ...
- ◇ intervalles contigus ? (**afin**, seulement **de**)
- ◇ intervalles uniques ? (alternatives de segmentation)

Exemples :

* ***etc.***

  ${}_0\text{etc}_{1\cdot 2} < 0 \ldots 2 > < 1 \ldots 2 >$

  <token form="etc." ranges="0_2"/>

  <token form="." ranges="1_2"/>

* ***afin, seulement de***

  ${}_0\text{afin}_1\text{,}_2\text{seulement}_3\text{de}_4$

  <token form="afin de" ranges="0_1 3_4"/>

  <token form="," ranges="1_2"/>

  <token form="seulement" ranges="2_3"/>

  Alternative : 2 *tokens* pour un *mot*

  (mais pbl de représentation des séquences des mots)

Articulation entre *tokens* et ≪*mots*≫ (*w*)

Mot = Transition entre deux sommets d'un *DAG*

- 1 séquence de **tokens** ⇔ 1 étiquette (entrée de lexique, unité distributionnelle, ...)
- Plusieurs transitions possibles pour la même séquence de tokens :
  - Alternative mot simple/mot composé
  - Alternative entre unités (≠ paradigme de chaque unité)
- Alternative d'étiquetage sur chaque transition (<alt >)

Un *mot* correspond à

- 1 token

```
<token form="vache" id="0">vache</token>
<w entry="vache" tokens="0"/>
```

- plusieurs tokens

```
<token form="à" id="0"/>
<token form="lequel" id="1">Auquel</token>
<w entry="auquel" tokens="0 1"/>
```

- plusieurs transitions sur un même token

```
<token form="à lequel" id="0">Auquel</token>
<w entry="à" tokens="0"/>
<w entry="lequel" tokens="0"/>
```

Graphe Connexe Acyclique

Ensemble de sommets et d'arcs

1. <state id="0">
   <transition target="1">
   ...
   </transition>
   </state>
2. Sans balise *state*
   <transition source="0" target="1">
   ...
   </transition>

   Mais :

   – Traitement automatique pas facilisé
   – Pas de décoration des états (marques inter-mots)

Un exemple d'application : tokenizer

Texte *brut* → Document XML

1. Segmentation par expressions régulières sur le texte
   - Chevauchements possibles ***etc.*** = ***etc.*** + ***.***
   - Éléments non segmentés (espaces, séparateur, . . .)
   - Segments vides ***auquel***= ***auquel*** + ∅

   ```
   <token id="t1" form="etc.">etc.</token> <token id="t2" form="."/>
   ```

2. Unités morpho-syntaxiques (quelques ajustements typographiques)
   - **-t-il** → **il**, **l'on** → **on**

     ```
     <token form="on" id="t1">-t-on</token>
     ```

3. Construction du graphe connexe
   - 1 arc du graphe = une entrée lexicale (par concaténation de *tokens*)

Marque morpho-syntaxique

- S'appuie sur norme FS (structure de traits)
- Alternatives possibles (élément **alt**)
- Forme compacte par bibliothèque de valeurs (attribut de **w**)

```
<transition source="0" target="1">
    <w entry="pomme de terre" tokens="t1 t2 t3" feats="pos@noun num@sg
gend@fem" / >
        </w></transition>
```

- Forme pleines (contenu de **w**)

```
<transition source="0" target="1">
    <w entry="pomme de terre" tokens="t1 t2 t3">
        <fs>
            <f name="pos"><sym value="noun"/></f>
            <f name="num"><sym value="sing"/></f>
            <f name="gen"><sym value="fem"/></f>
</fs></w></transition>
```

Contributions

- Manifestations
  - STTS (Stuttgart-Tuebingen Tagset)
  - CSAI University of Malta (Albert Gatt)
  - Japon
- Les ressources disponibles :
  - Tagsets
    - STTS
    - ELM-DE
    - TIGER
    - Verbobil
    - Multext
    - Eagle
    - Corpus Arboré Paris 7
    - ...
  - Segmentation
    - STTS
    - Corpus Arboré Paris 7
    - ...

Adéquation avec besoins manifestés – 1

Étiquetage : les besoins

- Liste plus ou moins classique des parties du discours ou catégories distributionnelles

  Présence de catégories hétérogènes (nombres, ponctuations, mots étrangers, abréviations, résiduels, classe unique, ...)
- Sous types très variés (catégories distributionnelles, sémantique, valence, ...)
- Marques morphologiques
- Présence inégale du **lemme**, **lexème**

Quelques Parties du discours

Multext	Eagles	STTS	Malte	P7
adj	adj	adj	adj	adj
adposition	adposition	adposition	adposition	
adv	adv	adv	adv	adv
art	art	art	art	
conj	conj	conj	conj	conj
dét	dét			dét
int	int	int	int	int
nom	nom	nom	nom	nom
num		card	num	
pro	pro	pro		pro
résid	résid			
uniq	uniq		uniq/non ass.	
verbe	verbe	verbe	verbe	verbe
abrév.				
		étranger		étranger
		part		
		préf		
		signe		
		ponct	ponct	ponct
			pro/dét	
				prép
				clitique

Adéquation avec besoins manifestés – 2

Segmentation

- Très généralement séquence
  $\text{token}_1/\text{tag}_1\ \text{token}_2/\text{tag}_2\ \ldots$
- Parfois :
  - organisation arborescente
    mot composé $< \{\text{composant}_1\ \text{composant}_2\ \ldots \text{composant}_k\}$
  - Éclatement d'un *token* sur deux *tags*
    $\text{token}_1\$\text{tag}_1\ \text{token}_1\$\text{tag}_2$ ou $\text{token}_1/\text{tag}_1{+}\text{tag}_2$
  - Plusieurs *tokens* sur le même *tag*
    $\text{token}_1\$\text{tag}_1\ \text{token}_2\$\text{tag}_2\ \ldots \text{token}_k\hat{}\text{tag}_1$
  - Plusieurs *tokens* qui doivent être recombinés
    $\text{token}_1\%\text{tag}_1\ \text{token}_2/\text{tag}_1$

Adéquation avec Besoins manifestés – 3

Étiquetage : les besoins

1. Parties du discours (catégorisation des mots du texte)
2. Sous-catégories (sémantiques, morphologiques, . . .)
3. Marques morphologiques (nombre, genre, cas, . . .)
4. Lemme, Lexème, . . .

◇ Typologie en fonction des langues & des projets

⇒ Pas de consensus sur les définitions

◇ Mais hiérarchie POS > sous-type > marques morpho.

⇒ Norme **FS** suffit pour encodage

\+ Typage des attributs

Adéquation avec Besoins manifestés – 4

Segmentation : les besoins (Corpus Paris 7 & STTS comme exemples)

- ↑ Plusieurs *tags* pour un m̂ *mot* (axe syntagm.) (STTS, P7)
- ↑ Plusieurs *mots* pour un m̂ *tag* (STTS, P7)
- ↑ Correction du *mot* (indication du *mot* bien formé) (STTS)
- ↑ Séquence vide (P7)
- ↑ Alternatives mot simple / mot composé (P7)
- ↑ Alternatives mot composé / séquence de composants (P7)
- ↑ Alternatives *tag*
- ↓ Discontinuité du *mot* (P7) (représentation d'un graphe)
- ↓ Correction de la segmentation (STTS) (e.g. si un espace manquait)

Adéquation avec Besoins manifestés

- Un noyau commun pour un répertoire de catégories ?
- Éléments d'annotation :
  - **token** Séquence vide, Correction du mot
  - **state** Marque inter mots (i.e. espaces manquant ou en trop)
  - **transition** Alternatives Mots simples/Mots composés, Composant/Composé, Ambiguïté d'étiquetage
  - **alt** Ambiguïté d'étiquetage
  - **w** Unité d'annotation (i.e. entrée lexicale)
  - **fs** Catégorie, sous-catégorie, marques morphologiques, lemme & lexèmes