Projet [gram]lab

CRÉATION D'ENVIRONNEMENT DE TRAVAIL POUR LINGUISTE

2010-2012

Lidia Varga – Adrien Durand

Projet [gram]lab

- **Projet open source**
- **Développement de ressources autour des grammaires locales**
- **Pour enrichir la communauté scientifique et industrielle**

Projet [gram]lab

Financé par

- **Fonds Européen de Développement Economique FEDER**
- **Conseil Général de Seine et Marne (Région-île-de France)**

Labélisé par

- **Pôle de Compétitivité francilien CAP DIGITAL**

Projet [gram]lab

Partenaires du projet

- **KWAGA (PME)**
- **ACTIMOS (PME)**
- **LIGM (Laboratoire, Université Paris-Est Marne-la-Vallée)**
- **KWAM (PME)**
- **LINGWAY (PME)**

Utilisateur sous-traitant

- **APIL (Association des Professionnels des Industries de la Langue)**

Projet [gram]lab

Plan de la présentation

- Objectifs du projets et critères de réussite
- Les principaux axes du projet (SP)
- Plate-forme [gram]lab
- Module de création automatique et semi-automatique de graphes de grammaire locale (automate de séquences)
- Diffusion et cas d'usage

Projet [gram]lab

Pourquoi une plateforme pour linguiste?

- **Création des grammaires (dictionnaires) extrêmement coûteuse**
- **Maintenance, gestion et évaluation difficiles**
- **Il n'existe pas de plateforme intégrée pour le développement de graphes en TAL comme pour les développements informatiques (Eclipse, NetBeans, KDevelop, Xcode, etc.)**

Projet [gram]lab

Objectifs du projet

- **Créer des outils et ressources lingistiques pour faciliter le travail des linguistes (développeur et linguiste en TAL, linguiste, professeur de FLE, etc.)**
- **Réduire considérablement le temps de création de grammaires (dictionnaires inclus) défiant toute concurrence**
- **Faciliter la maintenance et la gestion des grammaires**
- **Diffuser et partager des ressources linguistiques et outils : intégration dans une chaîne de traitement (UIMA, Maven)**

Projet [gram]lab

Critères de réussite

- **Mise à disposition de la plateforme Gramlab**
- **Gain d'efficacité (temps et ergonomie) dans la production de GL spécifiques**
- **Adéquation aux besoins utilisateurs**
- **Gain d'efficacité dans la maintenance des grands ensemble de GL**
- **Qualité et couverture des GL développées**

Projet [gram]lab

Principaux axes du projet

- **Constitution de corpus**
- **Développement d'outils pour la création automatique/semi-automatique de GL spécifiques**
- **Création de la plateforme Gramlab**
- **Développement et diffusion lié à des cas d'usage**

Projet [gram]lab

Constitution de corpus

- **Besoin de corpus pour créer les grammaires (corpus de test)**
  - **Types de textes : e-mail, CV, lettre, page web, flux RSS**
- **Besoin de corpus annotés pour évaluer les grammaires (corpus de référence)**
- **Besoin de corpus plus large pour enrichir les grammaires**
- **Besoin de corpus qualifiés pour la création automatique et semi-automatique de grammaires**

Projet [gram]lab

Constitution de corpus

- **Problème : incompatibilité de format, d'encodage des corpus et les paramètres d'entrée de la Plateforme Gramlab (et Unitex)**
- **Solution : chaîne de constitution de corpus efficace**
  - **Modules de conversion :**
    - **PDF, txt , odt, rtf, doc, docx, XML > XML/TEI lite) et un interfaçage UIMA**
  - **Module de Crawl**
    - **Détection automatique de pages similaires**

Projet [gram]lab

Création de corpus qualifié

- **Le corpus qualifié est à la base de la chaîne de création de grammaire**
- **Une séquence a au minimum un token**
- **Séquences d'expressions sémantiquement proches**
- **Reconnaître des sous-classes sémantiques pas des classes :**
  - **Jusqu'à demain, avant mardi, pour lundi, à partir de jeudi**
- **Maintenabilité des graphes**

Projet [gram]lab

Unitex : noyau de la Plateforme [gram]lab

- **Plateforme multilingue d'analyse automatique de textes**
- **Logiciel open source, téléchargeable sous licence LGPL (version libre d'Intex)**
- **Logiciel multi-système (Windows, Linux, Mac OS X)**
- **Développé à L'Université Paris-Est Marne-la-Vallée (S. Paumier, 2003)**
- **http://igm.univ-mlv.fr/~unitex/**

Projet [gram]lab

La plateforme Gramlab

- **Noyau de la plate-forme : Unitex (S. Paumier, 2008)**
  - **Intégration des fonctionnalités existantes d'Unitex**
  - **Définition et développement de l'interface et des nouvelles interactions (mode projet, partage, versionning)**
  - **Ergonomie, fonctionnalité, rapidité**

- **Outils de développement rapide des ressources linguistiques**
  - **Procédure de comparaison**
  - **Test de non-régression**
  - **Gestion de collections de graphes**
  - **Aide au débogage**

Projet [gram]lab

La plateforme Gramlab

- **Intégration d'Unitex dans un IDE open source**
  - **Gestion dans Eclipse des grammaires locales comme unité de programmation**
  - **Optimisation et test en charge**
  - **Efficacité et robustesse en vue d'une industrialisation**
- **Normalisation et standardisation des échanges**
  - **Formats standardisés pour échange de projets (héritage) et ressources (corpus, dictionnaires)**

Développement informatique pour LIGM :

Sébastien Paumier (outils et intégration)

Adrien Durand (Automate de Séquences)

Projet [gram]lab

Développement automatique et semi-automatique de grammaires locales :

Le module d'Automate de Séquences

Objectif

- **Produire automatiquement des graphes de grammaires locales décrivant les unités de base (heures, noms de personnes, adresses…)**
- **Accélérer la production de grammaires plus spécifiques à une application**

Projet [gram]lab

Module d'Automate de Séquences

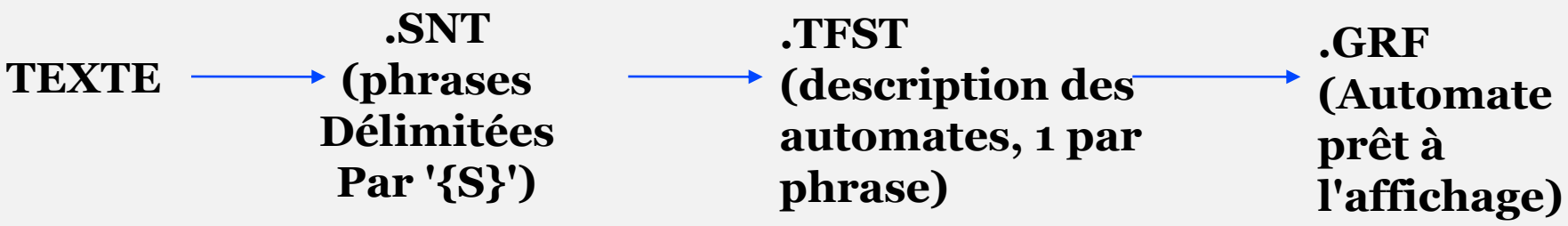
Corpus qualifié (corpus d'entrée)

- **Ensemble de séquences (pas nécessairement des phrases) regroupant des expressions proches (sens et type d'emploi)**
- **Sert à la génération automatique de graphes :**
- **Ensemble de séquences => graphe unique**
- **Formats**
  - **Liste de séquences séparées par retour à la ligne (txt)**
  - **Séquences délimitées par des balises XML dans un document XML/TEI :**

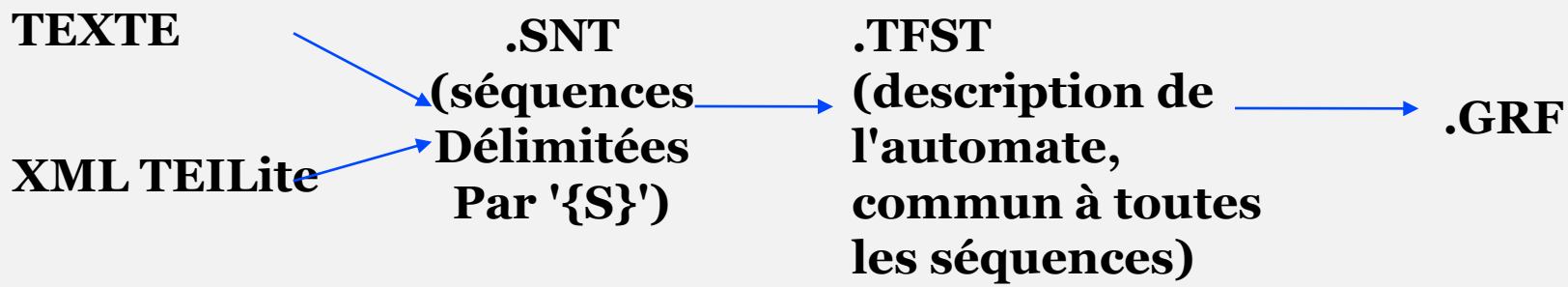
    <seg type="sequence">sequence</seg>

Projet [gram]lab

- **Chaîne pré-existante dans Unitex**



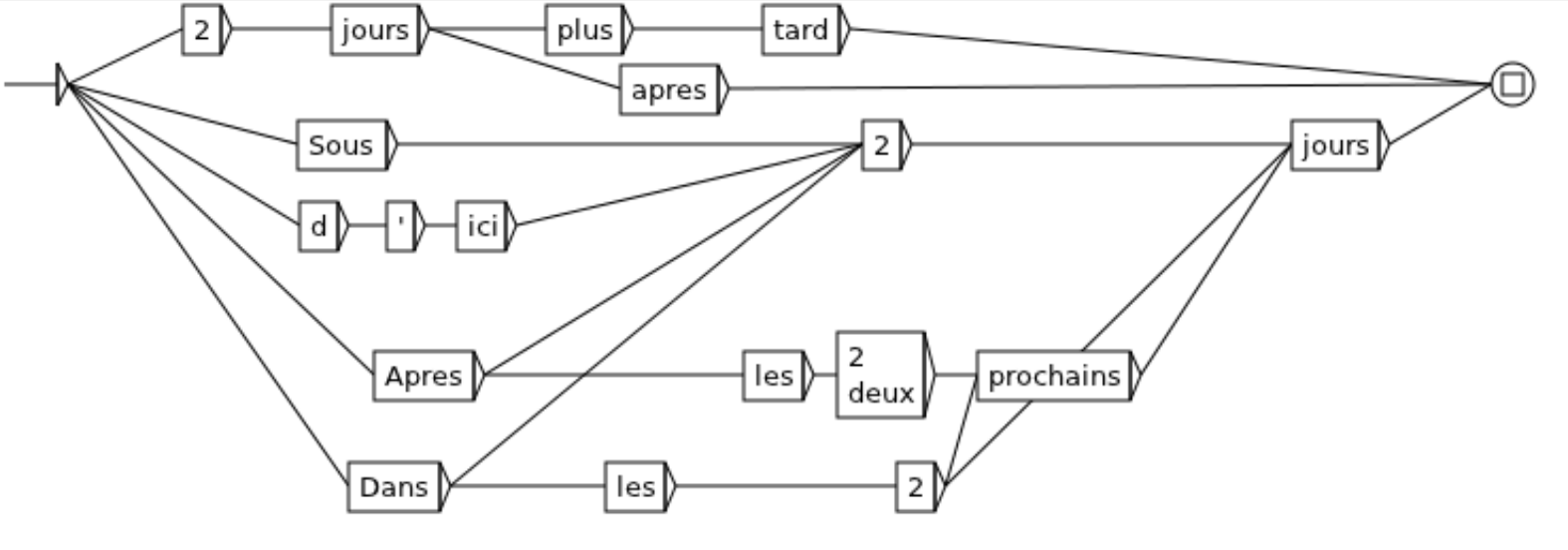
- **Chaîne modifiée pour Gramlab**



Projet [gram]lab

- **Exemple :**

Sous 2 jours
Apres 2 jours
Apres les 2 prochains jours
Apres les deux prochains jours
Dans 2 jours
Dans les 2 prochains jours
Dans les 2 jours
2 jours plus tard
2 jours apres
d ' ici 2 jours



Projet [gram]lab

- **Utilisation de jokers (1/2)**
  - **Problème : la liste des séquences reconnues (corpus qualifié) est incomplète**
  - **Objectif : reconnaître dans un corpus, de nouvelles séquences absentes de la liste donnée**
  - **Comment : produire à partir de chaque séquence une liste de séquences proches, à N opérations près.**
  - **Le nombre total d'opérations et le nombre de chaque opération (insertion / replacement / suppression) est borné.**
  - **Reconnaissance des séquences pertinentes précédemment ignorées à l'aide des graphes modifiés**

Projet [gram]lab

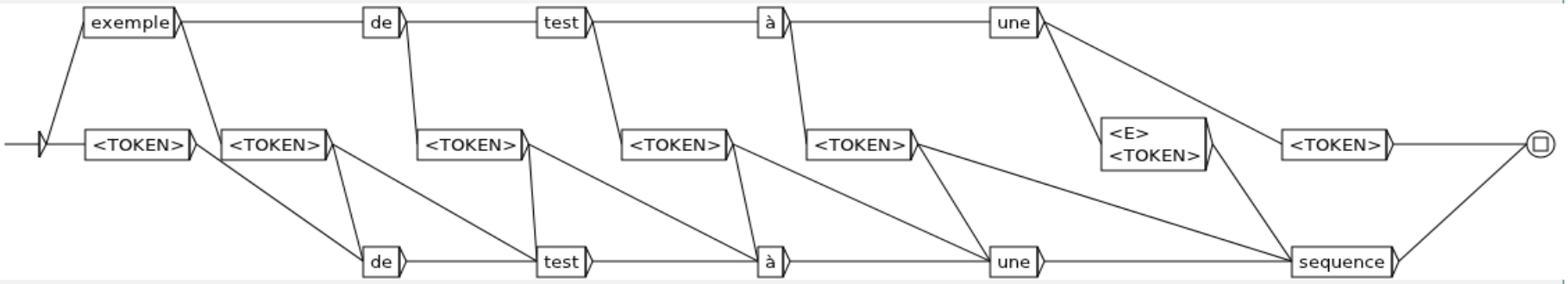
- **Utilisation de jokers (2/2)**

  Insertions, remplacements, suppressions de tokens

  - **Les séquences générées sont recherchées dans un corpus de test.**
  - **Les nouvelles séquences trouvées dans le corpus test sont ajoutées à la grammaire**
  - **Le nouvel automate créé gagne en couverture.**

Projet [gram]lab

- **Utilisation de jokers - Exemple**



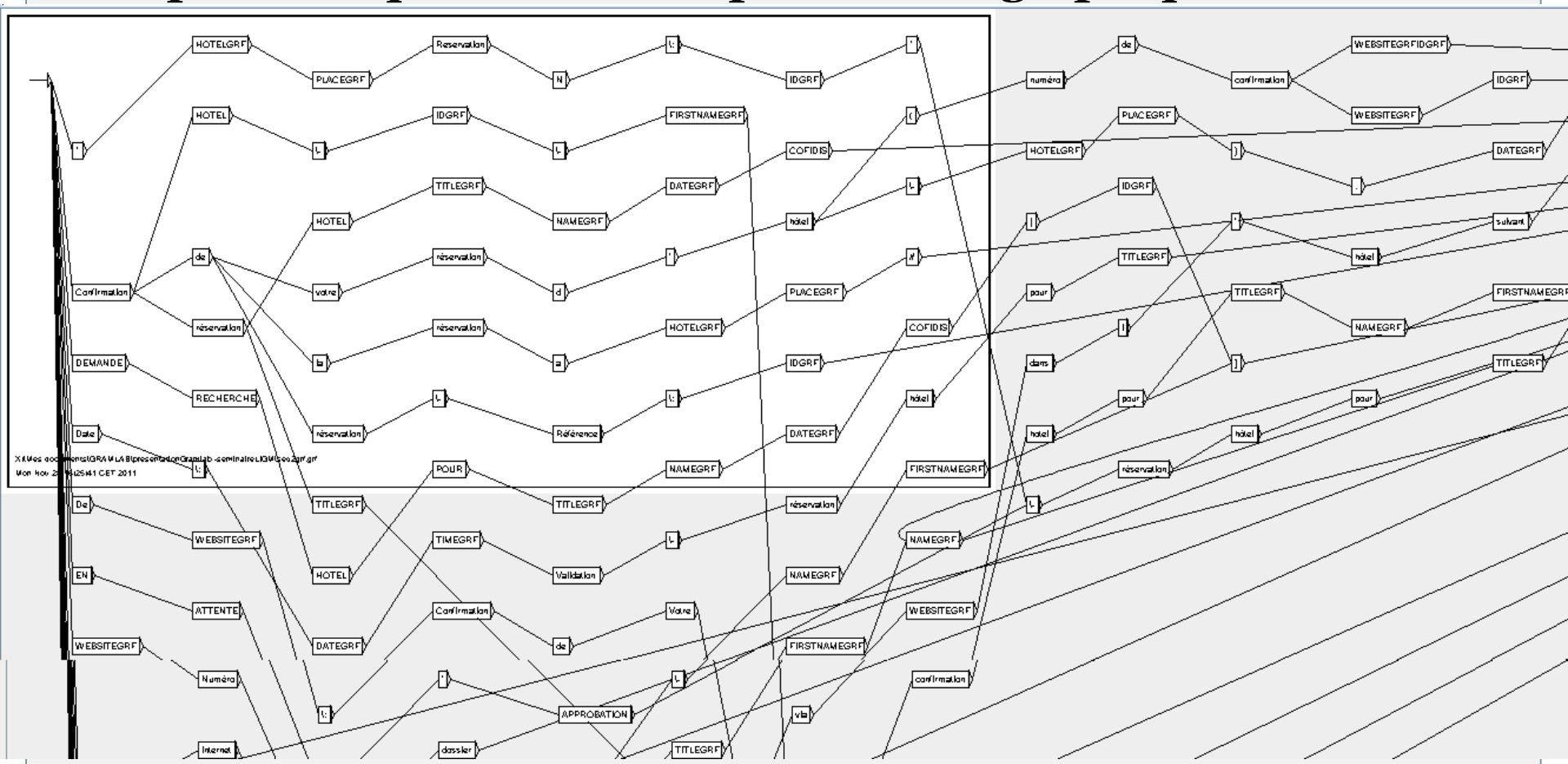
exemple de test à une séquence

max_opérations : 1
max_insertion : 1
max_suppression : 0
max_remplacement : 1

exemple de test à une séquence
<TOKEN> de test à une séquence
exemple <TOKEN> de test à une séquence
exemple <TOKEN> test à une séquence
exemple de <TOKEN> test à une séquence
exemple de <TOKEN> à une séquence
exemple de test <TOKEN> à une séquence
exemple de test <TOKEN> une séquence
exemple de test à <TOKEN> une séquence
exemple de test à <TOKEN> séquence
exemple de test à une <TOKEN> séquence
exemple de test à une <TOKEN>

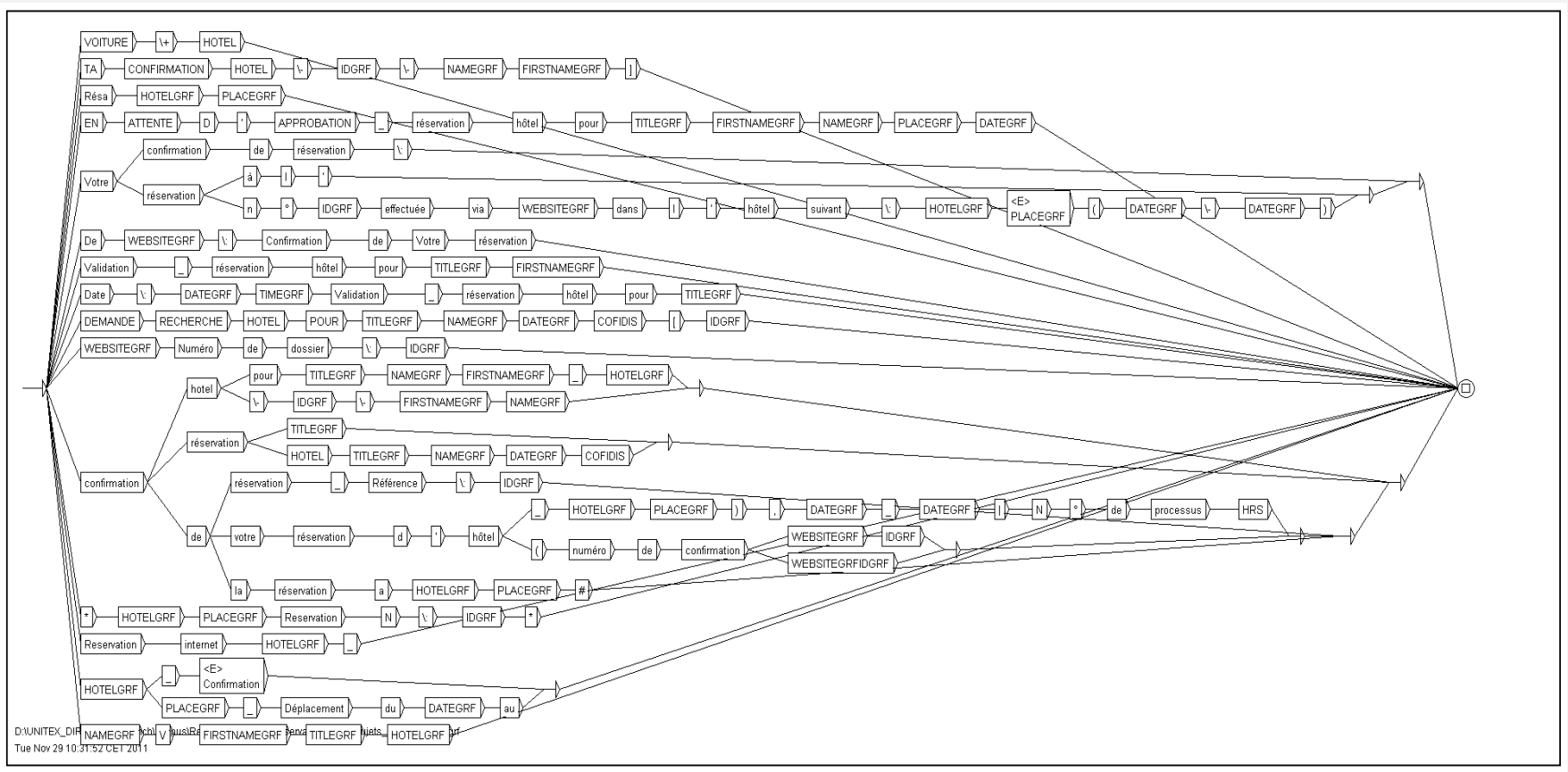
Projet [gram]lab

Graphe de séquences avant optimisation graphique



Projet [gram]lab

Graphe de séquences après optimisation graphique



Projet [gram]lab

Cas d'usage 1

- **CityAnnotator**

démontre l'utilisation d'une grammaire d'extraction de noms de villes françaises sur un corpus touristique

- **Le corpus de démonstration est un calendrier d'événements d'intérêt touristique en Othe Armance en Champagne. Il contient de nombreux noms de ville du département de l'Aube.**
- **La Grammaire reconnaît les noms de villes françaises présentes dans un texte et les annote avec la forme canonique et le numéro de département.**
- **Annotateur compatible avec le standard Apache UIMA**

Le nom de commune AIX-EN-OTHE, du département de l'Aube (10), donnera l'annotation suivante :

<annotGramLab commune="Aix-en-Othe" dep="Ville+10">AIX-EN-THE<annotGramLab>

Projet [gram]lab

Cas d'usage 2 :

Extracteur d'informations à partir de mails (des numéros de réservation de train, vol, hôtel)

Démontre la création et l'utilisation d'une grammaire conçue à l'aide de l'automate de séquences qui reconnaît le contexte (déclencheur) des occurrences de numéros de réservation

- **Le corpus d'entré : un corpus qualifiés des séquences de texte susceptibles de précéder un numéro de réservation**
- **Outils : module d'automate de séquences**
- **Sortie intermédiaire : grammaire en forme de graphe dans Unitex version 3.0 beta**
- **Application sur un corpus de test**
- **Modification du graphe possible par le linguiste**

Projet [gram]lab

Cas d'usage 2 - Corpus de test > corpus qualifié 1

Fichier Edition Format Affichage ?

```
Accord nécessaire pour la réservation : <Npers> <DATE>
Booking number:   <Num>
booking.com Numéro de réservation<Num>
code de réservation : <Num>
Communiquez ce N° de commande uniquement lors de vos échanges avec notre service clients  <Num>
Confirmation de votre réservation (N° de dossier <Num>)
La référence de votre réservation est <Num>
Merci de votre réservation: <Num>
Nouvelle confirmation de réservation: <Num>
Nouvelle réservation: <Num>
Numéro de confirmation Hotels.com <Num>
Numéro de référence: <Num>
Numéro de réservation : <Num>
Numéro de réservation : <Num>
Numéro de réservation voyages-sncf.com (vol): <Num> (7)
Numéro de vol: <Num>
Référence de la réservation:  <Num>Veuillez
référence de réservation: <Num>
référence de réservation: <Num>
Référence de votre dossier :* *<Num>
Référence de votre réservation <Ntrain> <Num>
Référence de votre réservation:  <Num>Votre
Références de réservation N° de commande Opodo : <Num>
RESERVATION NUMBER(S) TK/<Num>
votre réservation est en cours de traitement
voyageNuméro de réservation : <Num>IPassagerNuméro
YOUR RESERVATION NUMBER IS: <Num>
```

Projet [gram]lab

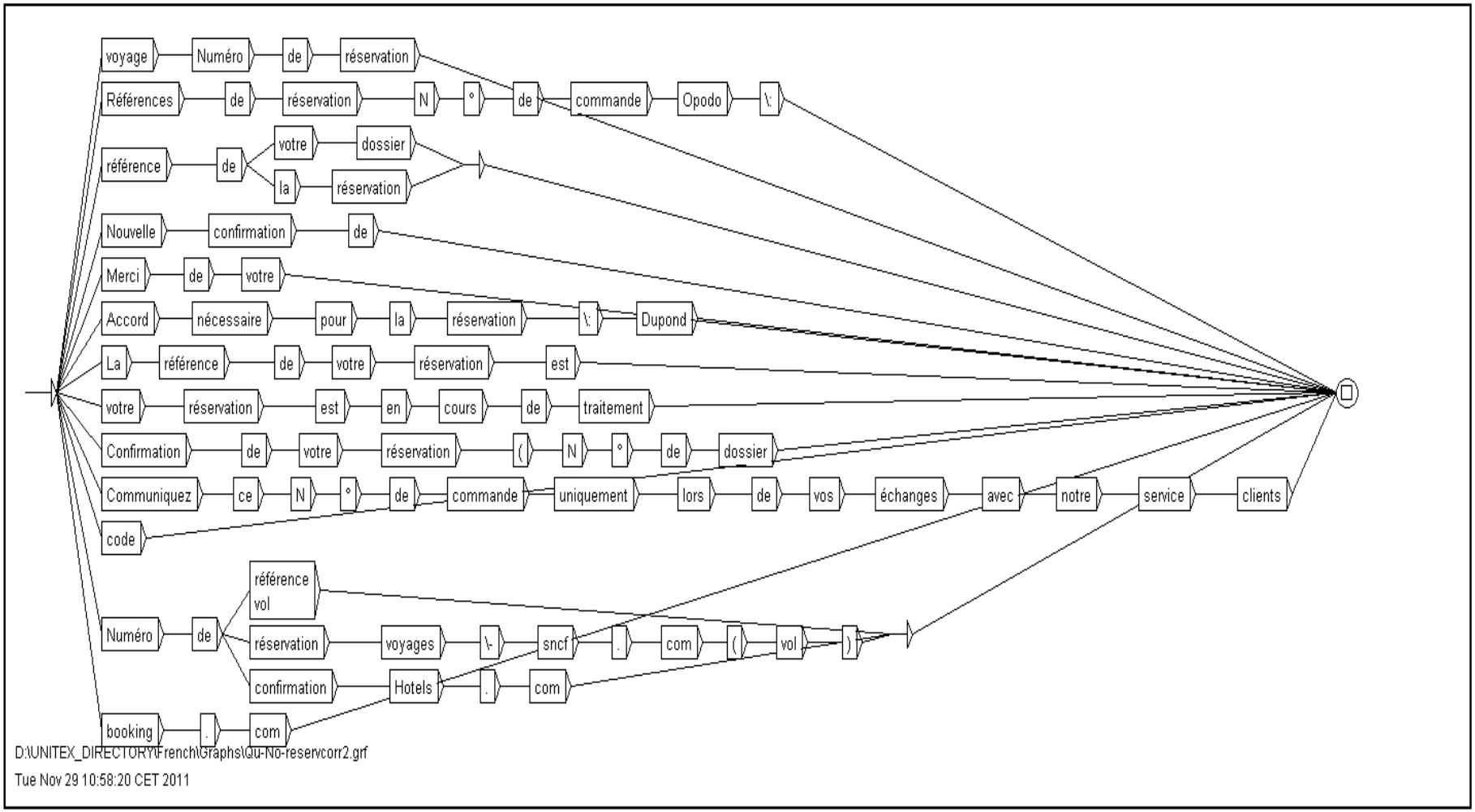
Cas d'usage 2 - Corpus de test > corpus qualifié 2

Fichier Edition Format Affichage ?

Accord nécessaire pour la réservation : Dupond
Booking.com Numéro de réservation <Num>
Code de réservation : <Num>
Communiquez ce N° de commande uniquement lors de vos échanges avec notre service clients <Num>
Confirmation de votre réservation (N° de dossier <Num>)
La référence de votre réservation est <Num>
Merci de votre réservation: <Num>
Nouvelle confirmation de réservation: <Num>
Nouvelle réservation: <Num>
Numéro de confirmation Hotels.com <Num>
Numéro de référence: <Num>
Numéro de réservation : <Num>
Numéro de réservation : <Num>
Numéro de réservation voyages-sncf.com (vol): <Num> (7)
Numéro de vol: <Num>
Référence de la réservation: <Num>
référence de réservation: <Num>
référence de réservation: <Num>
Référence de votre dossier :* *<Num>
Référence de votre réservation <Num>
Référence de votre réservation: <Num> Votre
Références de réservation N° de commande Opodo : <Num>
votre réservation est en cours de traitement
voyage Numéro de réservation : <Num>

Projet [gram]lab

Cas d'usage 2 - Sortie graphe



Projet [gram]lab

Cas d'usage 2

a) Traitement semi-automatique

- **Besoin de prétraitement spécifique des corpus (fautes d'orthographes typiques des mails,etc.)**
- **Graphes modifiables par le linguiste**
- **Rapide - réutilisable pour d'autres applications**

b) Traitement automatique

- **Rapide**
- **Grammaire pour une seule application**

Projet [gram]lab

Conclusion

- Nouvelles fonctionnalités d'Unitex
- Création de la plateforme intégrée Gramlab
- Utilisation de la plateforme Gramlab et d'Unitex dans des projets industriels de grande envergure
- Flexibilité d'intégration dans d'autres systèmes de traitement de l'information
- Gain de temps, robustesse et solution plus ergonomique
- Développements en cours

Projet [gram]lab

- **Site officiel : www.gramlab.org**

- http://sourceforge.net/projects/gramlab/