FOCUS 1
Grammaires locales pour la détection d'entités nommées

Denis Maurel, Nathalie Friburger (LIFAT)

Plan

Phénomènes et données

Méthodes et outils

Phénomènes : entités nommées

Entités nommées : objets linguistiques à la base de la recherche d'information : termes désignant de manière univoque une élément de l'univers du discours

- nom propres ou pas (personnes, lieux, organisations, productions humaines...)
- descriptions définies (temps, montant, quantités, fonctions)

Exemple : L'iPhone 4 a été annoncé à la conférence du 7 juin 2010 par Steve Jobs, PDG de la compagnie américaine Apple. Il pèse 140g.

Forme linguistique : entités souvent polylexicales, potentiellement enchâssées

Exemple : le [président du [conseil de la [région Centre Val de Loire]]]

Phénomènes : entités nommées

Deux caractérisations

- détecter les entités nommées (délimitation des frontières, encapsulation)
- catégoriser les entitées nommées (type)

Exemples

- *PERS* — *Paris and Nicky Hilton sont les enfants de Kathy Hilton*
- *LOC* — *Paris est située en aval du confluent de la Marne et de la Seine*
- *ORG* — *Paris a réaffirmé sa fermeté dans les négociations du Brexit*

Phénomènes : données

- Campagnes d'évaluation **Ester2** et **ETAPE** de reconnaissance des entités nommées en français (système CasEN : 1° ETAPE) : corpus journalistique

  Résultat ETAPE - 1° *CasEN* (grammaires locales), 4° MXs (fouille de texte)

- Projet **Renom** : entités nommées dans des textes de la Renaissance
- Projet **Istex** : entités nommées dans des collections de textes scientifiques
- Projet **Abliss**: veille scientifique dans des articles en biologie systémique : recherche de protéines, gènes, etc. dans des descriptions d'expériences

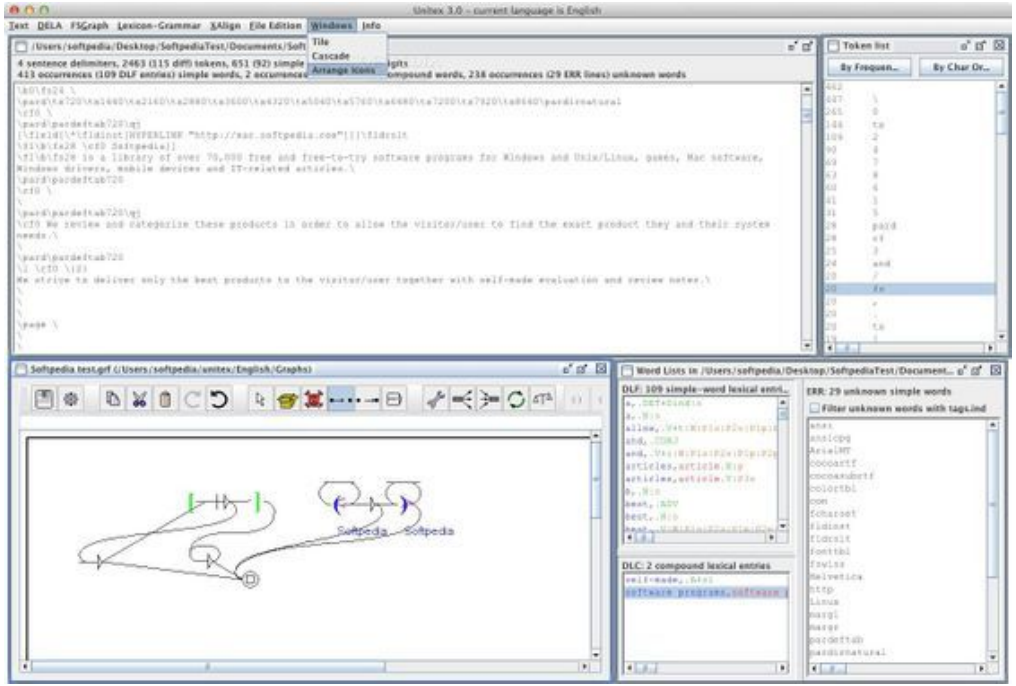
  Exemple : *phosphorylated ERK protein*

- Anonymisation de corpus

Méthodes et outils : grammaires locales

Description linguistique fine à l'aide de dictionnaires et de graphes définis par expertise linguistique

- **Analyse** - passage de **cascades de graphes** formalisés sous la forme de transducteurs : réseaux de transitions augmentés (**ATN**)
- Modélisation et traitement avec le logiciel **Unitex**

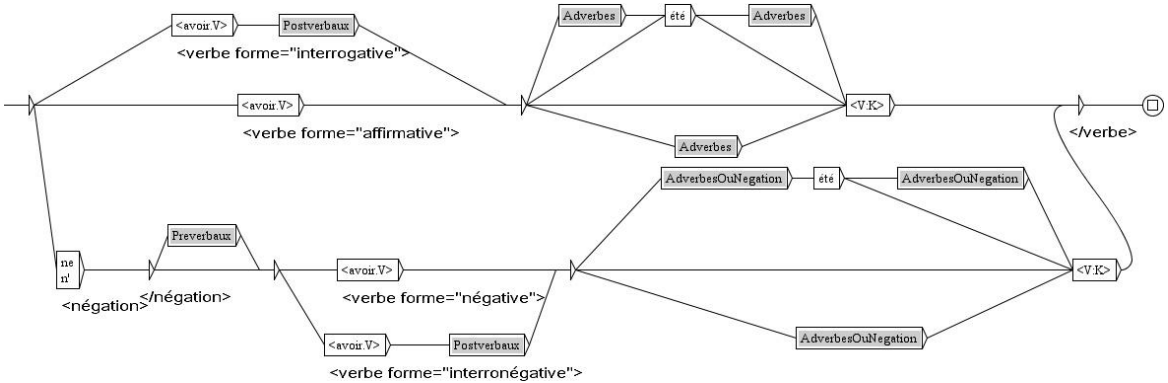


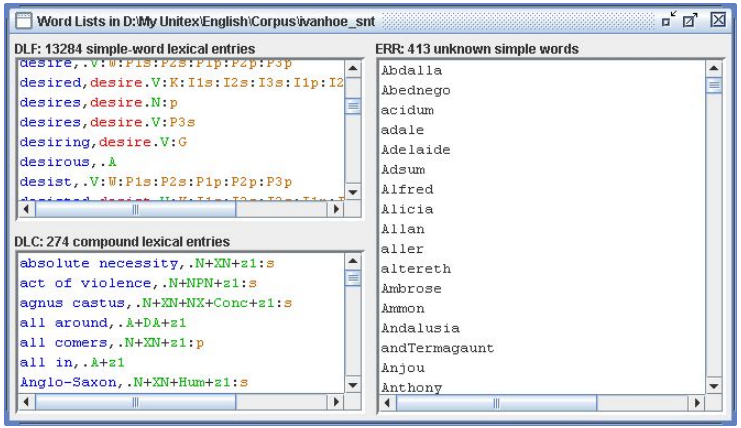
Méthodes et outils : Unitex

- Unitex est un logiciel libre d'analyse lexicale automatique auquel contribue le Lifat
- Unitex allie un système informatique performant : des réseaux de transitions "augmentées"
  - opérations sur le texte
  - utilisation de variables
  - compilation
  - cascades
- et une interface conviviale
  - dictionnaires
  - graphes

Méthodes et outils : Unitex

- Exemple de graphe : verbe *avoir* suivi d'un participe passé

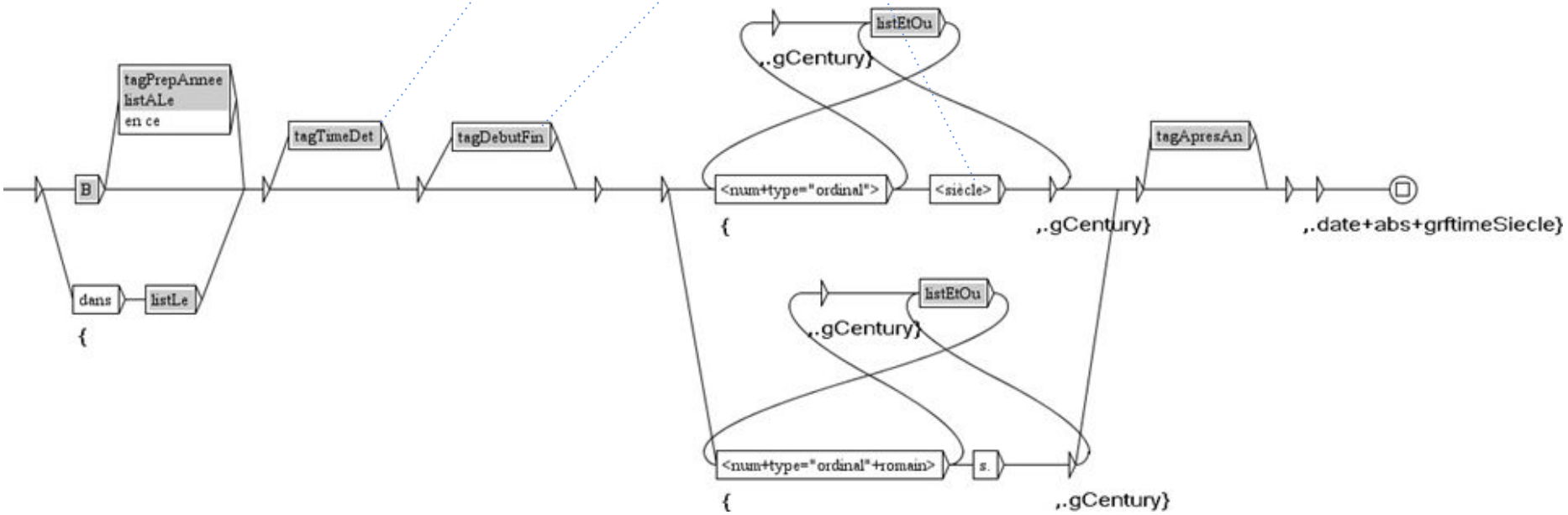




- Exemple de dictionnaire

Méthodes et outils : grammaires locales

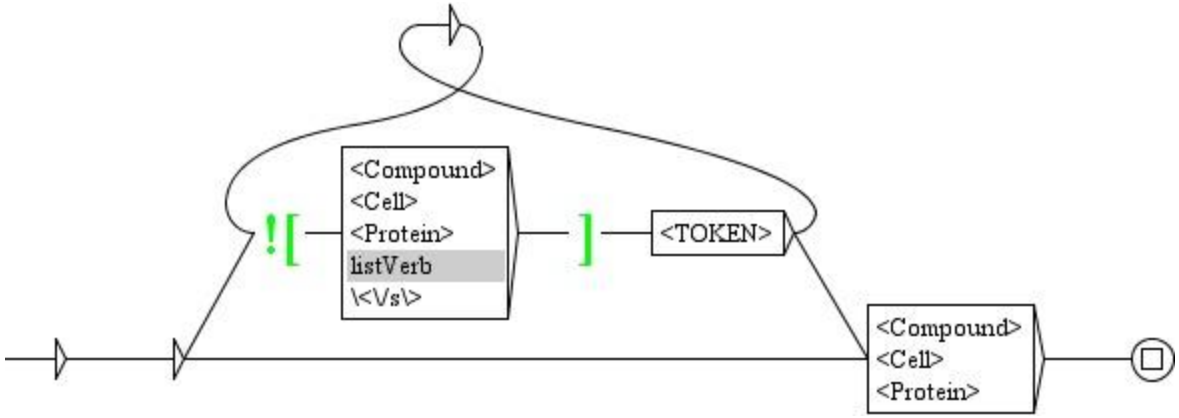
Exemple de graphe : *À peu près à la fin du XIXème siècle*



Méthodes et outils : grammaires locales

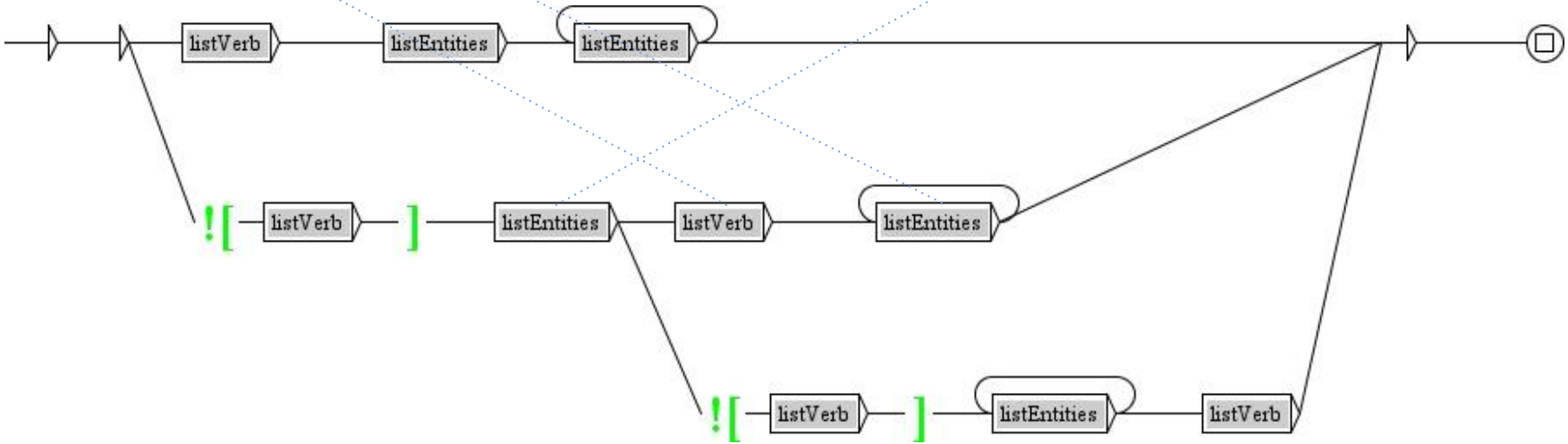
Exemple de graphe : La phrase *We found that only phosphorylated* **ERK** *protein bound to* **CDC25A** va être extraite pour un traitement ultérieur

1) Sous-graphe des entités (cellules, protéines…)



Méthodes et outils : grammaires locales

Exemple de graphe : La phrase *We found that only phosphorylated* **ERK** *protein* **bound** *to* **CDC25A** va être extraite pour un traitement ultérieur
2) graphe des descriptions d'expérience



Méthodes et outils : grammaires locales

Passage à l'échelle : projet Istex (français)

- 19 millions d'articles scientifiques traités
- 86 graphes dans la cascade, appelant 569 sous-graphes
- Dictionnaire français général : 682 000 entrées (DELA)
- Développement d'un dictionnaire de spécialité : 54 000 entrées

Passage à l'échelle : projet Istex (anglais)

- 19 millions d'articles scientifiques traités
- 64 graphes dans la cascade, appelant 305 sous-graphes
- Dictionnaire spécifique de spécialité : 52 000 entrées

Méthodes et outils : fouille de texte

Alternative - apprentissage automatique sur corpus annoté

- Doctorat de Damien Nouvel (2012)
- **Système mXs** : application à la REN de techniques de fouille de texte
- Fouille de motifs hiérarchiques séquentiels