université de TOURS

Des cascades Unitex pour la fouille bibliographique

Projet ANR Abliss sur les récepteurs biologiques

Denis Maurel, Sandy Chéry, Lifat

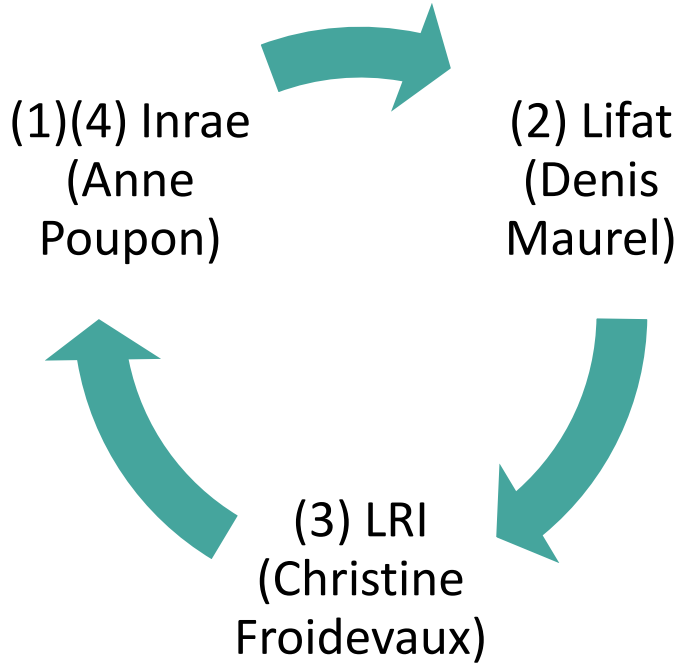
Présentation du projet Abliss

Présentation du projet

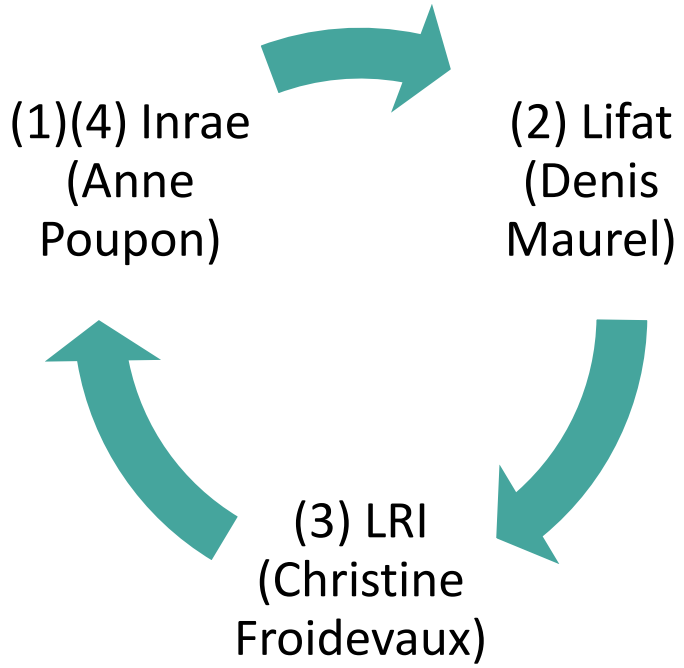
- **Projet interdisciplinaire avec des collègues biologistes**
  - Les récepteurs couplés aux *protéines G* sont des cibles de choix pour plus de 30% des médicaments
  - Mieux les connaître la signalisation permettrait
    - D'augmenter l'efficacité de ces médicaments
    - De diminuer les effets secondaires
  - Mais la littératures scientifiques est trop abondante
  - Et des résultats d'expériences réalisées restent cachés
    - Car ils ne concernent pas le sujet principal de l'article
    - Car ils ne figurent ni dans le titre, ni dans le résumé, ni dans les mots clés

Un cercle vertueux

(1) Décrire des expériences sous forme *prédicat-arguments*

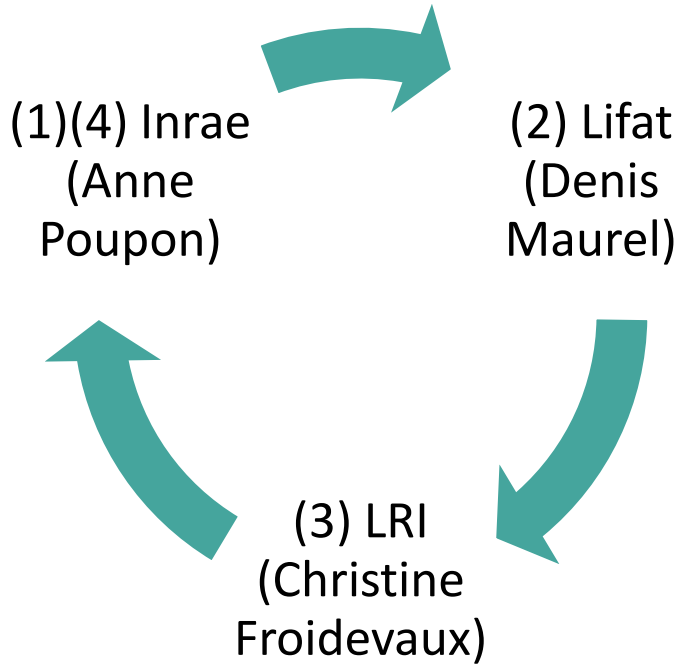


Un cercle vertueux



(2) Extraire des résultats expérimentaux dans les articles scientifiques par un module de traitement automatique des langues

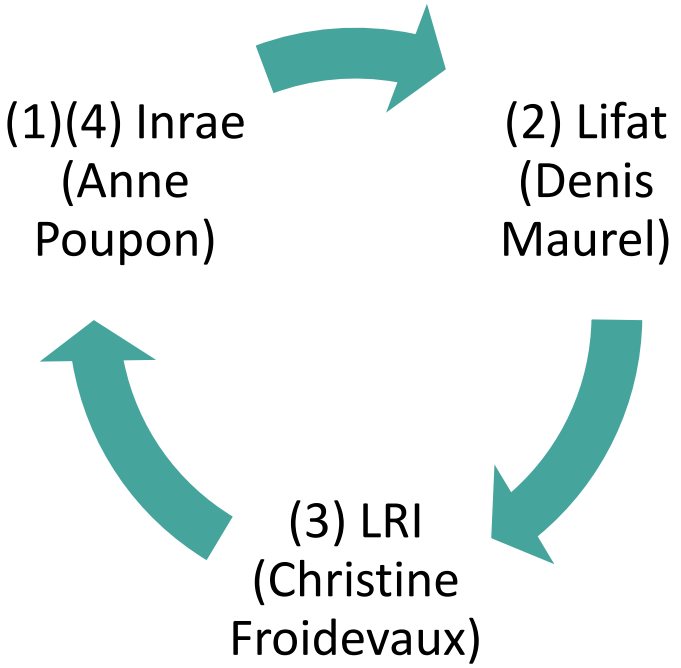
Un cercle vertueux



(3) Inférer les réseaux à partir de ces données par une méthode de raisonnement à base de connaissances

Un cercle vertueux

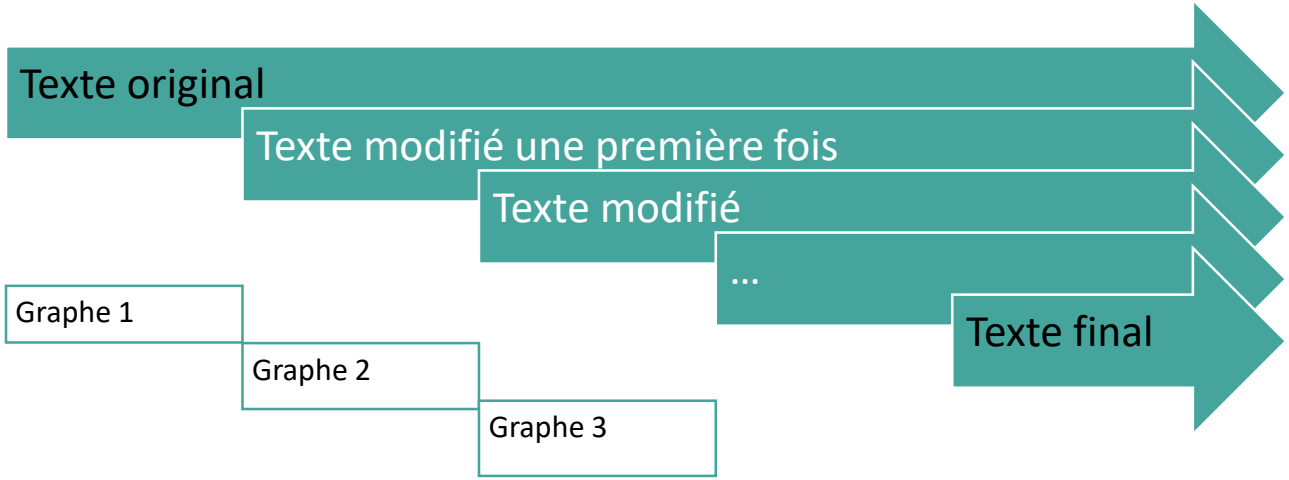
(4) Tester expérimentalement une proposition de réseau de signalisation



Détail de l'analyse

La méthode

- **Utilisation de cascades de graphes (CasSys)**
  - Il s'agit de créer un nouveau texte à partir du texte d'origine en le transformant par des graphes successifs, dans un ordre précis.



Le corpus

- **Un article scientifique en biologie**
  1. Résumé
     - principaux résultats
  2. Introduction
     - contexte de l'étude et état de l'art
  3. Matériel et méthodes
     - mise en pratique des expériences
  4. Résultats
     - liste et détail des expériences réalisées
  5. Discussion
     - interprétation des expériences et perspectives
  6. Conclusion

La chaine de traitement

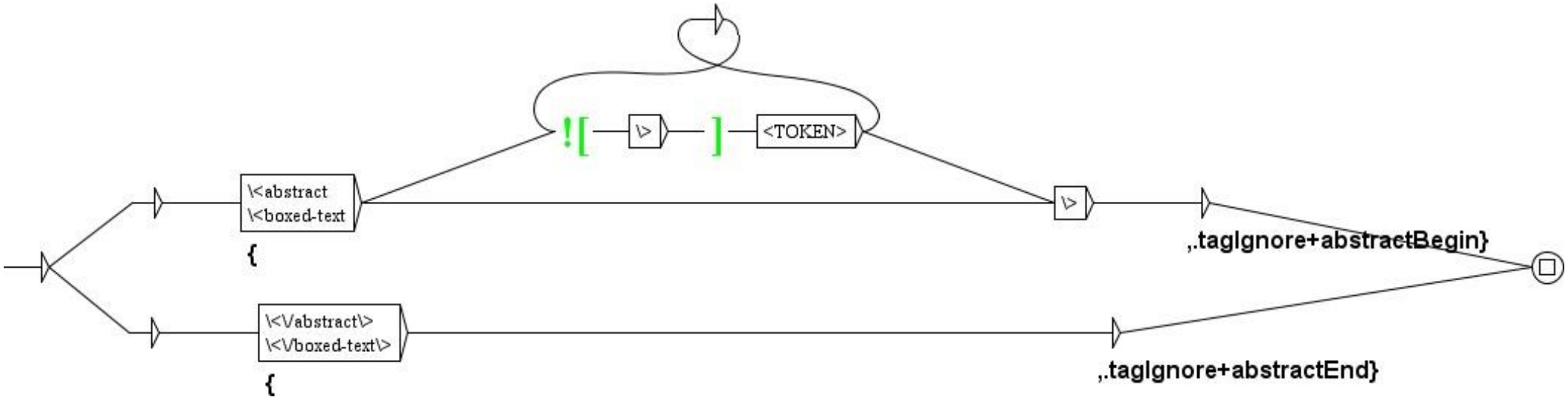
- **script PHP**
  - Interrogation de la base de publication PubMed
    - 3 mots clés (deux protéines et une réaction chimique)
    - -> liste des identifiants PMCID correspondant à la requête
  - Téléchargement de tous les articles à partir de la liste
  - Préparation de l'environnement des scripts Unitex
    - Remplacement des *LF* par des *CRLF*
  - Lancement de la 1ère cascade Unitex

La chaine de traitement

- **1ère cascade Unitex**
  - Cette cascade ne retient que la partie *Résultats* de chaque article
    - Utilisation d'un fichier de normalisation
      - Pour harmoniser les espaces, les tirets, les apostrophes, les points de suspensions
      - Pour transformer les codages de caractères HTML (par exemple &#x003B2 ; → β)
    - Utilisation du *Preprocessing* (en mode *Merge*) pour marquer le titre de la partie *Résultats* et le titre de la partie suivante
      - Remarque : le *Preprocessing* ne permet pas l'utilisation du début et de la fin du texte
    - Suppression du résumé, de la partie précédant la partie *Résultats* et de la partie suivant la partie *Résultats*

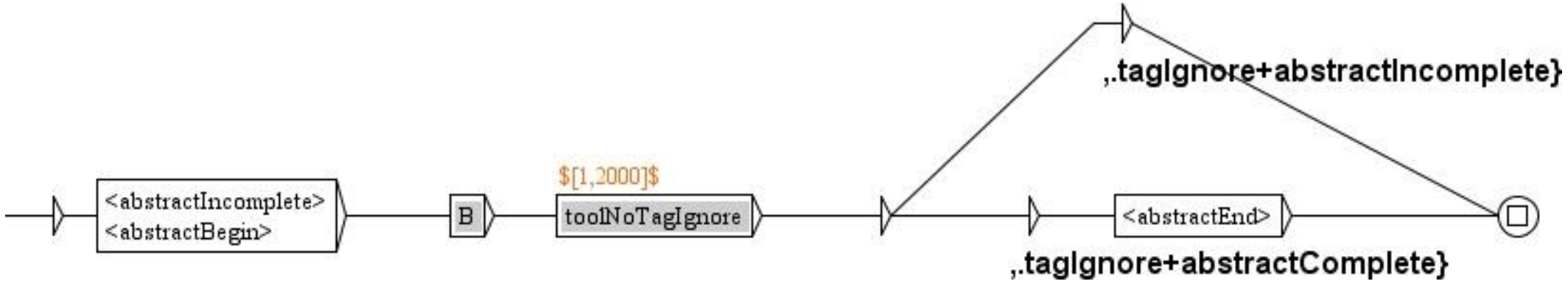
La chaine de traitement

- **1ère cascade Unitex**



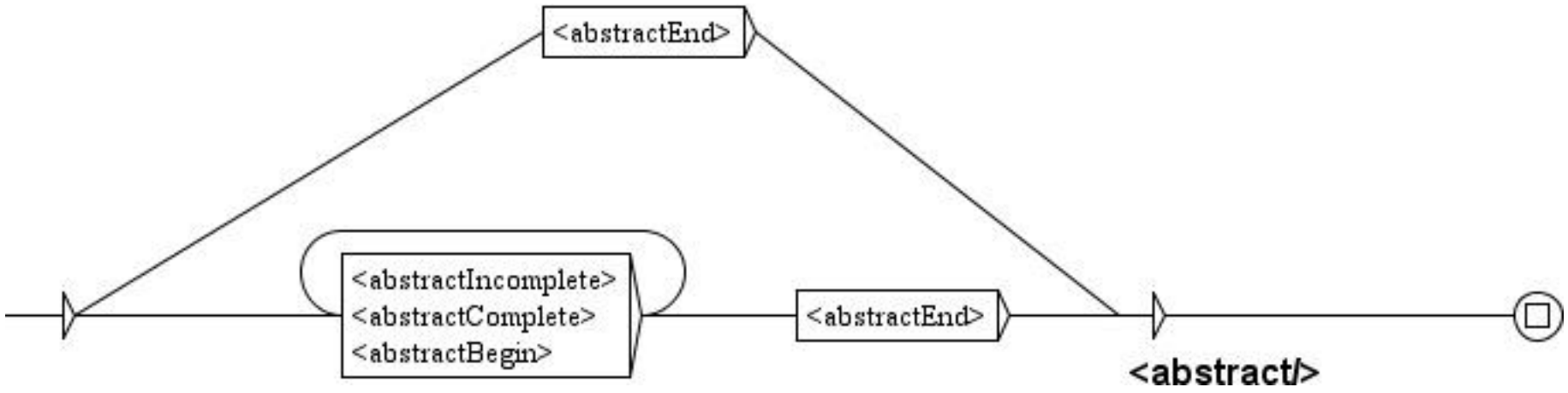
La chaine de traitement

- **1ère cascade Unitex**



La chaine de traitement

- **1ère cascade Unitex**



La chaine de traitement

- **Script PHP**
  - Élimination des fichiers sans intérêt
    - Les fichiers sans partie *Résultats*
    - Les fichiers sans les 3 mots clés encore présents
  - Lancement de la 2ème cascade Unitex

La chaine de traitement

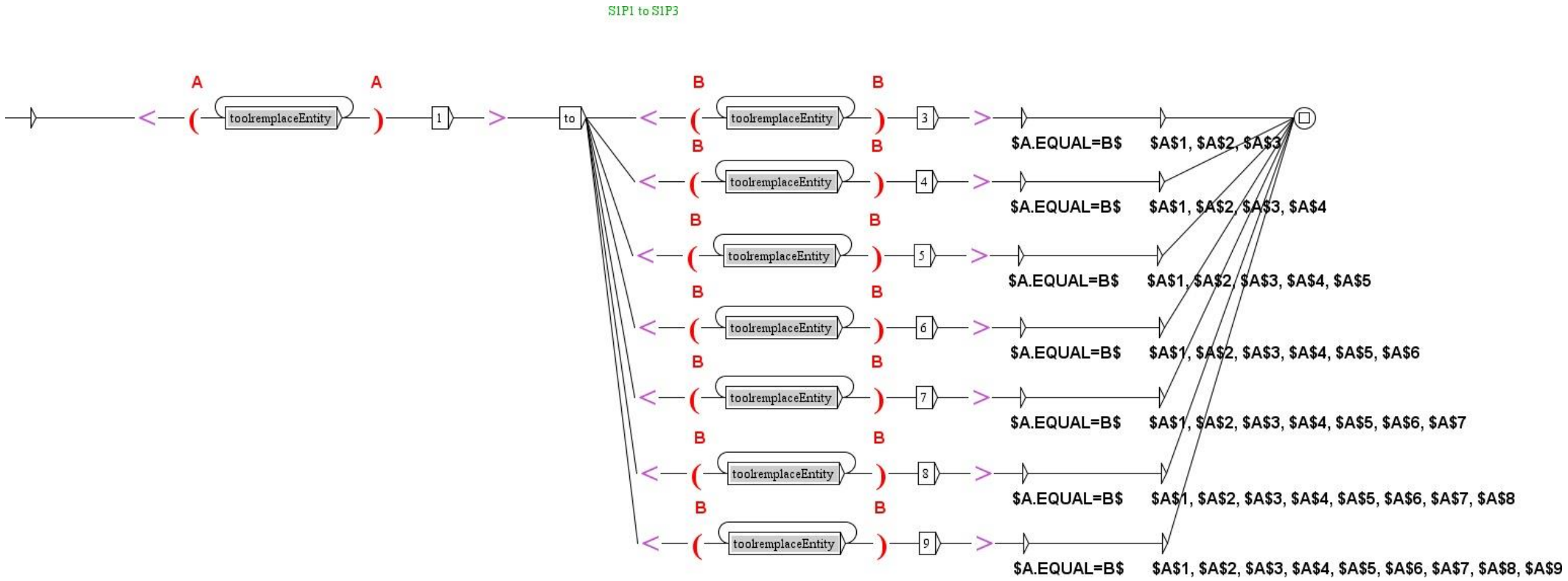
- **2ème cascade Unitex**
  - Cette cascade détaille, fichier par fichier, paragraphe par paragraphe, phrase par phrase, les structures prédicat-arguments correspondant à des résultats d'expérience
    - Pas de fichier de normalisation
    - Pas de *Preprocessing*

La chaine de traitement

- **2ème cascade Unitex**
  - Prétraitements
    - Balises Xml
    - Phrases
    - Verbes contractés
    - Renvois (figures, bibliographie)
    - Création de liste (par exemple : *S1P1 to S1P3* → *S1P1, S1P2, S1P3*)

La chaine de traitement

- **2ème cascade Unitex**



La chaine de traitement

- **2ème cascade Unitex**
  - Analyse biologique
    - Entités
    - Coordinations d'entités
    - Verbes
    - Prédicats simples
    - Statut hypotétique pour les affirmations non expérimentées
    - Prédicats complexes
  - Synthèse
    - Mise en forme TEI

La chaine de traitement

- **Script PHP**
  - Création du fichier d'analyse (au format XML-TEI)

    <teiCorpus>
    <teiHeader>*(entête présentant le projet Abliss)*</teiHeader>
      <tei>*(premier texte)*
        <teiHeader>*(entête contenant les métadonnées du premier fichier)*</teiHeader>
        <text>*(texte du premier fichier et structures prédicat-arguments)*</text>
      </tei>
    (textes suivants)
    </ teiCorpus >

  - Lancement de la 3ème cascade Unitex

La chaine de traitement

- **3ème cascade Unitex**
  - Cette cascade utilise l'option *standOff* pour donner des résultats quantitatifs
    - Un deuxième fichier de résultats est créé

```
<background>
  <fact>protein("ERK")</fact>
  <frequency value="2847"/>
 </background>
```

La chaine de traitement

- **Quelques chiffres**
  - 5 141 articles après interrogation de PubMed
  - 548 articles avec une partie Résultat et contenant encore les 3 mots clés
  - 15 132 arguments (protéines, gènes…)
  - 3 560 prédicats

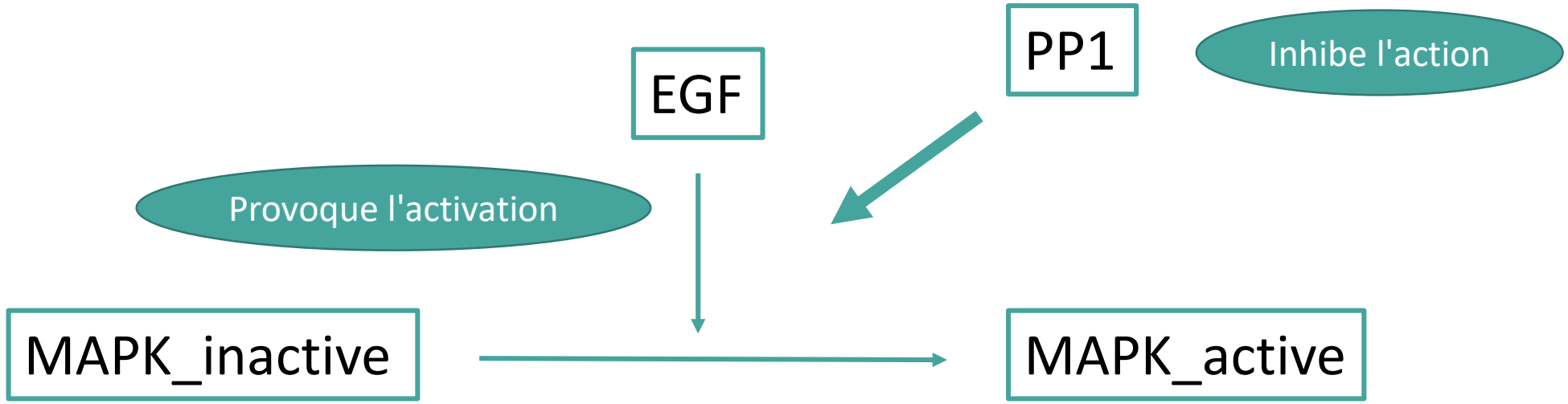
Un exemple

PP1 also had a significant inhibitory effect on MAPK activation by EGF.

(Goparaju et al., 2005)

Détail de l'exemple

PP1 also had a significant inhibitory effect on MAPK activation by EGF.



Les métadonnées

```
<teiHeader>
  <titleStmt>
    <title>The S1P2 Receptor Negatively Regulates Platelet-Derived Growth Factor-Induced Motility and Proliferation</title>
    <author>Goparaju SK, Jolly PS, Watterson KR, Bektas M, Alvarez S, Sarkar S, Mel LIshii I, Chun J, Milstien S, Spiegel S</author>
  </titleStmt>
  <publicationStmt>
    <publisher>Molecular and Cellular Biology</publisher>
    <date>2005 May</date>
    <biblScope unit="vol">25</biblScope>
    <biblScope unit="issue">10</biblScope>
    <biblScope unit="pp">4237-4249</biblScope>
    <idno type="PMID">15870293</idno>
    <idno type="DOI">10.1128/MCB.25.10.4237-4249.2005</idno>
    <idno type="PMCID">PMC1087716</idno>
  </publicationStmt>
</teiHeader>
```

Les structures prédicat-arguments

** **
PP1 also had a significant inhibitory effect on MAPK activation by EGF.
<desc type="entity">protein("PP1")</desc>
<desc type="entity">protein("MAPK")</desc>
<desc type="entity">protein("EGF")</desc>
<desc type="predicate">modifiedForm("MAPK"_active, "MAPK"_inactive")</desc>
<desc type="predicate">associationModulation("PP1", "MAPK"_active, "MAPK"_inactive, "EGF", increase, confirmed, unknownCell, unknownMethod)</desc>
** **

Détail de l'exemple

PP1 also had a significant inhibitory effect
on MAPK activation by EGF.

```
protein("PP1")
protein("MAPK")
protein("EGF")
```

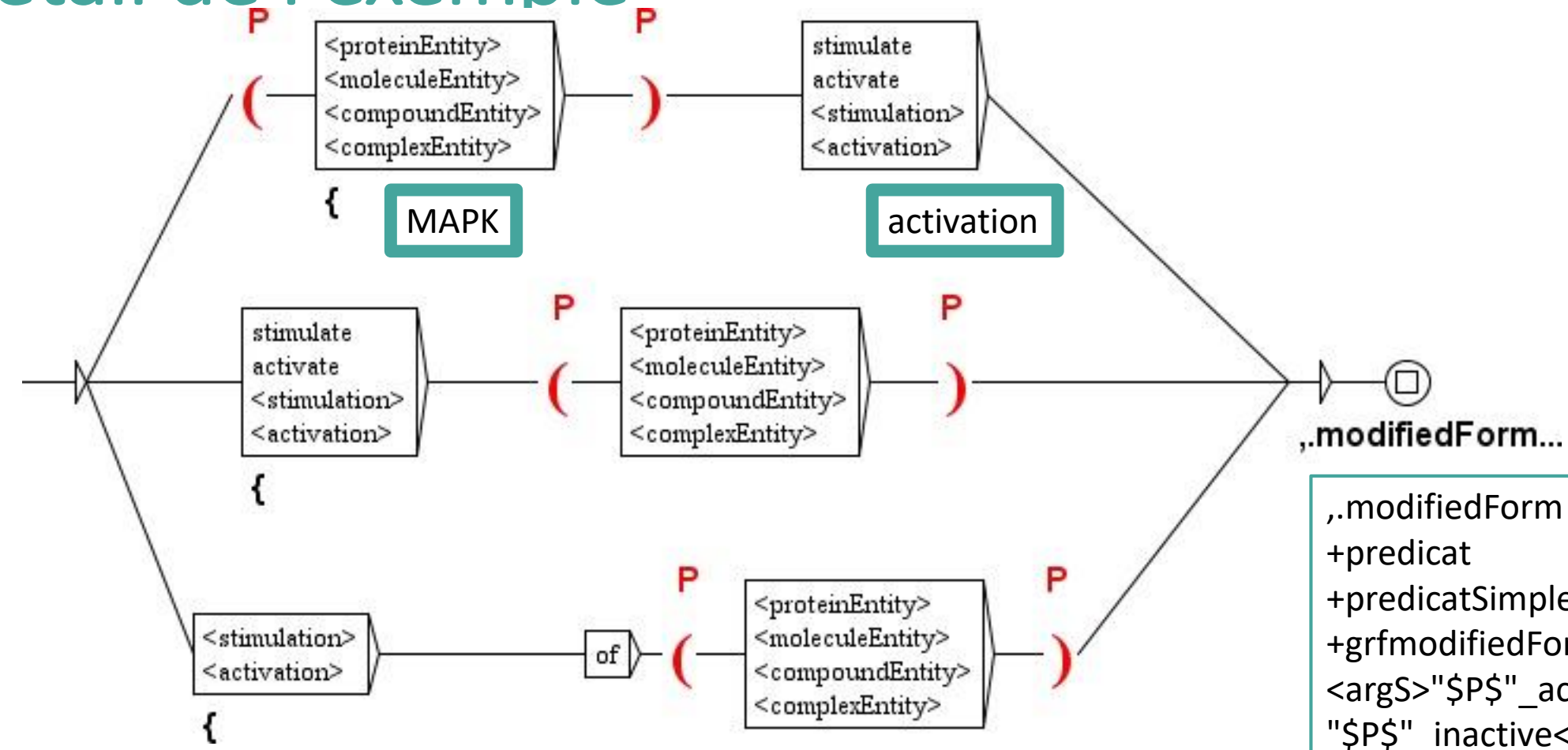
Détail de l'exemple

PP1 also had a significant inhibitory effect
on MAPK activation by EGF.

modifiedForm("MAPK"_active, "MAPK"_inactive")

modifiedForm("MAPK"_active, "MAPK"_inactive")

Détail de l'exemple




```
,.modifiedForm
+predicat
+predicatSimple
+grfmodifiedForm}
<argS>"$P$"_active,
"$P$"_inactive</argS>
```

Détail de l'exemple

PP1 also had a significant inhibitory effect
on MAPK activation by EGF.

associationModulation("PP1", "MAPK"_active, "MAPK"_inactive, "EGF", increase, confirmed, unknownCell, unknownMethod)

Poursuite du travail

Poursuite du travail

- **Finalisation de la chaîne**
- **Évaluation de la création des prédicats sur des textes non étudiés**
- **Recherche sur le paragraphe (?) des informations manquantes**
  - associationModulation("PP1", "MAPK"_active, "MAPK"_inactive, "EGF", increase, confirmed, *unknownCell*, *unknownMethod*)

Merci !