Rabelais Tours

LI
Laboratoire d'informatique
EA 6300

Recherche d'information par cascade de graphes Unitex

Denis Maurel
Anubhav Gupta
Nathalie Friburger

LI
Laboratoire d'informatique
EA 6300

Présentation des outils

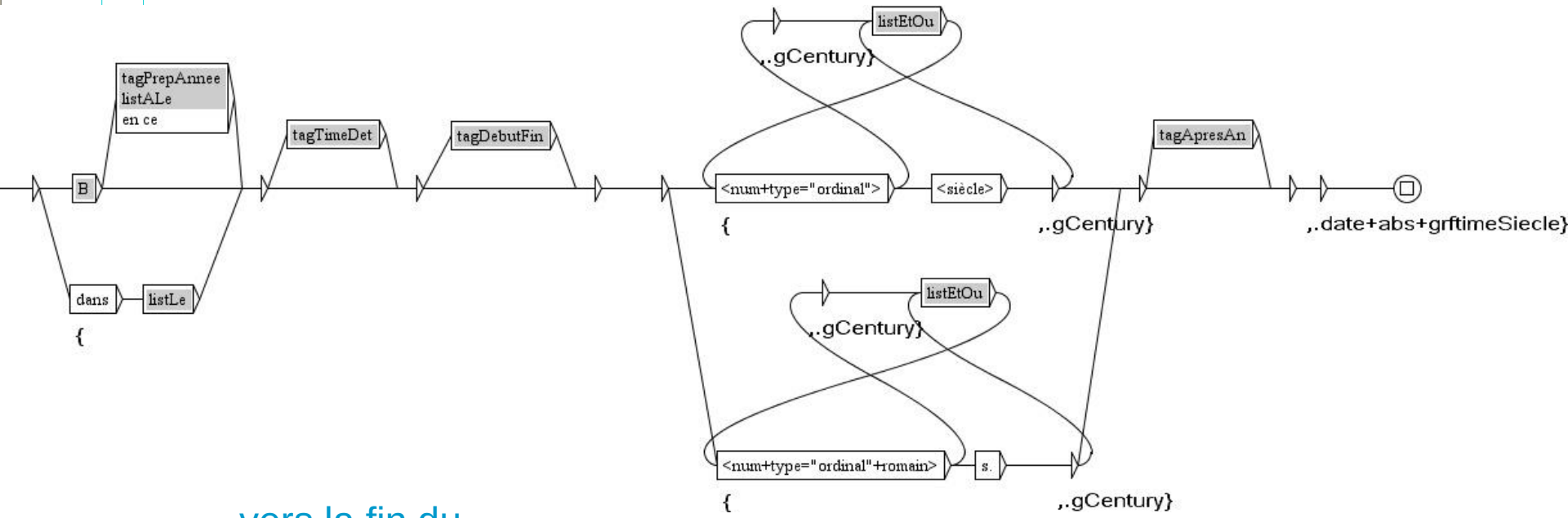
UNITEX ET CASSYS

Unitex

- Unitex est un logiciel libre d'analyse lexicale automatique
- Unitex allie
  - des réseaux de transitions "augmentées"
  - et une interface conviviale (des graphes)

Laboratoire d'informatique
EA 6300

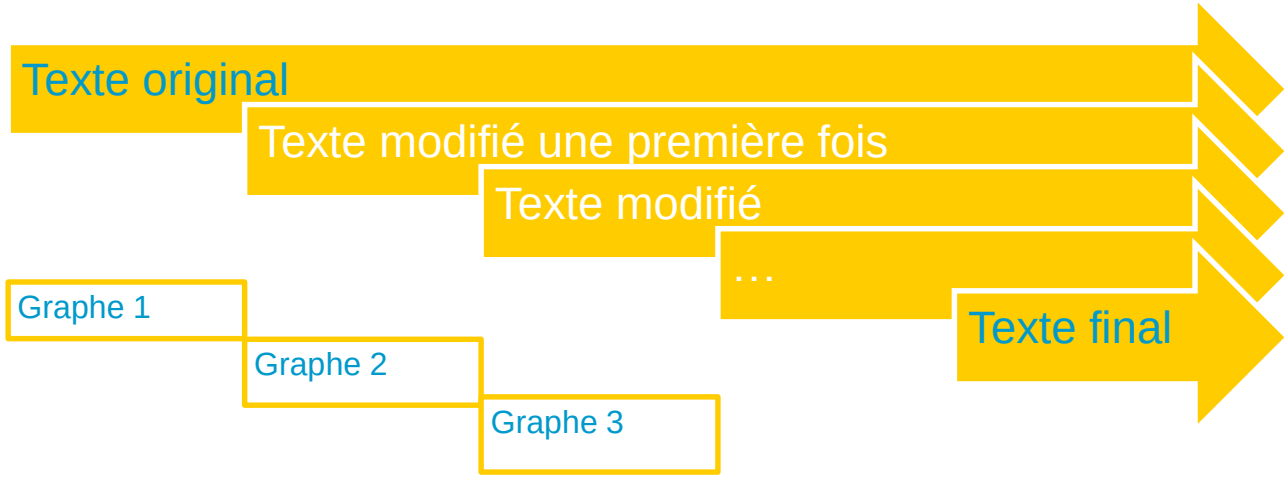
Un graphe Unitex



vers la fin du 20ème siècle de notre ère

CasSys

- Un module pour la constitution et l'utilisation de cascades de graphes intégré à Unitex



rue du
11 novembre 1918

15 millions de textes scientifiques à analyser

PROJET ISTEX

Recherche d'information

- Exemple

  The polyclonal antibody anti-c-Raf-1 (C-12, the epitope corresponding to the carboxy terminal amino acids of Raf-1 p74) was obtained from Santa Cruz Biotechnology (Santa Cruz, CA)

elsevier_raw_IST19970010000078_03037207_v135i2_S0303720797002049_main.xml

Recherche d'information

- Exemple

  The polyclonal antibody anti-c-Raf-1 (C-12, the epitope corresponding to the carboxy terminal amino acids of Raf-1 p74) was obtained from <orgName type="provider">Santa Cruz Biotechnology</orgName> (<placeName>Santa Cruz, CA</placeName>)

elsevier_raw_IST19970010000078_03037207_v135i2_S0303720797002049_main.xml

Recherche d'information

```
<standOff>
  <teiHeader>
    <fileDesc> … </fileDesc>
    <revisionDesc> … </revisionDesc>
  </teiHeader>
  <listAnnotation type="placeName" xml:lang="en">
    <annotationBlock corresp="text" xmls="https//www.tei-c.org/ns/1.0">
      <placeName change="#Unitex-3.1" resp="#istex-rd" scheme="http://orgName-entity.lod.istex.fr">
        <term>Santa Cruz, CA</term>
        <fs type="statistics">
          <f name="frequency">
            <numeric>1</numeric>
          </f>
        </fs>
      </placeName>
    </annotationBlock>
     …
  </listAnnotation>
  …
</standOff>
```

Premiers résultats

- Passage à l'échelle
  - 2 385 948 documents
    - temps d'exécution: environ une journée pour 500 000 documents
    - soit une moyenne de 0,17 s par document

Premiers résultats

- Création d'une cascade d'analyse pour l'anglais [améliorations en cours]
  - cascade de 55 graphes
  - tests réalisés sur 49 documents contenants 5 414 entités nommées

Premiers résultats

- Utilisation de la cascade d'analyse existante pour le français [reprise complète en cours]
  - cascade de 130 graphes
  - tests réalisés sur 40 documents contenants 4 695 entités nommées

Laboratoire d'informatique
EA 6300

Premiers résultats

		Anglais	Français
SER	$\frac{\#I + \#D + 0{,}5 * \#T + 0{,}5 * \#E + \#TE}{\#R}$	38,6%	34,9%
Rappel	$\frac{\#S - \#I}{\#R}$	55,7%	71,5%
Précision	$\frac{\#S - \#I}{\#S}$	91,5%	87,1%
Précision du typage	$\frac{\#S - (\#I + \#T + \#TE)}{\#S}$	85,6%	84,8%
Précision du balisage	$\frac{\#S - (\#I + \#E + \#TE)}{\#S}$	79,5%	80,2%

Merci !