Une introduction à XML pour les métiers de la traduction

Alexandre Roulois (Université Paris-Cité, IRIF, CNRS)

Table of contents

Décrire des données

> « Le concret, c'est ce qui est intéressant, la description d'objets, de paysages, de personnages ou d'actions ; en dehors, c'est du n'importe quoi. » (Claude Simon)

De l'interprétation

> « Et l'on peut me réduire à vivre sans bonheur,
> Mais non pas me résoudre à vivre sans honneur. » (*Le Cid* de Corneille, Acte II scène 1)

Point de vue humain

- suite de caractères
- structure complexe (mots, phrase, ponctuation)
- avec signification

Point de vue machine

- suite d'octets (…`0101 00110010 11011110 01100001 100001`…)
- structure simple (tout est octet)
- aucune signification

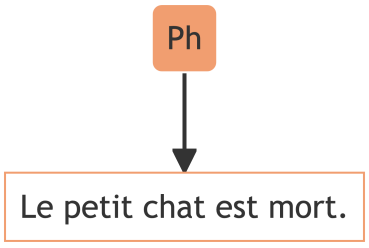
Question : Comment structurer l'information ?

Du point de vue sémantique, aucun octet n'est supérieur à un autre.

Exemple de l'analyse syntaxique de surface

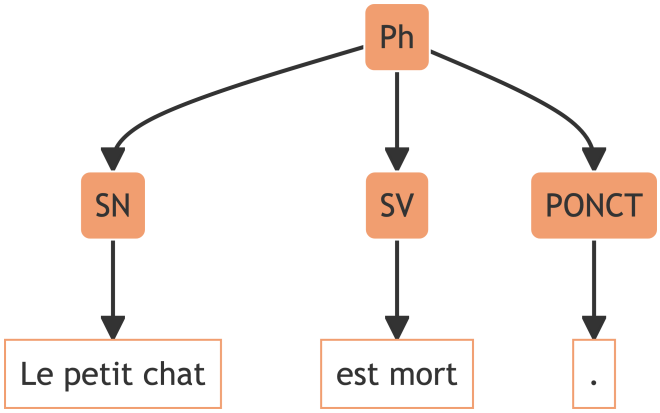
> « Le petit chat est mort. » (*L'école des femmes* de Molière, Acte II scène 5)

Représentation sous forme de graphe :



- message identique
- information supplémentaire par étiquetage

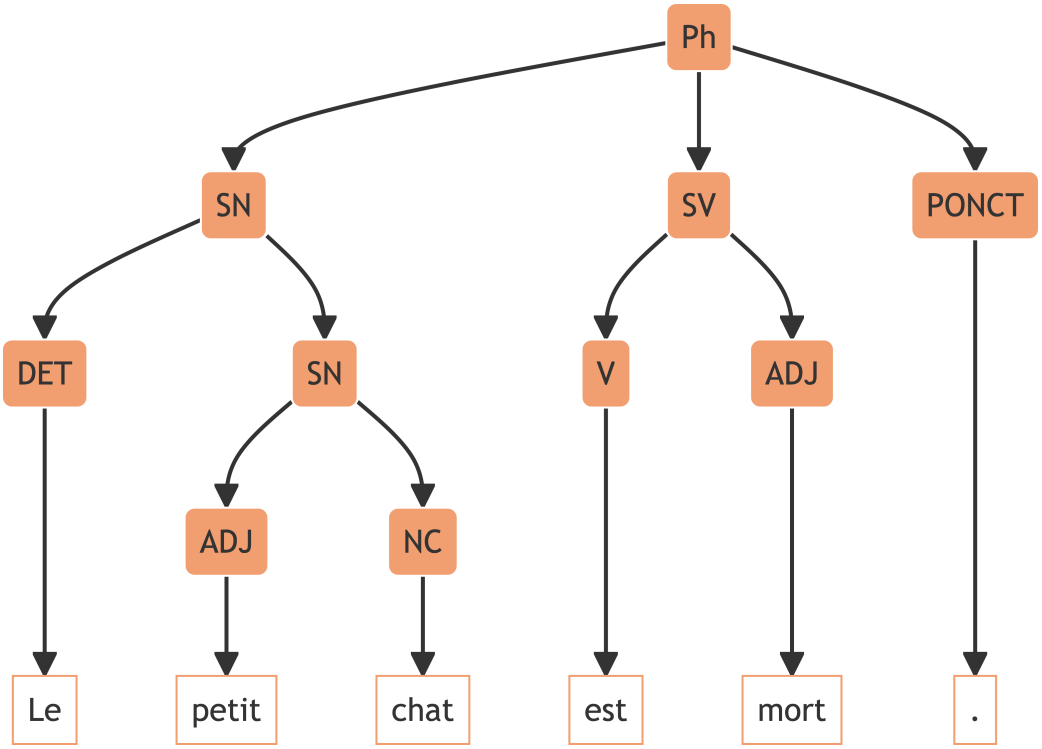
Granularité plus fine par segmentation en constituants :



- *Le petit chat* : syntagme nominal
- *est mort* : syntagme verbal
- *.* : ponctuation

Phrase identique mais davantage d'informations !

Segmentation en mots + étiquetage morphosyntaxique :



- phrase identique
- volume analyse > au volume phrase
- modélisation sous forme d'arbre

Pourquoi une analyse syntaxique ?

- répond au besoin de mettre en évidence la structure d'une phrase
- succède à une analyse lexicale
- préalable à une analyse sémantique

Face à une phrase, différents besoins :

- lire
- déclamer
- apprendre
- comprendre
- analyser
- ...

Observations

- agent humain effectue opérations selon ses besoins
- machine ne fait rien car aucun besoin

Questions

- pourquoi utiliser une machine ?
- comment traduire ses besoins à une machine ?

Analyse syntaxique pour décrire la structure d'une phrase :

- langage de description
- étiquettes (*tags*)
- vocabulaire (`SN V DET NC …`)
- grammaire :
  - $\mathrm{V} \not\supset \mathrm{SV}$
  - $\mathrm{V} + \mathrm{ADJ} = \mathrm{SV}$
  - $\mathrm{ADJ} \in \mathrm{SN}$ ou $\mathrm{ADJ} \in \mathrm{SV}$
  - …

Avec XML, processus identique !

Un langage de balisage

> « Oh ! Quel dommage Mme Chombier ! » (Jean-Marie Bigard, *La valise RTL*)

Objectif : organiser de façon logique et hiérarchisée un ensemble afin de faciliter l'accès à ses constituants

Problématique : un fichier étant une suite d'octets, comment en représenter la structure ?

Solution : utiliser un langage informatique de description

- balises (étiquettes de description)
- syntaxe (règles d'écriture)
- vocabulaire (libellé des étiquettes)

- grammaire (agencement des balises)

Description de la phrase *Le petit chat est mort* :

```
<!-- part-of-speech tagging -->
<Ph>
  <SN num="1">
    <DET>Le</DET>
    <SN num="2">
      <ADJ num="1">petit</ADJ>
      <NC>chat</NC>
    </SN>
  </SN>
  <SV>
    <V>est</V>
    <ADJ num="2">mort</ADJ>
  </SV>
  <PONCT>.</PONCT>
</Ph>
```

- balises pour étiqueter un segment : `<NC>chat</NC>`
- syntaxe : `<NC>` et non pas `?NC++`
- vocabulaire : `NC` plutôt que `Ceci est un nom commun`
- grammaire : $NC \notin V$

XML : *Extensible Markup Language*

- Standard du W3C : https://www.w3.org/TR/REC-xml/
- Méta-langage pour créer des langages de description :
  - formats personnalisés
  - formats normalisés (SVG, MathML…)
  - formats consensuels (TEI, TMX…)
- Souplesse accrue :
  - balises et attributs personnalisables
  - exploitations multiples
- Rigidité opportune :

– syntaxe stricte XML vs HTML permissif
– idéal pour conserver et échanger des données

Balise : dispositif de signalisation autour d'un segment

- syntaxe XML stricte
- étiquette d'un format défini ou personnalisé
- grammaire de validation normée, personnalisée ou absente

```
<!-- customised format -->
<Ph>Le petit chat est mort.</Ph>
<!-- HTML (standard) -->
<p>Le petit chat est mort.</p>
<!-- TEI -->
<sp>
    <l>Le petit chat est mort.</l>
</sp>
```

Formats HTML et TEI font référence à des grammaires :

- norme W3C pour HTML
- directives du consortium TEI pour le format TEI

Pour le format personnalisé, la grammaire est à créer !

La galaxie XML

« Il y a longtemps dans une galaxie lointaine, très lointaine… » (*Star Wars*)

Un écosystème de technologies

Pour produire, analyser, stocker…

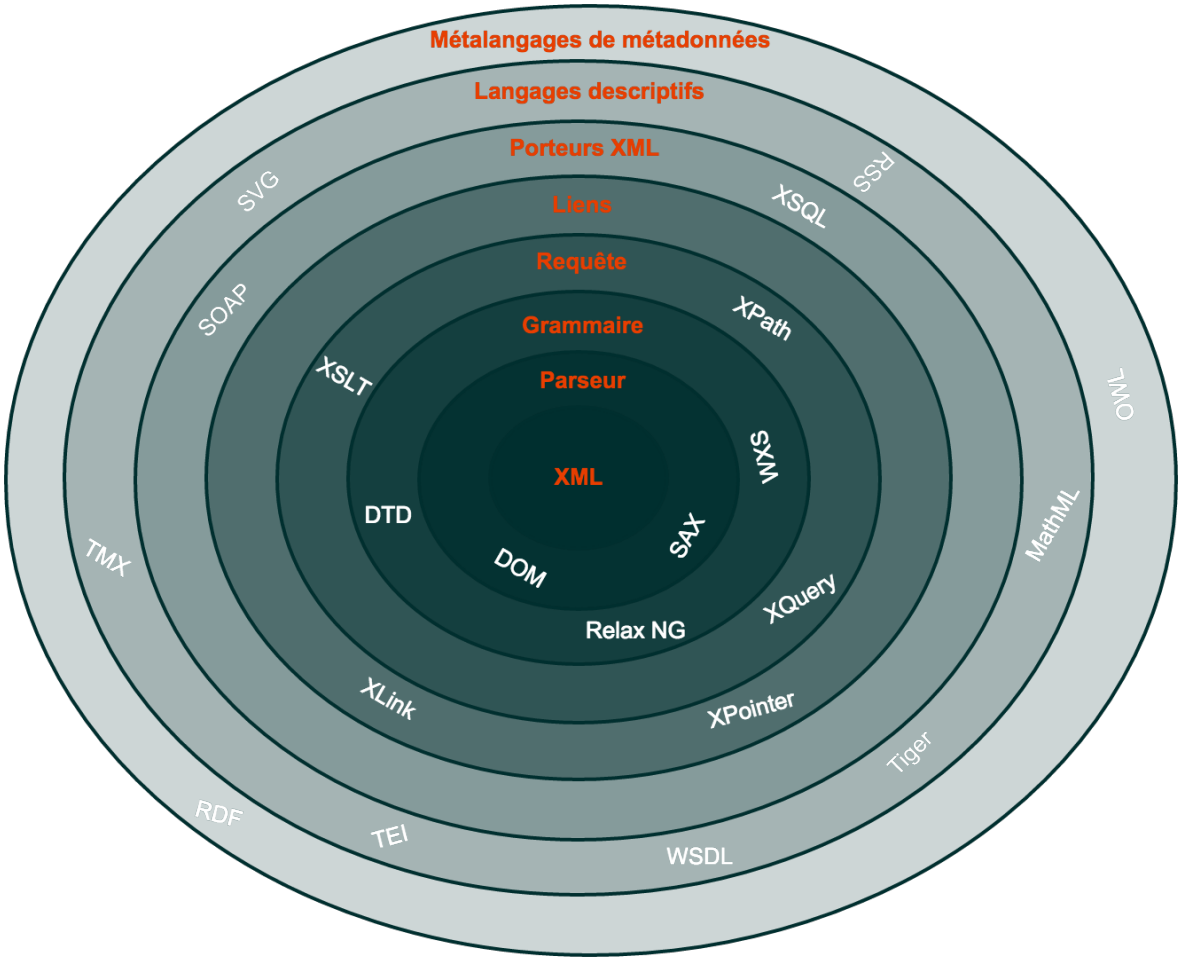


Figure 1: Représentation de la galaxie XML

Parseur : analyse syntaxique et modélisation du document XML

Grammaire : règles de composition d'un document XML

Requête : sélection et transformation de fragments XML

Liens : lier des fragments entre eux

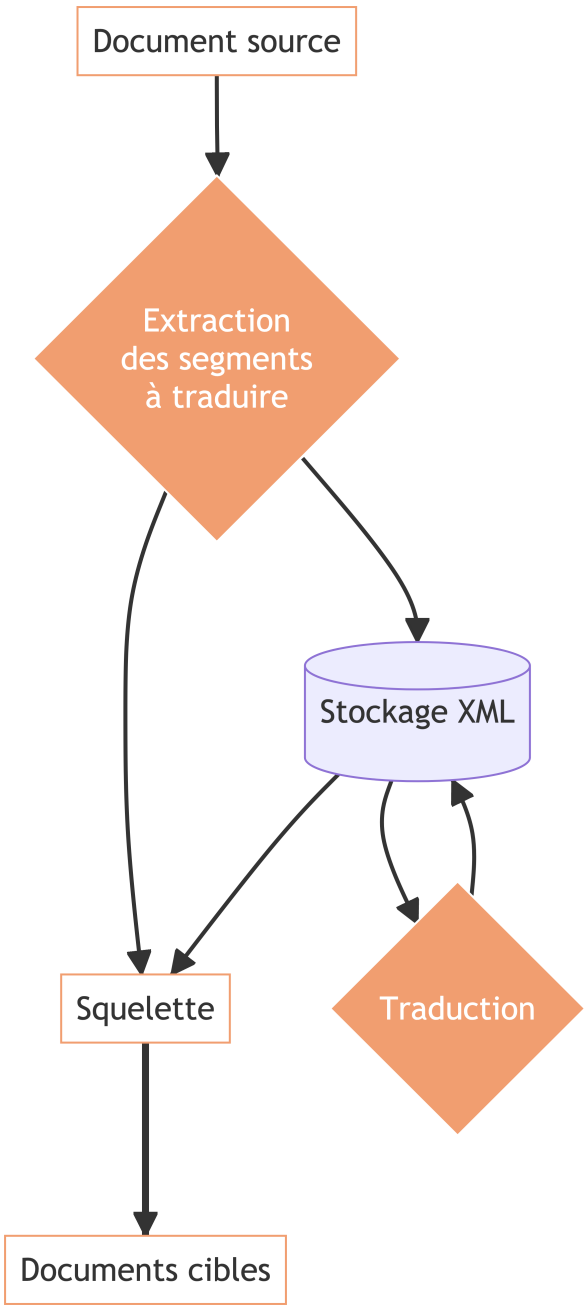
Porteurs : protocoles pour l'échange de données XML

Langages descriptifs : formats de données pour applications spécifiques

Métadonnées : décrire des langages descriptifs

Intérêts pour la traduction

- langage extensible
- règles syntaxiques strictes facilitent la structuration des données
- grammaires spécifiques :
  - TMX (mémoires de traduction)
  - TBX (glossaires)
  - XLIFF (organisation du flux de travail pour la localisation)
- intégration aux outils de TAO :
  - Trados
  - MemoQ
  - OmegaT
- organiser le traitement automatique des savoirs :
  - RDFS (Web sémantique)
  - OWL (ontologies)



Exemple de chaîne de production

Un gouvernement en perte de vitesse

Face à une série de crises politiques, économiques et sociales, le gouvernement semble perdre son influence et sa capacité à répondre aux attentes des citoyens.

Alors que la popularité des dirigeants s'effrite, les réformes peinent à s'imposer dans un climat de défiance généralisée. L'absence de solutions concrètes aggrave les tensions, laissant place à un sentiment croissant de déconnexion entre les gouvernants et les gouvernés.

1. 3 zones identifiées (titre, chapeau, texte)

```
% start heading
  unit_1 = Un gouvernement en perte de vitesse
% end heading
% start lead
  unit_2 = Face à une série de crises politiques, économiques et sociales, le gouvernement se
% end lead
% start text
  unit_3 = Alors que la popularité des dirigeants s'effrite, les réformes peinent à s'imposer
% end text
```

2. base de données :

```
<unit id="1">Un gouvernement en perte de vitesse</unit>
<unit id="2">Face à une série de crises politiques, économiques et sociales, le gouvernement
<unit id="3">Alors que la popularité des dirigeants s'effrite, les réformes peinent à s'impos
```

3. traduction :

```
<unit id="2">
  <variant lang="fr">Face à une série de crises politiques, économiques et sociales, le gouve
  <variant lang="en">In the face of a series of political, economic, and social crises, the g
</unit>
```

4. extraction des variantes d'une unité
5. production du document de sortie