ATOLL - INRIA Rocquencourt

http://atoll.inria.fr

Activités de RNIL / Normalangue

Éric de la Clergerie

Eric.De_La_Clergerie@inria.fr

Journée Normalangue

26 Septembre 2003

Objectifs de RNIL

Contribuer à définir des normes internationales dans le domaine de l'ingénierie linguistique au sein du comité ISO TC37/SC4, ainsi qu'à les valider et à les diffuser.

Objectifs de RNIL

Contribuer à définir des normes internationales dans le domaine de l'ingénierie linguistique au sein du comité ISO TC37/SC4, ainsi qu'à les valider et à les diffuser.

- **Participation** à la définition de normes internationales pour représenter et gérer des ressources linguistiques, sous la forme d'un groupe miroir français au TC37/SC4
- **Validation** des propositions normatives par le développement de bibliothèques informatiques (API) et production de jeux de test ;
- **Communication** vers la communauté nationale et francophone sur les travaux et les résultats obtenus au sein du TC37/SC4

Partenaires

Vaste consortium permettant une bonne expertise (sur ressources et outils) et une bonne diffusion.

Partenaires :

- **Académiques :** INRIA (2 Équipes), ATILF, LLF, TALaNa/Lattice, IRIN, LIMSI, Clips, RESO
- **Industriels :** CEA, XRCE, EDF R&D, Systran, EADS, Softissimo, Sinequa, Lucid-IT, J-Way
- **Associatif :** AFNOR (support dans les processus de normalisation)

RNIL comme miroir français de ISO TC37 SC4

- TC37 SC4 : Sous comité ISO en charge de la normalisation pour les ressources linguistiques
  Présidence française (Laurent Romary) ⇒ impose un rôle moteur

RNIL comme miroir français de ISO TC37 SC4

- TC37 SC4 : Sous comité ISO en charge de la normalisation pour les ressources linguistiques
  Présidence française (Laurent Romary) ⇒ impose un rôle moteur

- Suivi des initiatives SC4
  Votes, Examen des nouvelles propositions, désignation d'experts, participation aux évènements SC4

- Organisation d'évènements SC4
  Réunion plénière TC37SC4 en France en 2004
  Réunions techniques

- Proposition de nouveaux chantiers de normalisation

Organisation et communication

- Réunions régulières : 6 depuis Décembre 2002
  Effort pour être en phase avec les échéances ISO (réunions et votes)

- Support AFNOR pour la communication officielle
  Documents ISO, Convocations et comptes-rendus

- Site WEB `http://atoll.inria.fr/RNIL`
  Suivi des chantiers en cours, liens, documents (ISO, internes RNIL, externes), archives

- Liste électronique de discussion technique `rnil@inria.fr`
  Mise en place éventuelle d'autres modes de communication (plus ciblés, plus flexibles)

Groupes de travail

Mise en place progressive de plusieurs groupes de travail dans SC4
associés à des groupes de travail dans RNIL

WG1 Mécanismes de base pour les ressources linguistiques et terminologie du domaine

WG2 Schémas de représentation dont annotations (morpho-syntaxiques)

WG3 Représentation de données textuelles multilingues
Mémoires de traduction

WG4 Bases lexicales

WG5 Environnement de gestion de ressources lexicales

Groupes de travail

Mise en place progressive de plusieurs groupes de travail dans SC4
associés à des groupes de travail dans RNIL

WG1 Mécanismes de base pour les ressources linguistiques et terminologie du domaine

WG2 Schémas de représentation dont annotations (morpho-syntaxiques)

WG3 Représentation de données textuelles multilingues
Mémoires de traduction

WG4 Bases lexicales

WG5 Environnement de gestion de ressources lexicales

Actuellement, pour RNIL comme pour SC4, un seul réel groupe de travail (WG1)
avec émergence des autres groupes (WG2, WG3 et WG4)

Chantiers SC4 et RNIL

En cours :

DCR Data Category Registery (WG1)

LAF Linguistic Annotation Framework (WG1)

FSR Feature Structure Representation (WG1)

MSAF Morpho-Syntactic Annotation Framework (WG2)

Potentiels SC4 et émergents RNIL :

Lexiques adaptés au TAL (WG4)

Mémoires de traduction (WG3)

Références Gestion des annotations référentielles (WG2)

Chantiers SC4 et RNIL

En cours :

DCR Data Category Registery (WG1)

LAF Linguistic Annotation Framework (WG1)

FSR Feature Structure Representation (WG1)

MSAF Morpho-Syntactic Annotation Framework (WG2)

Potentiels SC4 et émergents RNIL :

Lexiques adaptés au TAL (WG4)

Mémoires de traduction (WG3)

Références Gestion des annotations référentielles (WG2)

Souhait de voir un **déploiement** sur le terrain de ces propositions pour validation et retour.

LAF – Cadre pour l'annotation linguistique

- Proposer un cadre générique pour l'annotation linguistique
  Qu'est ce qu'une annotation ? Sur quoi porte-t-elle ? Liens avec les documents annotés ?

- Proposition de Nouvel Item de Travail au sein de SC4
  (Vote Décembre 2003)

- Deux séminaires (Pont-à-Mousson 2002, Sapporo 2003),
  énonçant des grands principes guidant nos propositions

LAF – Cadre pour l'annotation linguistique

- Proposer un cadre générique pour l'annotation linguistique
  Qu'est ce qu'une annotation ? Sur quoi porte-t-elle ? Liens avec les documents annotés ?

- Proposition de Nouvel Item de Travail au sein de SC4
  (Vote Décembre 2003)

- Deux séminaires (Pont-à-Mousson 2002, Sapporo 2003),
  énonçant des grands principes guidant nos propositions
  - Utilisation technologies XML (XML, DTD, XML Schema, XSL, . . .)
  - Formats lisibles pour des humains et flexibilité (niveaux de complexité)
  - Prise en compte de mécanismes de conversion
  - Existence d'un format pivot pour l'interopérabilité
  - Prise en compte du flux dans les documents primaires
  - Prise en compte de aspects multimédia (adressage)
  - Identification et réutilisation uniforme de modèles génériques :
    structure de traits, structures de graphes (DAG), . . .
  - . . .

FSR – Structures de Traits

Les structures de traits [Feature Structure] sont un mécanisme très générique (neutre) en Linguistique pour décrire des paires (propriété,valeur).

- Déjà un chapitre sur la Représentation de FS (FSR) dans TEI (Text Encoding Initiative)
- Proposition de normalisation dans SC4, conduite par la Corée
  Nouvel item de travail
  Statut de « Committee Draft » [CD] au niveau ISO en attente de vote
- Réunion de coordination avec TEI (Nancy, Novembre 2003)
  Réunion technique en Corée (Février 2004)

FSR – Structures de Traits

Les structures de traits [Feature Structure] sont un mécanisme très générique (neutre) en Linguistique pour décrire des paires (propriété,valeur).

- Déjà un chapitre sur la Représentation de FS (FSR) dans TEI (Text Encoding Initiative)
- Proposition de normalisation dans SC4, conduite par la Corée
  Nouvel item de travail
  Statut de « Committee Draft » [CD] au niveau ISO en attente de vote
- Réunion de coordination avec TEI (Nancy, Novembre 2003)
  Réunion technique en Corée (Février 2004)

Au niveau français :

- Préparation de commentaires français (Eric de la Clergerie)
- Basés sur des travaux sur les Structures de Traits (INRIA Rocquencourt, LORIA)
  Outils, DTDs, Schéma XML
- Participation aux réunions techniques internationales

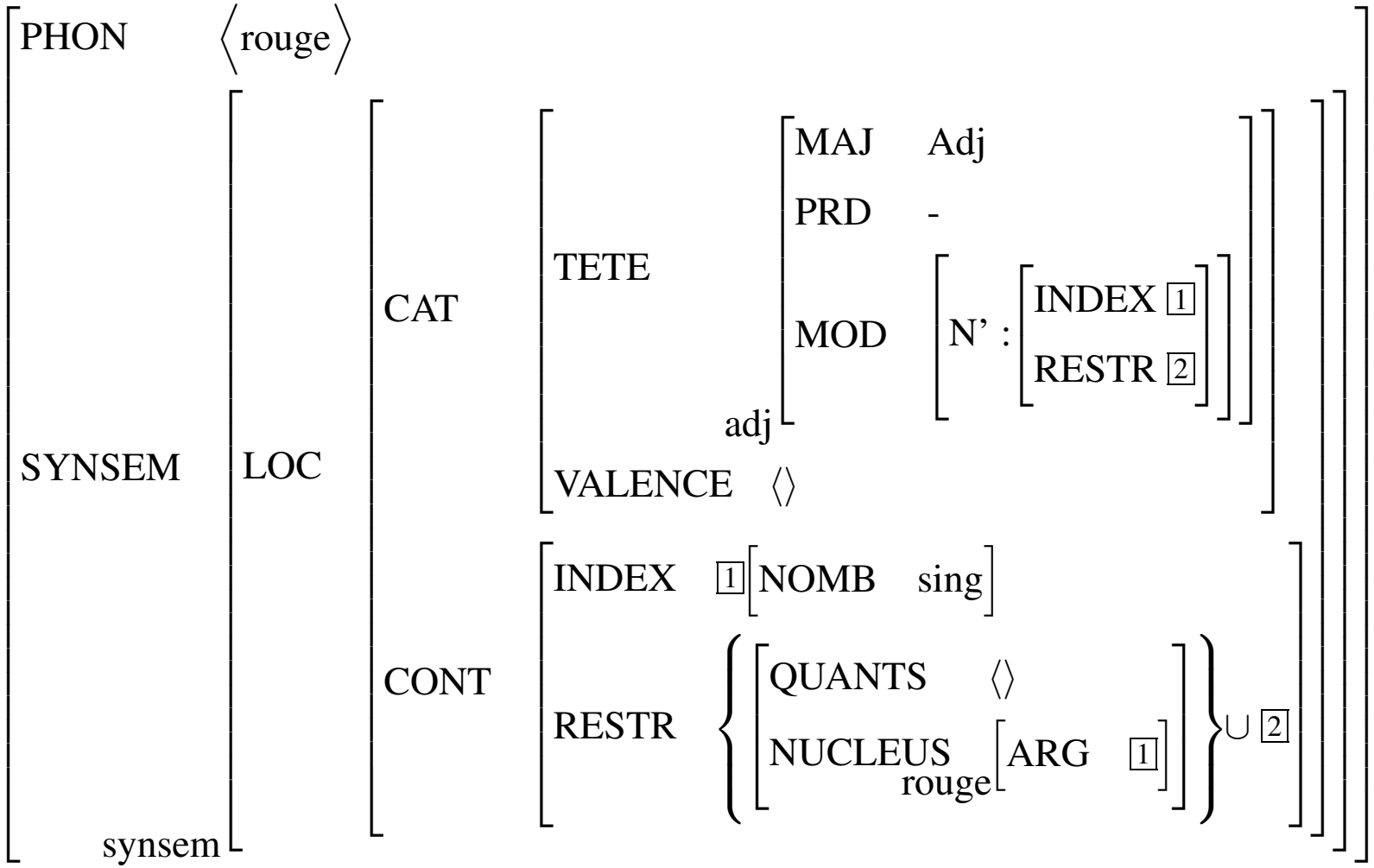
Structures de Traits – forme simple

Une simple liste de propriétés / valeurs atomiques,

number	sing
pers	3
mode	ind
tense	present
transitive	+
passive	-

```
<fs>
  <f name="number"><sym value="sing"/></f>
  <f name="pers"><sym value="3"/></f>
  <f name="mode"><sym value="ind"/></f>
  <f name="tense"><sym value="present"/></f>
  <f name="transitive"><plus/></f>
  <f name="passive"><minus/></f>
</fs>
```

FS complexe (Lexique HPSG)



DCR – Répertoire de Catégories de Données

- Mécanisme d'enregistrement de concepts linguistiques
  avec définition et liste éventuelle de valeurs attachées
  Équivalents et variations dans différentes langues

- Usage non normatif mais informatif

- Favoriser des références vers ces catégories dans les ressources linguistiques
  Par exemple, dans un jeu d'étiquettes morpho-syntaxiques pour le français

- Piloté par Laurent Romary (SC4 et RNIL).
  Vote « Committee Draft » en Novembre 2003

- Mise en place et premières expérimentations d'un répertoire au LORIA

- Réflexions dans RNIL sur l'utilisation de tels répertoires
  Enregistrement, Gestion, Exploitation

MSAF – Cadre d'Annotation Morpho-Syntaxique

- Initiative française lancée par RNIL
  Pilotée par Lionel Clément (IE financé sur RNIL) et Eric de la Clergerie
- Motivée par l'importance et la maturité des premiers niveaux de traitement linguistique (segmentation, étiquetage, morphologie, . . .)
  Diversité mais convergences possibles dans les outils et ressources

MSAF – Cadre d'Annotation Morpho-Syntaxique

- Initiative française lancée par RNIL
  Pilotée par Lionel Clément (IE financé sur RNIL) et Eric de la Clergerie
- Motivée par l'importance et la maturité des premiers niveaux de traitement linguistique (segmentation, étiquetage, morphologie, . . .)
  Diversité mais convergences possibles dans les outils et ressources
- Appel à contributions : jeux d'étiquettes, guides d'annotions, corpus annotés
  Quelques contributions : France, Allemagne, Malte, Japon
- MSAF accepté comme nouvel item de travail (NWI) dans SC4 (SC4 WG2)
- Première proposition de schéma de représentation
  présentée au sein de RNIL et en réunion TC37SC4 (Sapporo, Juillet 2003)
  Plutôt bien accueillie
- Rédaction progressive d'un « Committee Draft » pour vote
  Réunions techniques (Décembre et Février)
  Objectif : une décision lors de LREC (Mai 2004)
- Développement d'outils de validation
  Chaîne de traitement morpho-syntaxique pour le français ; accès serveur morpho-syntaxe

Morpho-Syntaxe : modèle

Proposition de modèle avec deux niveaux de structure et un niveau d'articulation

- Segmentation des données
  Indépendance du média, linéarité document, niveaux de granularité (mot, intra mots, groupe de mots)

- Représentation des « étiquettes » (tagging)
  S'appuie sur les Catégories de Données (DCR) et les Structures de traits (FSR)

- Automates finis reliant segmentation et étiquettes
  Gestion des ambiguïtés – Treillis de mots

Proposition de plusieurs variantes plus ou moins simplifiées
pouvant se ramener à la forme complète.

Le point sensible reste la définition des jeux d'étiquettes :

- Promouvoir l'utilisation de catégories enregistrées et définies
  Faciliter le référencement de ces catégories
- Laisser la possibilité d'utiliser son jeu d'étiquettes
- Travail d'examen de jeux existants pour repérer les convergences

Jeux d'étiquettes

Multext	Eagles	STTS	Malte	Paris 7
adj	adj	adj	adj	adj
adposition	adposition	adposition	adposition	
adv	adv	adv	adv	adv
art	art	art	art	
conj	conj	conj	conj	conj
det	det			det
int	int	int	int	int
noun	noun	noun	noun	noun
num		card	num	
pro	pro	pro		pro
resid	resid			
single	single		single/non ass.	
verb	verb	verb	verb	verb
abrev.				
		foreign		foreign
		part		
		pref		
		sign		
		punct	punct	punct
			pro/det	
				prep
				clitic

MSAF – Illustration

```
<token id="0">il</token>
<token id="1">sort</token>
<token id="2">les</token>
<token id="3">pommes</token>
<token id="4">de</token>
<token id="5">terre</token>
<token id="6">.</token>
```

```
<w tokens="0" entry="he" tag="pos@cl ␣ ..."/>
<w tokens="1" entry="dig_out" tag="pos@v ␣ ..."/>
<w tokens="2" entry="the" tag="pos@det ␣ ..."/>
<fsm>
    <state id="s1" type="init"/>
    <state id="s2"/><state id="s3"/>
    <state id="s4" type="final"/>
    <transition source="s1" target="s4">
        <w tokens="3 ␣ 4 ␣ 5" entry="potato" .../>
    </transition>
    <transition source="s1" target="s2">
        <w tokens="3" entry="apple" .../>
    </transition>
    <transition source="s2" target="s3">
        <w tokens="4" entry="from" .../>
    </transition>
    <transition source="s3" target="s4">
        <w tokens="5" entry="earth".../>
    </transition>
</fsm>
<w tokens="6" entry="dot" tag="pos@punct ␣ ..."/>
```

MSAF – Évolution

- Rédaction d'une proposition de « Committee Draft »
- Développement d'un schéma XML
- Développement de script de conversion (XSL) pour les variantes simplifiées
- Développement d'outils de validation
  Chaîne de traitement morpho-syntaxe en accès serveur
  Mise à disposition des outils de cette chaîne
- Poursuite du travail de comparaison et d'adéquation pour une large palette de langues
- Prise en compte des commentaires

Scénario de deploiement

- Coordination et synergie des divers axes.
- La Morpho-Syntaxe s'appuie sur les Structures de traits et les catégories
- Constitution d'un jeu d'étiquettes pour le français exploitant les catégories de données.
- Plongement du schéma d'annotation morpho-syntaxique dans LAF
  respect des principes LAF, possibilité de conversion vers le format pivot de LAF

Axes émergents

Suite à une liste de domaines évoqués lors de la dernière réunion TC37SC4 (Sapporo), mise en place de groupes de travail au sein de RNIL sur :

- Lexiques pour le TAL (Gil Francopoulo / INRIA)
  Peut s'appuyer sur DCR, MSAF, FSR

- Mémoires de traduction (Nasredine Semmar / CEA)
  Point de départ potentiel : Norme TMX (Translation Memory eXchange)

- Annotations référentielles (Susanne Salmon Alt / ATILF)

Interactions avec d'autres actions

- Doivent être renforcées !

Interactions avec d'autres actions

- Doivent être renforcées !

- Suivi de l'action EASY / EVALDA
  Participations croisées : Coordinateur Patrick Paroubek, équipes INRIA, . . .
  morpho-syntaxique ; aussi concerné par LAF

- Suivi de l'action INRIA SYNTAX
  Gestion de documents (dont extraction terminologique)

- Représentant RNTL Outilex (Eric Laporte)

Interactions avec d'autres actions

- Doivent être renforcées !

- Suivi de l'action EASY / EVALDA
  Participations croisées : Coordinateur Patrick Paroubek, équipes INRIA, . . .
  morpho-syntaxique ; aussi concerné par LAF

- Suivi de l'action INRIA SYNTAX
  Gestion de documents (dont extraction terminologique)

- Représentant RNTL Outilex (Eric Laporte)

- Visites dans les laboratoires et entreprises

Évolutions

- Poursuivre les chantiers en cours
  Mettre en place les chantiers émergents

- Mieux structurer en sous-groupes de travail
  mais garder une interaction forte entre groupes

- Faire émerger des outils, des ressources et des services WEB
  suivant la « stabilisation » des propositions de normes