



Tutoriel Unitex-CasSys

Denis Maurel

Université de Tours

Ce tutoriel suppose de connaître Unitex et la création de graphe, par exemple en ayant fait le tutoriel *Prise en main d'Unitex pour l'annotation de corpus* disponible sur <https://tln.univ-tours.fr/tln/version-francaise/navigation/ressources/tutoriels-unitex>. On trouvera à cette même URL la préparation et la correction complète des exercices ci-dessous.

A titre d'exemple réel, on peut télécharger la cascade CasEN sur la reconnaissance d'entités nommées, réalisée par le Lifat, disponible sur : <https://tln.univ-tours.fr/tln/version-francaise/navigation/ressources/casen/>

Pour Unitex, choisir la version 3.2 alpha : <http://unitexgramlab.org/releases/latest-alpha/>.

Préparation

Le plus simple est de télécharger le fichier de préparation¹ sur le site TLN :

https://tln.univ-tours.fr/medias/fichier/preparation-tutoriel-unitex-cassys-denis-maurel_1562936277186-zip?ID_FICHE=321996&INLINE=FALSE

et de le dézipper dans votre dossier personnel Unitex (les fichiers se placeront au bon endroit).

Puis de passer à la section 1, page 3.

En cas d'impossibilité, poursuivre ci-dessous.

Création de dossiers

Plaçons-nous dans notre dossier personnel Unitex, dans *French*.

1. Dans *CasSys* : Créons un premier dossier pour l'ensemble du tutoriel, nommé *Tutoriel_Unitex_CasSys*². Créons, dans ce dossier, trois dossiers nommés *TexteBrut*, *TexteXML* et *Nombres*.
2. Dans *Corpus* : Créons un premier dossier pour la première partie du tutoriel, nommé *Tutoriel_Unitex_CasSys*. Créons, dans ce dossier, deux dossiers nommés *TexteBrut* et *TexteXML*. Puis, créons un second dossier nommé *80jours*, dans lequel nous faisons glisser le fichier *80jours.txt* présent dans la distribution d'Unitex.
3. Dans *Graph* : Créons un premier dossier pour l'ensemble du tutoriel, nommé *Tutoriel_Unitex_CasSys*. Créons, dans ce dossier, deux dossiers nommés *Texte* et *Nombres*. Dans chacun de ces deux dossiers, créons deux nouveaux dossiers nommés *Analyse* et *Synthese*.

¹ Sauf si vous l'avez déjà fait pour le Tutoriel Unitex, dans ce cas passez directement à la section 1, page 3.

² Remarque importante : comme Unitex est un logiciel multiplateformes, les noms de dossiers et de fichiers utilisés doivent ne comporter ni diacritiques, ni espaces.

Copie des textes à analyser

Le texte ci-dessous est à recopier via l'éditeur d'Unitex (menu *File Edition/New File*) et à enregistrer sous le nom *texte.txt* dans le dossier *French\Corpus\Tutoriel_Unitex_CasSys\TexteBrut*³.

Le maire de la ville de Prédetours a organisé une grande cérémonie le mardi 11 novembre 2014 pour modifier le nom de la rue de Prédetours en rue du 11 novembre 1918.

Ce mardi 11 novembre prouve que Prédetours n'a pas oublié ses anciens combattants.

D'ailleurs il n'est pas impossible qu'une Maison communale du 11 novembre voit le jour pour célébrer le 11 novembre 2018...

Puis utiliser le menu *File/Save As...* pour l'enregistrer à nouveau sous le même nom, *texte.txt*, dans le dossier *French\Corpus\Tutoriel_Unitex_CasSys\TexteXML*. Compléter alors le texte pour obtenir :

```
<tei>
<title>La Gazette de l'hôpital Bretonneau</title>
<date>Publié le mercredi 19 novembre 2014</date>
<body>
<p>Le maire de la ville de Prédetours a organisé une grande cérémonie le mardi 11 novembre 2014 pour
modifier le nom de la rue de Prédetours en rue du 11 novembre 1918.</p>
<p>Ce mardi 11 novembre prouve que Prédetours n'a pas oublié ses anciens combattants.</p>
<p>D'ailleurs il n'est pas impossible qu'une Maison communale du 11 novembre voit le jour pour célébrer le 11
novembre 2018...</p>
</body>
</tei>
```

Enregistrer le fichier puis fermer la fenêtre.

³ Si il apparaît, ne pas tenir compte de l'avis *This is not necessarily the text being processed by Unitex*, cliquer sur OK.

1 Texte brut à analyser en entités nommées

Dans cet exercice, le but est la création d'un fichier où seront balisés les dates, les villes, les adresses et les bâtiments du fichier *TexteBrut.txt*.

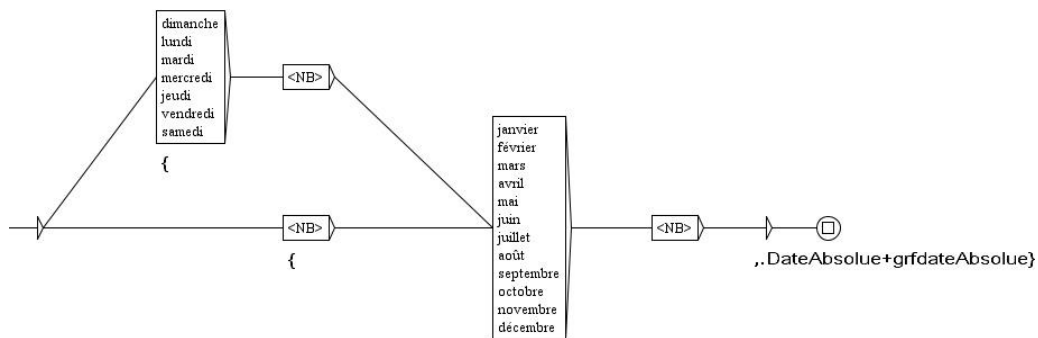
1.1 Cascade d'analyse

Commençons par ouvrir, dans le menu *Text* (et non *File Edition* comme précédemment), le fichier *French\Corpus\Tutoriel_Unitex_CasSys\TexteBrut\Texte.txt*, en répondant *Yes* à la question *Do you want to preprocess the text* et en décochant les deux premières coches pour ne laisser que l'option *Apply all default Dictionaries*.

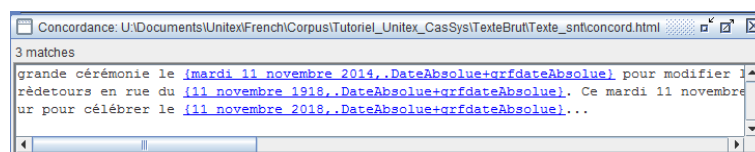
Créons un premier graphe qui sera enregistré (comme les suivants) dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Texte\Analyse*.

1.1.1 Graphe *dateAbsolue.grf*

Ce premier graphe va reconnaître les *dates absolues*, c'est-à-dire les dates contenant une année⁴. Pour simplifier les numéros de jours et d'années ne sont pas détaillés, mais simplement repérés par le code `<NB>` qui désigne une séquence de chiffres⁵. La date reconnue sera placée à l'intérieur des symboles `{` et `..DateAbsolue}`, afin de constituer une étiquette lexicale⁶, qui simule une entrée de dictionnaire de catégorie *DateAbsolue*. On ajoutera un trait pour le débogage, avec le nom du graphe, sous la forme `+grfdateAbsolue`.



Enregistrons et compilons⁷ (bouton *Compile*) ce graphe. Utilisons le menu *Text/Locate Pattern...* en choisissant ce graphe et en cochant l'option *Merge with input text*. Cliquons sur le bouton *Search*, puis sur *OK*. Règlons la taille du contexte : *Left: 20 chars* et *Right: 255 chars* ; puis l'ordre : *Sort according to:Text Order*. Cliquons sur le bouton *Build concordance*.



1.1.2 Graphe *dateRelative.grf*

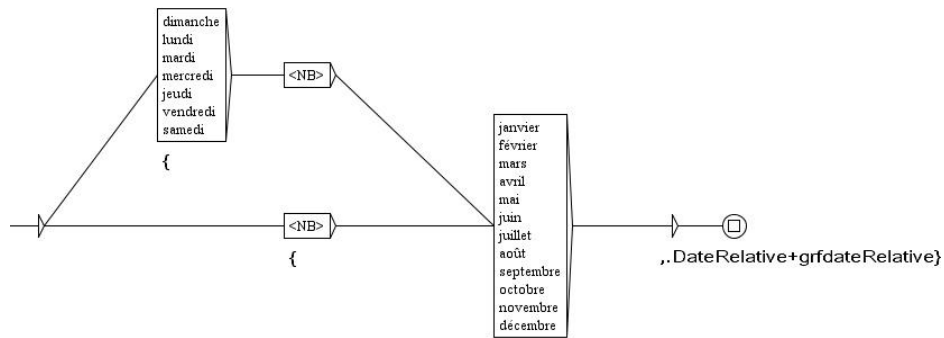
Créons un deuxième graphe pour reconnaître les *dates relatives*, c'est-à-dire les dates sans mention de l'année. Ouvrons le graphe *dateAbsolue.grf* et enregistrons-le sous le nom *dateRelative.grf*. Il nous reste à supprimer l'année et à modifier la sortie.

⁴ Cette définition et la suivante sont empruntées à la campagne d'évaluation *Ester*.

⁵ Pour la liste des symboles spéciaux, voir le manuel, section 4.3.1.

⁶ Voir le manuel, section 2.5.4.

⁷ Pour créer une cascade il est obligatoire de compiler les graphes que nous souhaitons y placer.



Enregistrons et compilons ce graphe.

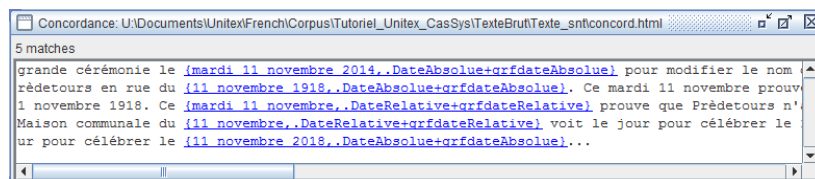
1.1.3 Cascade *analyse.csc*

Nous avons donc deux graphes que nous allons passer successivement sur le texte, grâce à une première cascade, que nous appellerons *cascade d'analyse*. Ouvrons le menu *Text/Apply CasSys Cascade...* et cliquons sur le bouton *New*. Plaçons-nous dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Texte\Analyse*. Avec la souris, faisons glisser le graphe *dateAbsolue.fst2*⁸, puis le graphe *dateRelative.fst2*, dans la partie droite de la fenêtre.

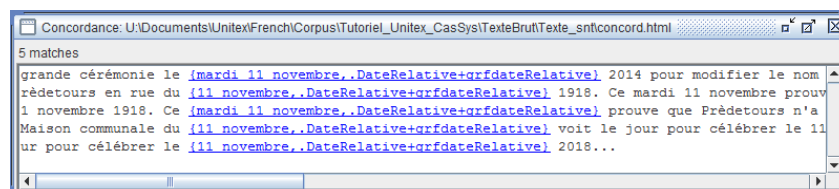
Voir la vidéo : http://tln.li.univ-tours.fr/Tln_VideosTutorielUnitex/creationCascade.mp4.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	dateAbsolue.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	dateRelative.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Enregistrons cette cascade dans le dossier *French\CasSys\Tutoriel_Unitex_CasSys\TexteBrut* sous le nom *analyse.csc*. Fermons la fenêtre et cliquons sur le bouton *Launch*. Règlons la taille du contexte : *Left: 20 chars* et *Right: 255 chars* ; puis l'ordre : *Sort according to:Text Order*. Cliquons sur le bouton *Build concordance*.



Remarquons que l'ordre des graphes est important, car, si nous inversons (avec la souris) ces deux graphes, la concordance obtenue n'est plus la même.

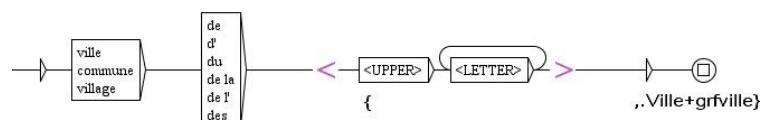


En effet, le graphe *dateRelative* a reconnu les cinq dates et donc, le graphe *dateAbsolue* n'avait plus rien à reconnaître !

1.1.4 Graphe *ville.grf*

Le graphe suivant va reconnaître les noms de ville, commune et village. Après l'un de ces trois mots, suivi de la préposition *de*, nous placerons un mot commençant par une majuscule, reconnu en utilisant le mode morphologique d'Unitex⁹ et les codes *<UPPER>* et *<LETTER>*.

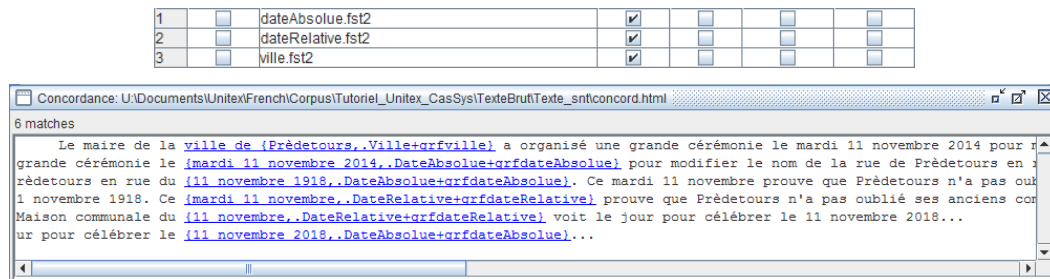
Voir la vidéo : http://tln.li.univ-tours.fr/Tln_VideosTutorielUnitex/ville.mp4.



⁸ Les fichiers *.fst2* sont des graphes compilés.

⁹ Voir le manuel, section 6.4.

Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *analyse.csc* (remise dans le bon ordre).

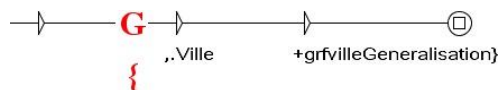


Nous pouvons remarquer sur cette concordance que *Prédetours* a été reconnu dans *ville de Prédetours*, mais, ni dans *rue de Prédetours*, ni dans *Prédetours n'a pas*. En effet, le premier terme est dans un contexte qui a permis sa reconnaissance, les deux autres non.

1.1.5 Graphe de généralisation d'étiquetage

Un graphe de généralisation d'étiquetage permet d'étiqueter des mots hors contexte s'ils ont déjà été étiquetés ailleurs dans le texte, grâce à un contexte. Par exemple, ici, la reconnaissance de *Prédetours* par le contexte *ville de* entrainera celui des deux autres occurrences de ce mot par le graphe *villeGeneralisation.grf*. Pour créer un graphe générique simple¹⁰, il faut sélectionner la boîte qui contient le code à généraliser et cliquer, comme pour les variables, les contextes et le mode morphologique, sur le bouton *insert generic graph mark before the selected box*.

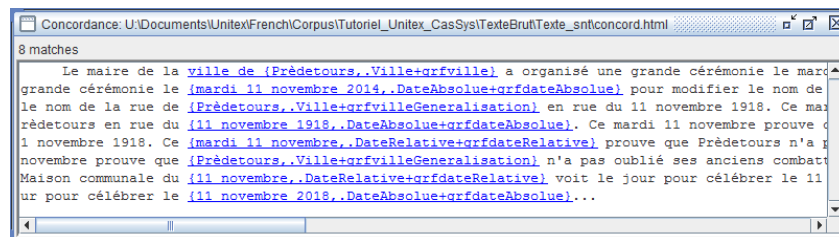
Voir la vidéo : http://tln.li.univ-tours.fr/Tln_VideosTutorielUnitex/villeGeneralisation.mp4.



Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *analyse.csc*, en cochant, en plus de la case *Merge*, la case *Generaliz...*

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	dateAbsolue.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	dateRelative.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	ville.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	villeGeneralisation.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

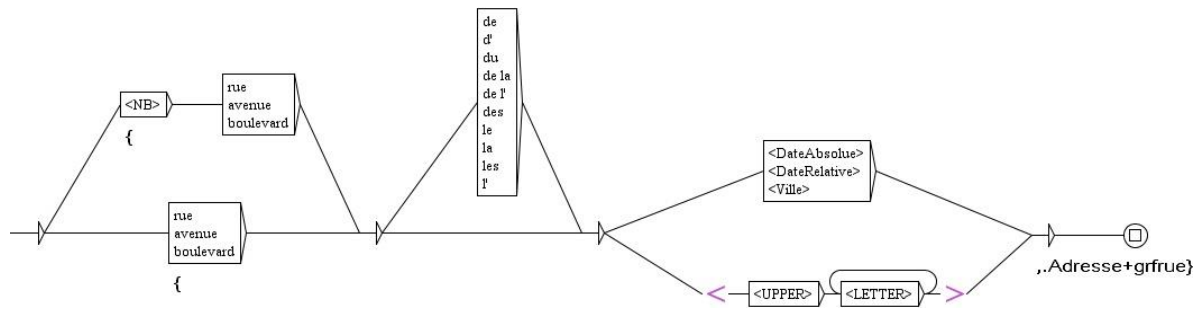
On obtient ce qu'on souhaitait : trois reconnaissances du mot *Prédetours*.



1.1.6 Graphe rue.grf

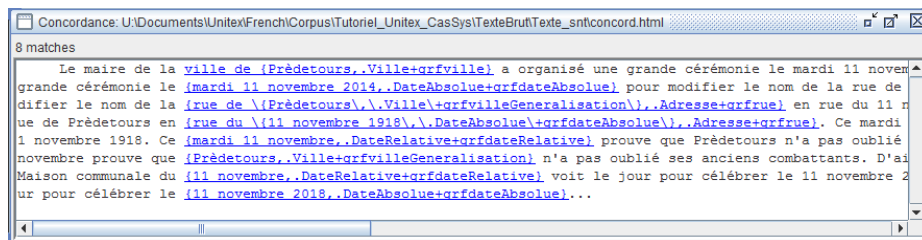
Créons maintenant un graphe pour reconnaître les noms de rue. Après les mots *rue*, *avenue*, *boulevard*..., éventuellement précédé d'un numéro et éventuellement suivi d'une préposition ou d'un déterminant, nous allons considérer deux possibilités : une date ou une ville, déjà reconnue par la cascade et donc désignée par les codes *<DateAbsolue>*, *<DateRelative>* ou *<Ville>* ; un mot commençant par une majuscule.

¹⁰ Des graphes de généralisation d'étiquetage plus sophistiqués sont possibles aussi. Voir le manuel, section 12.3.



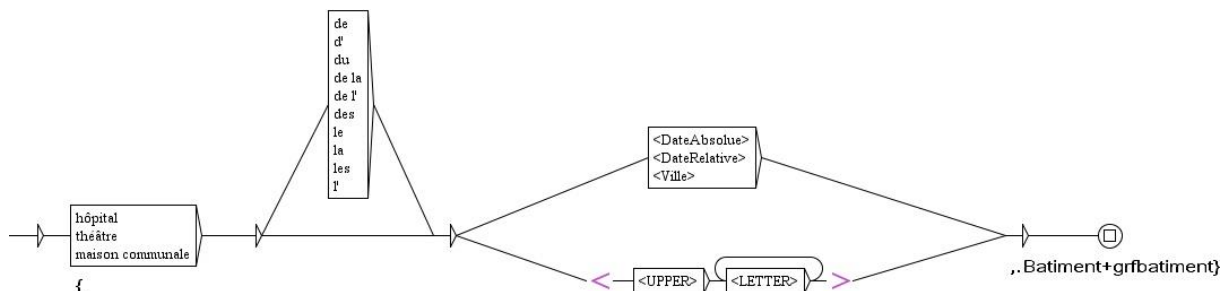
Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *analyse.csc*.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	dateAbsolute.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	dateRelative.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	ville.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	villeGeneralisation.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	<input type="checkbox"/>	rue.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



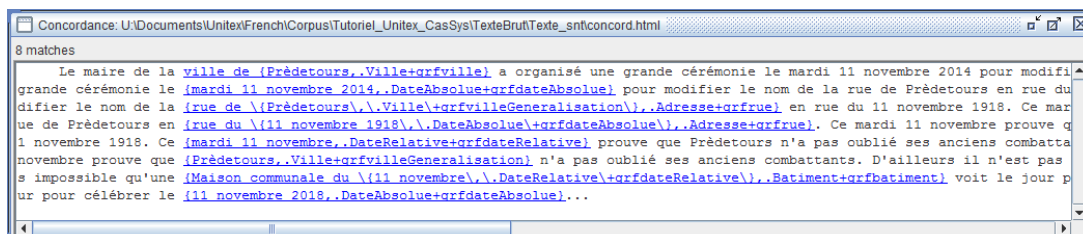
1.1.7 Graphe *batiment.grf*

Enfin, cliquons sur le menu *FSGraph/Save as...* et appelons ce graphe *batiment.grf*. Il nous suffit de modifier le début et la fin du graphe pour reconnaître les bâtiments.



Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *analyse.csc*.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	dateAbsolute.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	dateRelative.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	ville.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	villeGeneralisation.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	<input type="checkbox"/>	rue.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input type="checkbox"/>	batiment.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

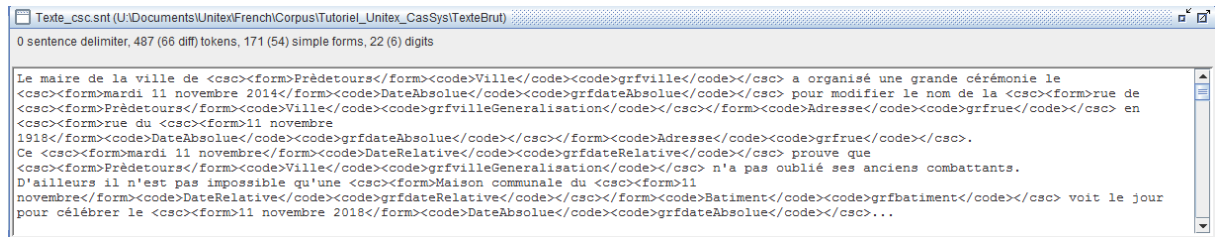


Le passage de la cascade sur le fichier *Texte.txt* génère (entre autres) un fichier XML, *Texte_csc.txt*, où les accolades sont remplacées par les balises `<csc>...</csc>`, la forme reconnue par `<form>...</form>` et le code et les traits par `<code>...</code>`¹¹.

¹¹ Voir le manuel, section 12.4.3.

1.2 Cascade de synthèse

Ouvrons dans le menu *Text* le fichier *French\Corpus\Tutoriel_Unitex_CasSys\TexteBrut\Texte_csc.txt*, en répondant *No* à la question *Do you want to preprocess the text*¹².



La cascade de synthèse sera passée sur ce fichier et permettra de créer un fichier balisé au format XML de notre choix. Créons un premier graphe qui sera enregistré (comme les suivants) dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Texte\Synthese*.

1.2.1 Graphe *balisage.grf*

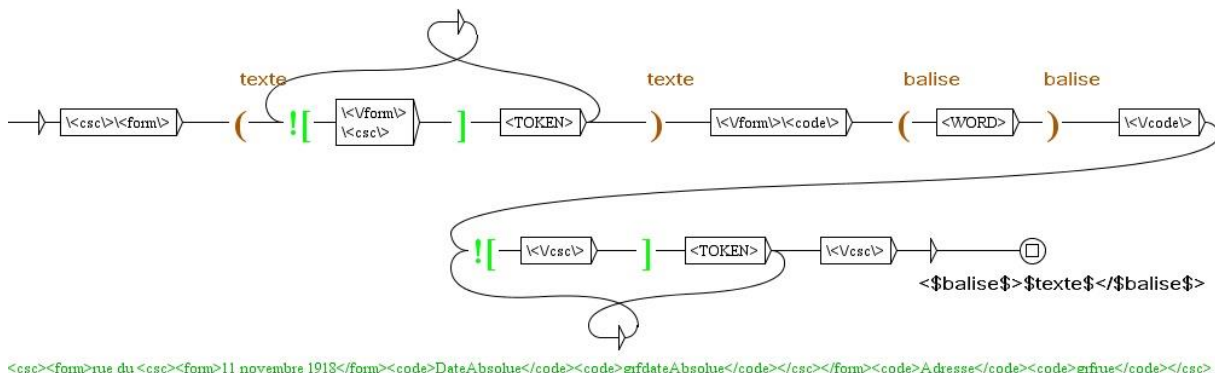
Le graphe *balisage* sera passé en mode *Replace* et aura pour but de :

1. supprimer les balises *csc*, *form* et *code* ;
2. sauvegarder le texte dans une variable¹³ *\$texte\$* et la catégorie dans une variable *\$balise\$* ;
3. remplacer l'ensemble par le texte balisé : *<\$balise\$>\$texte\$</\$balise\$>*.

Pour récupérer le texte, nous utilisons une boucle qui reconnaît tous les éléments (balise `<TOKEN>`) dont le contexte droit¹⁴ n'est, ni une balise `</form>`, ni une balise `<csc>`, qui débiterait une cascade imbriquée dans celle que nous analysons. Puis nous reconnaissons la catégorie (supposée être une séquence de lettres) par le code `<WORD>` et nous cherchons la balise `</csc>` de fin de cascade.

Une remarque importante : comme Unitex utilise des chevrons pour ses codes, nous devons, pour reconnaître un chevron dans un texte le *protéger* (c'est-à-dire le faire précéder) par un antislash¹⁵. Par exemple, la balise `<csc>` sera écrite dans une boîte Unitex `<csc>`. De même pour le slash qui sert, dans Unitex, à marquer le début des sorties. Par exemple, la balise `</csc>` sera écrite dans une boîte Unitex `</csc>`.

Pour nous aider à bien écrire ce graphe complexe, nous allons insérer un exemple dans une boîte de commentaire¹⁶, c'est-à-dire une boîte non reliée aux autres et qui commence par un slash.



Voir la vidéo : http://tln.li.univ-tours.fr/Tln_VideosTutorielUnitex/balilage.mp4.

Enregistrons et compilons ce graphe, puis créons une nouvelle cascade, que nous appellerons *synthese.csc*, et ajoutons-le en première position, en cochant le mode *Replace*.

¹² La synthèse ne nécessite pas de dictionnaires, car il s'agit juste d'une modification des balises XML.

13 Une variable dans Unixte mémorise la partie du texte reconnue par les boîtes entre parenthèses. Ces parenthèses se place comme les contextes et le mode morphologique, par le bouton *surround box selection with an input variable*. Le nom de la variable est libre. En sortie, ce nom est encadré par un caractère \$. Voir le manuel, section 5.2.5.

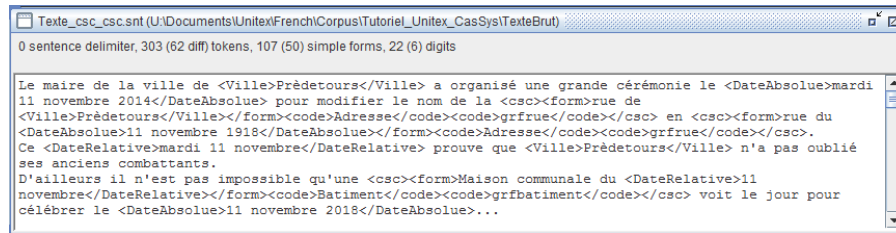
¹⁴ Voir le manuel, section 6.3.1.

¹⁵ Voir le manuel, section 5.2.7. Ceci ne concerne pas les sorties pour lesquelles aucune protection n'est nécessaire.

¹⁶ Voir le manuel, section 5.2.1.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

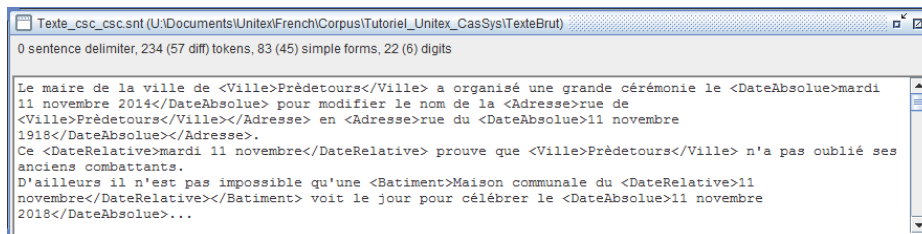
Lançons cette cascade. Nous obtenons la création d'un nouveau fichier XML, *Texte_csc_csc.txt*.¹⁷



La transformation du balisage des éléments non imbriqués est réussie. Par exemple, la première date devient `<DateAbsolue>mardi 11 novembre 2014</DateAbsolue>`. On remarque que, pour les éléments imbriqués, la partie intérieure est bien transformée. Il suffit donc de relancer une deuxième fois le graphe. Mais, s'il y avait trois niveaux d'imbrication, il faudrait le lancer trois fois... Nous allons donc cocher la case *Until Fix Point* qui relance le graphe jusqu'à ce que le texte ne soit plus modifié¹⁸.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

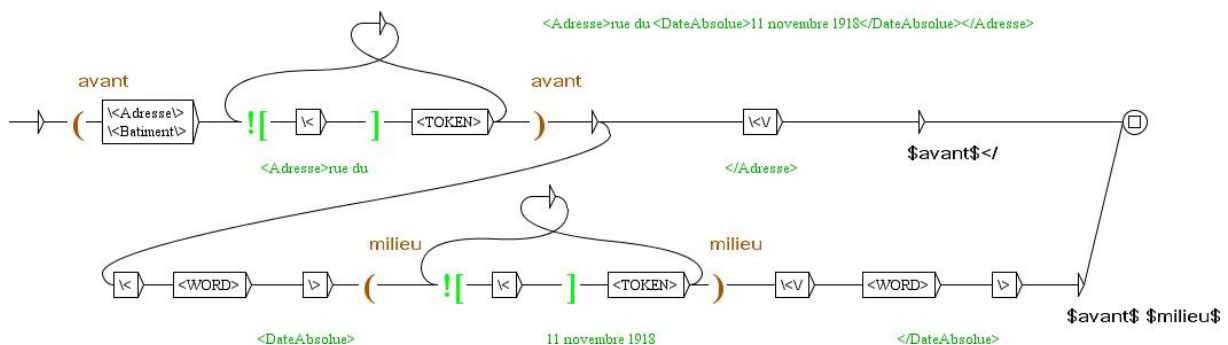
En lançant à nouveau cette cascade, le fichier *Texte_csc_csc.txt* devient correct, avec, par exemple : `<Adresse>rue du <DateAbsolue>11 novembre 1918</DateAbsolue></Adresse>`.



1.2.2 Graphe suppressionInterne.grf

Bien sûr, la pertinence de baliser une date ou une ville dans une adresse n'est pas évidente ! Indispensable pour reconnaître l'adresse, nous souhaiterions maintenant supprimer ce balisage. De même pour les noms de bâtiment.

Le graphe *suppressionInterne.grf* commence par placer dans la variable *\$avant\$* la balise et ce qui la suit jusqu'à l'ouverture d'une autre balise. Si cette balise est une balise fermante, il n'y a pas d'imbrication et on recopie le tout (c'est-à-dire *\$avant\$</>*). Sinon, si cette balise est une balise ouvrante, il y a imbrication et on mémorise juste le texte dans une variable *\$milieu\$*, puis on réécrit simplement *\$avant\$ \$milieu\$*.



Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *synthese.csc*.

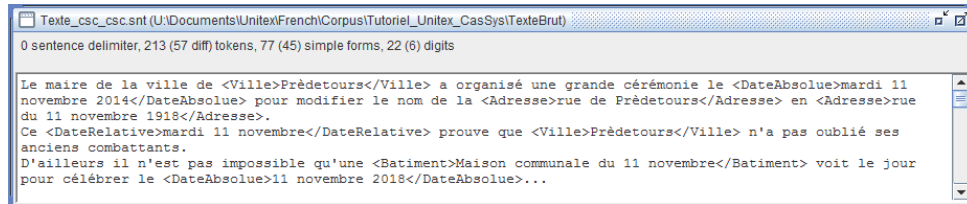
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	suppressionInterne.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

¹⁷ Lorsqu'il y a des suppressions (mode *Replace*) le résultat est difficile à visualiser dans une concordance. Nous afficherons ici systématiquement le fichier résultant, suffixé par *_csc_csc.txt*.

¹⁸ Voir le manuel, section 12.2.2.

1.2.3 Résultat final

En lançant cette cascade, le fichier *Texte_csc_csc.txt* devient tel que nous le souhaitons, avec, par exemple :
 <Adresse>rue du 11 novembre 1918</Adresse>.



Cependant une difficulté peut surgir, traitée en annexe (section 4, page 21), pour les utilisateurs avancés.

2 Texte XML à analyser en entités nommées

Dans cet exercice, le but est le même que pour l'exercice précédent, mais le fichier est un fichier XML.

Une remarque : ce fichier ne doit pas avoir l'extension *.xml*, sinon Unitex supprime les balises pour analyser le texte. Il doit être renommé¹⁹ si c'est le cas.

Nous allons utiliser les mêmes cascades, complétées par quelques graphes supplémentaires. Copions les deux cascades (*analyse.csc* et *synthese.csc*) du dossier *French\CasSys\Tutoriel_Unitex_CasSys\TexteBrut* dans le dossier *French\CasSys\Tutoriel_Unitex_CasSys\TexteXML*.

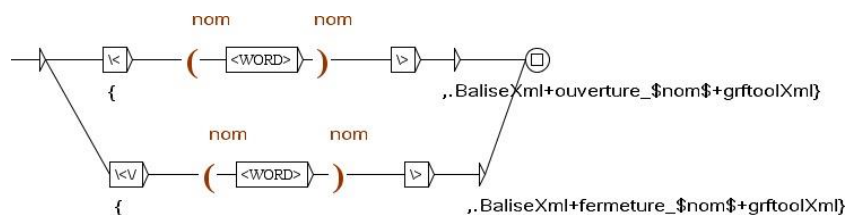
2.1 Cascade d'analyse

Pour lancer l'analyse, commençons par ouvrir, toujours dans le menu *Text*, le fichier *French\Corpus\Tutoriel_Unitex_CasSys\TexteXML\Texte.txt*, en répondant *Yes* à la question *Do you want to preprocess the text* et en décochant les deux premières coches pour ne laisser que l'option *Apply all default Dictionaries*.

Créons un premier graphe qui sera enregistré (comme les suivants) dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Texte\Analyse*.

2.1.1 Repérer les balises XML : graphe *toolXml.grf*

Pour simplifier, nous ne traitons ici que des balises *<nom>...</nom>*. Un graphe qui reconnaît toutes les balises XML est disponible à l'intérieur de la cascade CasEN²⁰. Le nom de la balise sera supposé être une séquence de lettres, reconnue par le code Unitex *<WORD>*. La catégorie de toutes les balises sera notée *baliseXml*. Cependant, pour garder le nom de la balise, nous allons utiliser une variable *\$nom\$* et indiquer aussi dans un trait s'il s'agit d'une balise ouvrante ou fermante. Ce trait sera donc *ouverture_\$nom\$* ou *fermeture_\$nom\$*.

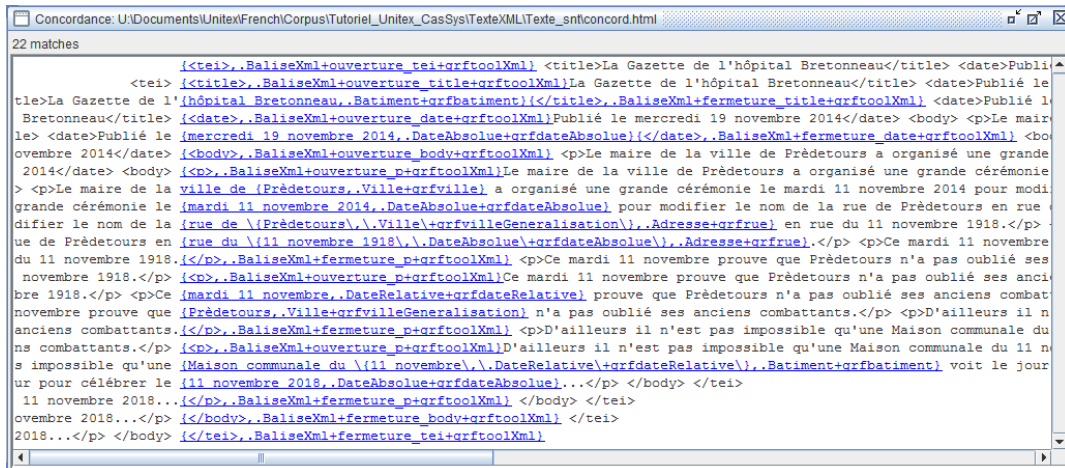


Enregistrons et compilons ce graphe, puis ajoutons-le en première position à la cascade *analyse.csc*.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	toolXml.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	dateAbsolute.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	dateRelative.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	ville.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	villeGeneralisation.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
6	<input type="checkbox"/>	rue.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	<input type="checkbox"/>	batiment.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

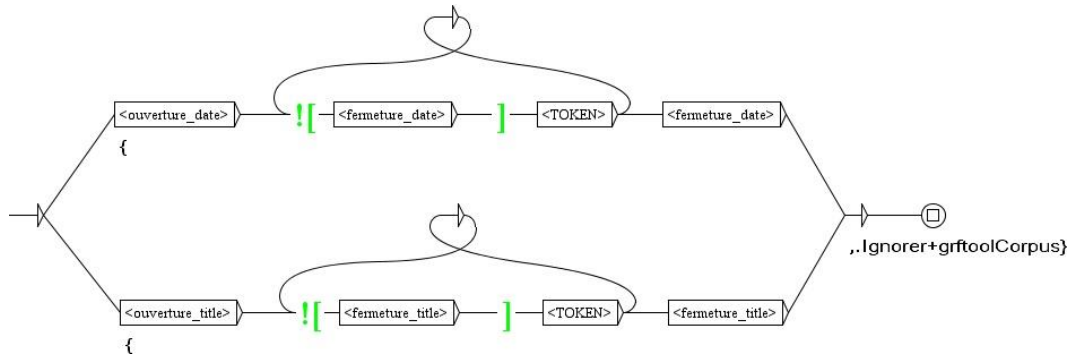
¹⁹ C'est-à-dire qu'un fichier *Texte.xml* doit être renommé *Texte.txt* ou *Texte.xml.txt* avant utilisation sous Unitex.

²⁰ <https://tln.univ-tours.fr/tln/version-francaise/navigation/ressources/casen/>



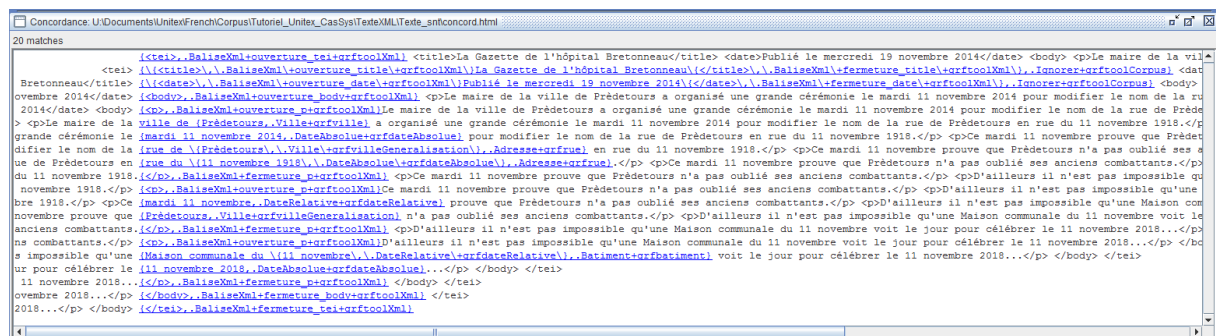
2.1.2 Cacher l'entête : graphe *toolCorpus.grf*

Dans l'exemple précédent, le bâtiment *hôpital Bretonneau* et la date *mercredi 19 novembre 2014* est reconnue dans l'entête alors qu'elle n'est pas dans le texte. Pour éviter cela, nous allons cacher le contenu des balises de l'entête, en les catégorisant *Ignorer*. En effet, les contenus de catégorie *Ignorer* ne seront pas traités par les graphes qui suivent.



Enregistrons et compilons ce graphe²¹, puis ajoutons-le en deuxième position à la cascade *analyse.csc*.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz.
1	<input type="checkbox"/>	toolXml.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	toolCorpus.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	dateAbsolute.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	dateRelative.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	ville.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input type="checkbox"/>	villeGeneralisation.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
7	<input type="checkbox"/>	rue.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	<input type="checkbox"/>	batiment.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



Remarque : nous avons utilisé le code Unitex *<ouverture_date>*, ce qui ne pose pas de problème, car il n'y a pas d'ambiguïté. Si le trait avait été nommé *ouverture*, le code *<ouverture>* aurait fonctionné aussi, mais aurait été ambigu avec le lexème *ouverture* (une *ouverture*, des *ouvertures*). Pour éviter cette ambiguïté, il faudrait utiliser la code *<BaliseXml+ouverture>*.

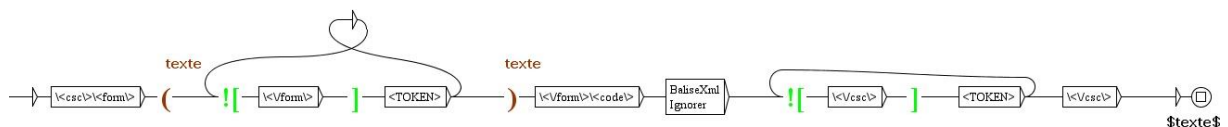
²¹ Si le texte que nous voulons cacher était trop long, ce graphe ne conviendrait pas. Il faudrait utiliser un graphe qui le cache petit à petit, avec l'option *Until Fix Point* cochée. Un exemple d'un tel graphe se trouve dans la cascade CasEN.

2.2 Cascade de synthèse

Ouvrons dans le menu *Text* le fichier *French\Corpus\Tutoriel_Unitex_CasSys\TexteXML\Texte_csc.txt*, en répondant *No* à la question *Do you want to preprocess the text*.

2.2.1 Graphe suppressionCodes.grf

Créons un nouveau graphe qui sera enregistré (comme les suivants) dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Texte\Synthese*. Ce graphe va supprimer les codes *BaliseXml* et *Ignorer* en récupérant le texte dans une variable *\$texte\$*.



Enregistrons et compilons ce graphe, puis ajoutons-le, en mode *Replace*, en première position à la cascade *analyse.csc*.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	suppressionCodes.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	suppressionInterne.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

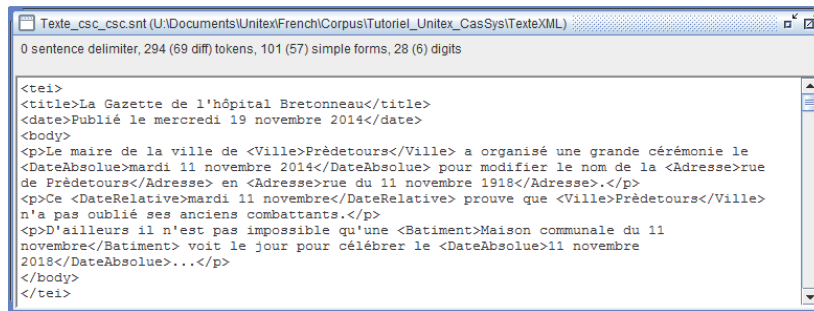
Si ce graphe n'est passé qu'une fois, il ne traite que les balises XML imbriquées dans les balises *Ignorer* :

Pour que la suppression soit complète, il faut cocher la case *Until Fix Point* :

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	suppressionCodes.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	suppressionInterne.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

2.2.2 Résultat final

La deuxième cascade génère le fichier *Texte_csc_csc.txt* qui correspond à notre souhait.



Texte_csc_csc.snt (U:\Documents\Unitex\French\Corpus\Tutoriel_Unitex_CasSys\TexteXML)

0 sentence delimiter, 294 (69 diff) tokens, 101 (57) simple forms, 28 (6) digits

```
<tei>
<title>La Gazette de l'hôpital Bretonneau</title>
<date>Publié le mercredi 19 novembre 2014</date>
<body>
<p>Le maire de la ville de <Ville>Prédetours</Ville> a organisé une grande cérémonie le
<DateAbsolue>mardi 11 novembre 2014</DateAbsolue> pour modifier le nom de la <Adresse>rue
de Prédetours</Adresse> en <Adresse>rue du 11 novembre 1918</Adresse>.</p>
<p>Ce <DateRelative>mardi 11 novembre</DateRelative> prouve que <Ville>Prédetours</Ville>
n'a pas oublié ses anciens combattants.</p>
<p>D'ailleurs il n'est pas impossible qu'une <Batiment>Maison communale du 11
novembre</Batiment> voit le jour pour célébrer le <DateAbsolue>11 novembre
2018</DateAbsolue>...</p>
</body>
</tei>
```

3 Nombres et mesures

Le but de cet exercice est d'annoter les nombres et les mesures dans le roman de Jules Verne distribué avec Unitex ("Le tour du monde en 80 jours"). Nous allons réutiliser, parfois en les modifiant, les graphes du tutoriel *Prise en main d'Unitex pour l'annotation de corpus*, disponibles sur <https://tln.univ-tours.fr/tln/version-francaise/navigation/ressources/tutoriels-unitex/>. Mais, si vous avez téléchargé et dézippé le fichier de préparation dans votre dossier personnel Unitex, les graphes que nous allons réutiliser ont déjà été copiés au bon endroit, à savoir dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Nombres\Analyse*.

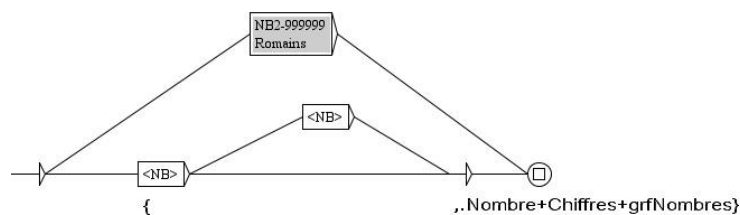
3.1 Cascade d'analyse

Commençons tout d'abord par ouvrir, dans le menu *Text*, le fichier intitulé *French\Corpus\80jours\80jours.txt*, en répondant *Yes* à la question *Do you want to preprocess the text* et en décochant les deux premières coches pour ne laisser que l'option *Apply all default Dictionaries*.

Créons un premier graphe qui sera enregistré (comme les suivants) dans le dossier intitulé *French\Graphs\Tutoriel_Unitex_CasSys\Nombres\Analyse*.

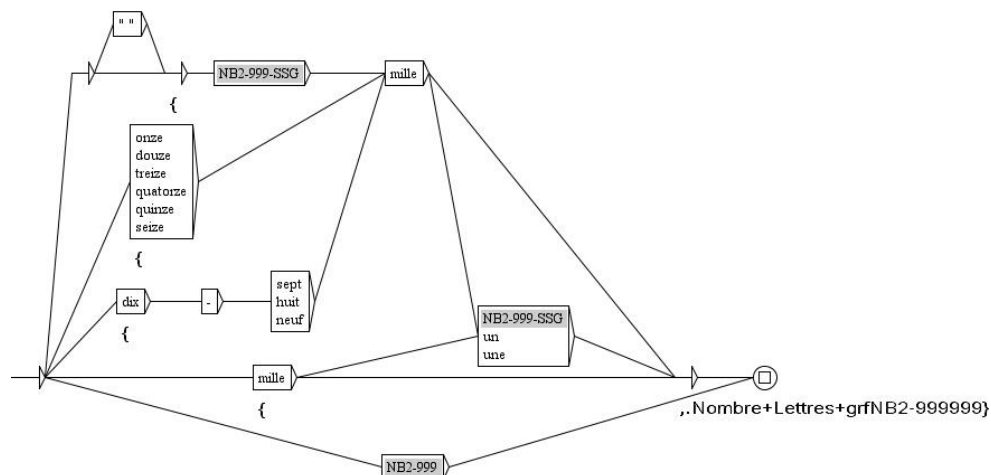
3.1.1 Graphe *Nombres.grf*

Le premier graphe de la cascade sera chargé de reconnaître et d'annoter les nombres entiers inférieurs au million, qu'ils soient écrits en chiffres arabes, en chiffres romains ou en toutes lettres. Rappelons que le symbole *<NB>* désigne une séquence de chiffres. Dans ce fichier, l'espace joue le rôle de séparateur de milliers.

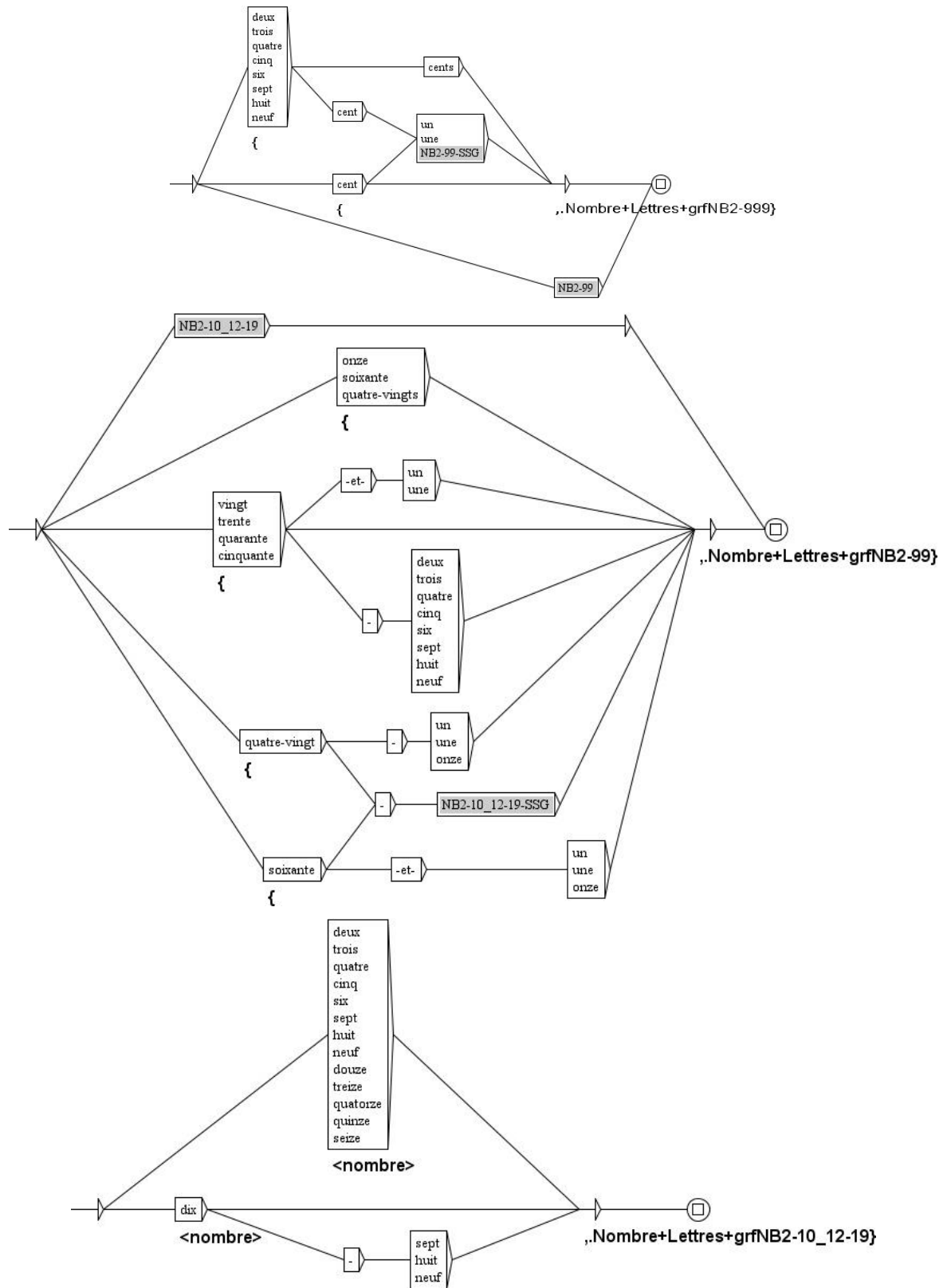


Les sous-graphes *NB2-999999* et *Romains* sont ceux construits dans le premier tutoriel. Nous allons les modifier rapidement en remplaçant les ouvertures de balise par *{* et les fermetures par *„Nombre}²²*. Dans la suite, les graphes suffixés *-SSG* sont les mêmes que les graphes non suffixés, mais sans sorties.

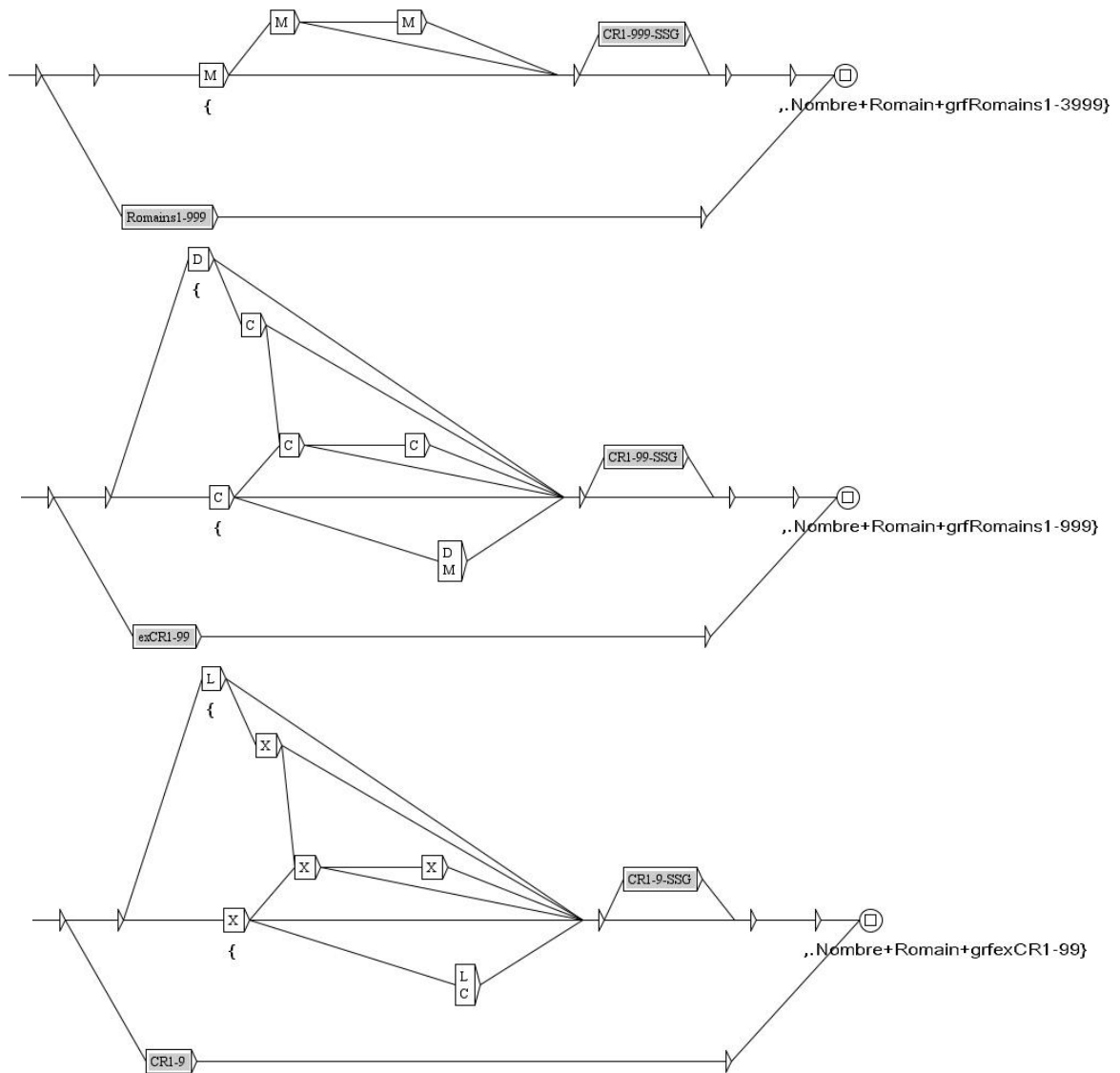
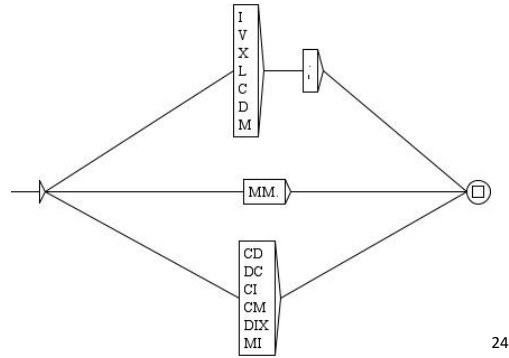
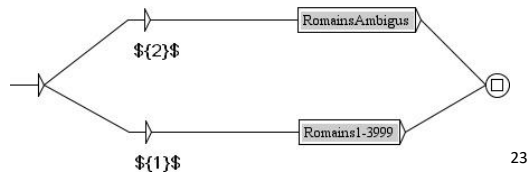
3.1.1.1 Sous-graphes des nombres écrits en toutes lettres



²² Comme d'habitude, nous ajouterons des traits à ces catégories.

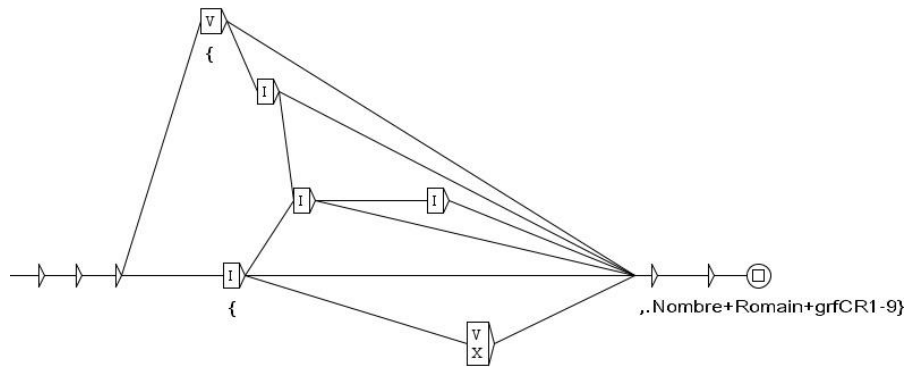


3.1.1.2 Sous-graphes des nombres écrits en chiffres romains



²³ Sous-graphe *Romains.grf*.

²⁴ Sous-graphe *RomainsAmbigus.grf*.



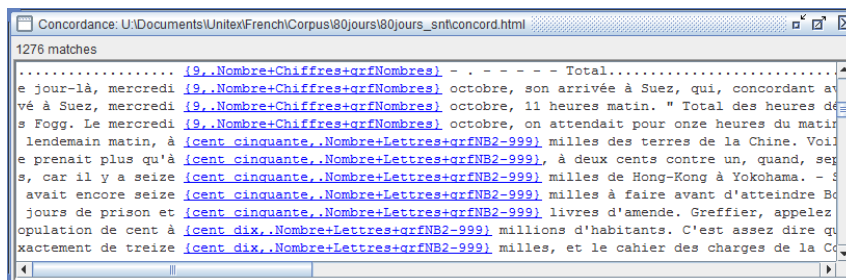
3.1.2 Création de la cascade d'analyse

Compilons le graphe *Nombres.grf*.

Ouvrons le menu *Text/Apply CasSys Cascade...* et cliquons sur le bouton *New*. Plaçons-nous dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Nombres\Analyse*. Avec la souris, faisons glisser le graphe *Nombre.fst2* dans la partie droite de la fenêtre.

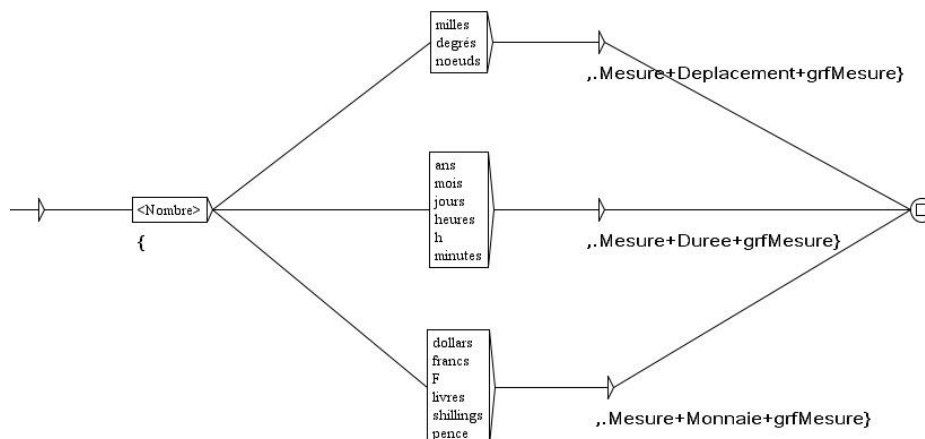
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	Nombres.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Enregistrons cette cascade dans le dossier *French\CasSys\Tutoriel_Unitex_CasSys\Nombres* sous le nom *analyse.csc*. Fermons la fenêtre et cliquons sur le bouton *Launch*. Réglons la taille du contexte : *Left: 20 chars* et *Right: 255 chars*. Cliquons sur le bouton *Build concordance*.



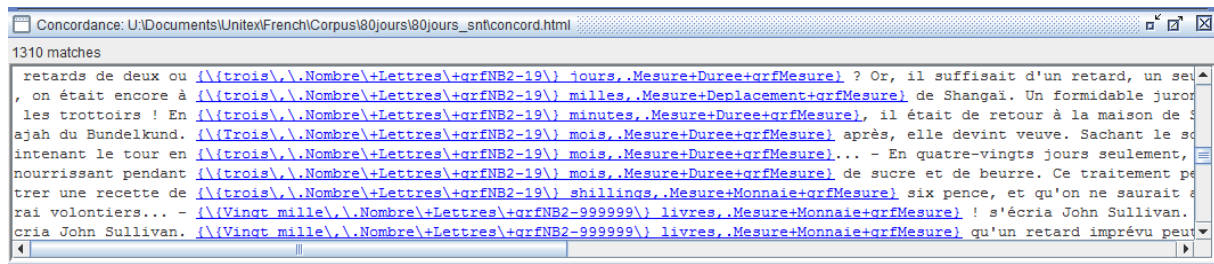
3.1.3 Graphe *mesure.grf*

Ce graphe, pour simplifier, sera *ad hoc* par rapport au roman. Un graphe des mesures est disponible dans la distribution de la cascade CasEN. Trois catégories de mesure apparaissent : les déplacements (*milles, degrés* et *noeuds*), la durée (*ans, mois, jours, heures, h, minutes*) et les monnaies (*dollars, francs, F, livres, shillings, pence*).



Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *analyse.csc*.

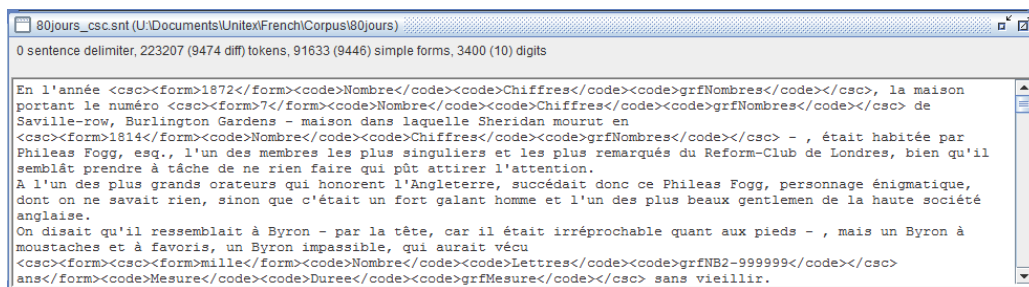
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	Nombres.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	Mesure.fst2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



3.2 Cascade de synthèse

3.2.1 Création de la cascade

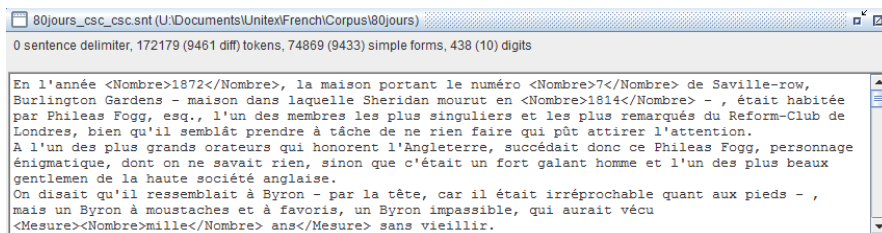
Ouvrons dans le menu *Text* le fichier *French\Corpus\80jours\80jours_csc.txt*, en répondant *No* à la question *Do you want to preprocess the text*.



Dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Texte\Synthese*, copions les graphes *balisage.grf* et *balisage.fst2*. Collons-les dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Nombres\Synthese*. Ouvrons le menu *Text/Apply CasSys Cascade...* et cliquons sur le bouton *New*. Plaçons-nous dans le dossier *French\Graphs\Tutoriel_Unitex_CasSys\Nombres\Synthese*. Avec la souris, faisons glisser le graphe *balisage.fst2* dans la partie droite de la fenêtre et cochons *Replace* et *Until Fix Point* (car il y a des balises imbriquées).

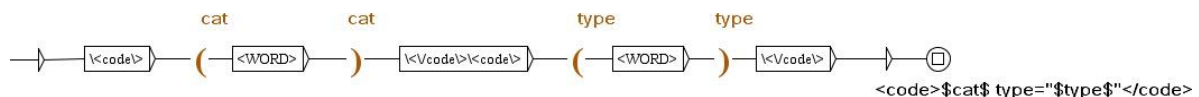
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Enregistrons cette cascade dans le dossier *French\CasSys\Tutoriel_Unitex_CasSys\Nombres* sous le nom *synthese.csc*. Fermons la fenêtre et cliquons sur le bouton *Launch*. Inutile de lancer la concordance, regardons plutôt le fichier *French\Corpus\80jours\80jours_csc_csc.txt*.



3.2.2 Le graphe type.grf

L'inconvénient de cette synthèse est que nous perdons toute l'information sur les catégories de nombre (lettres, chiffres, chiffres romains) et de mesure (déplacement, durée et monnaie). Pour éviter cela, nous allons créer un graphe *type.grf* qui va transformer le trait placé dans Unitex en un type XML. Par exemple, nous allons transformer `<code>Mesure</code><code>Monnaie</code>` en `<code>Mesure type="Monnaie"</code>`.

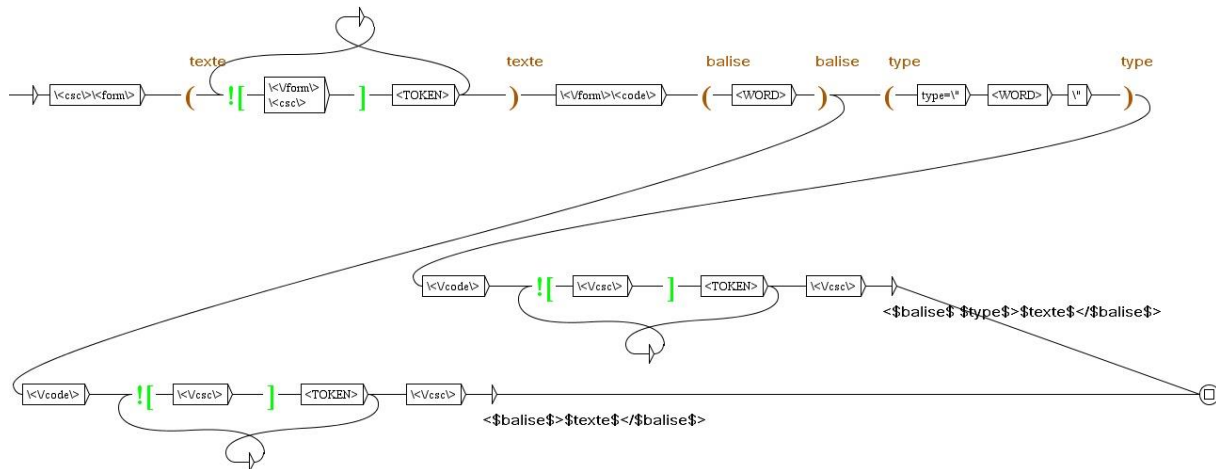


Enregistrons et compilons ce graphe, puis ajoutons-le en première position à la cascade *synthese.csc*, en mode *Replace*.

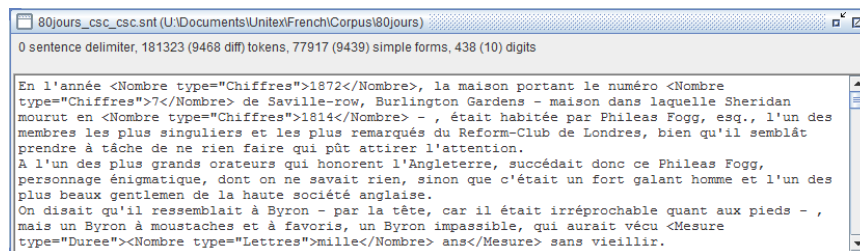
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	type.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

3.2.3 Le nouveau graphe *balisage.grf*.

Il nous faut alors compléter le graphe *balisage.grf* par l'éventuelle présence d'un type. Remarquons que, dans une boîte, les guillemets doivent être protégés par un antislash : \"²⁵.



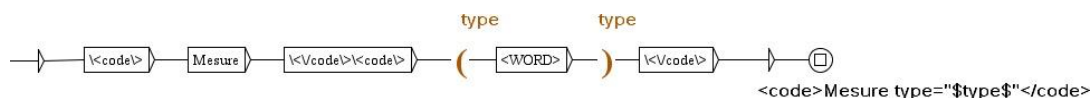
Enregistrons et compilons ce graphe.



Un graphe plus élégant est présenté en annexe (section 0, page 22) pour les utilisateurs avancés.

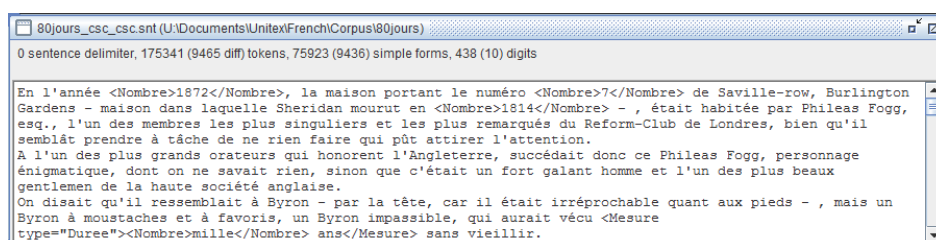
3.2.4 Des types à placer uniquement sur les mesures

Montrons combien la cascade de synthèse permet un résultat "sur mesure", avec deux autres présentations des (mêmes) résultats : imaginons que nous ne souhaitions pas placer de type sur la balise Nombre, mais seulement sur la balise Mesure. Une solution simple serait d'écrire un graphe *typeMesure.grf* en remplaçant le code <WORD> par le mot *Mesure* et, donc, en supprimant la variable \$cat\$.



Enregistrons et compilons ce graphe, puis ajoutons-le en deuxième position à la cascade *synthese.csc*, en mode *Replace*. Sur la ligne du graphe *type.grf*, cochons la case *Disabled*, qui nous permet de choisir les graphes actifs ou non dans la cascade.

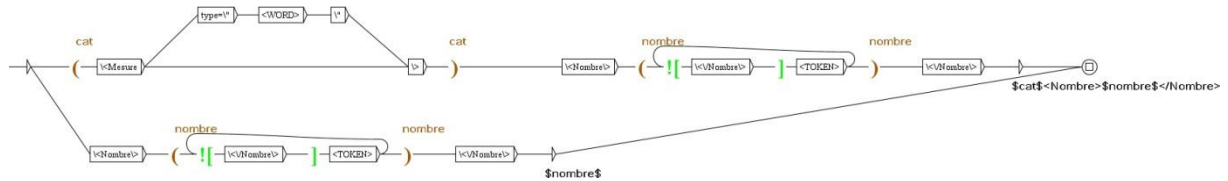
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz.
1	<input checked="" type="checkbox"/>	type.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	typeMesure.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>



²⁵ Cette remarque ne concerne pas les sorties pour lesquelles aucune protection n'est nécessaire.

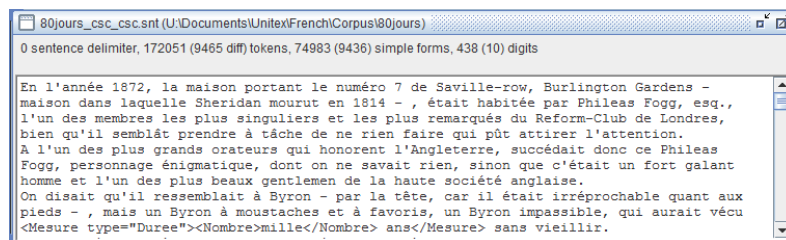
3.2.5 Suppression du balisage des nombres hors mesure

Ajoutons une contrainte : ne baliser les nombres que lorsqu'ils font partie d'une mesure. Le graphe *baliseNombre.grf* ne recopie les balises *<Nombres>* que si elles sont précédées d'une balise *<Mesure>*.



Enregistrons et compilons ce graphe, puis ajoutons-le en dernier à la cascade *synthese.csc*.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz.
	<input checked="" type="checkbox"/>	type.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	typeMesure.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	baliseNombre.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



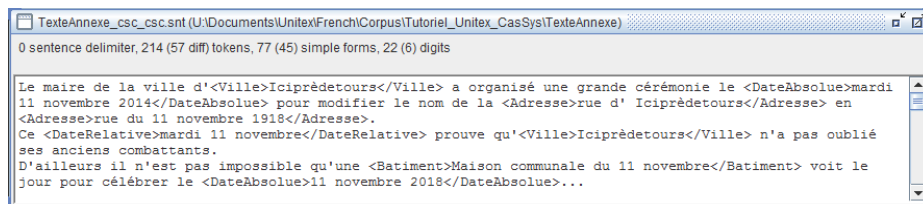
Un graphe plus élégant est présenté en annexe (section 4.3, page 22) pour les utilisateurs avancés.

4 Annexe pour utilisateurs avancés

4.1 Complément de la section 1.2.2, page 8

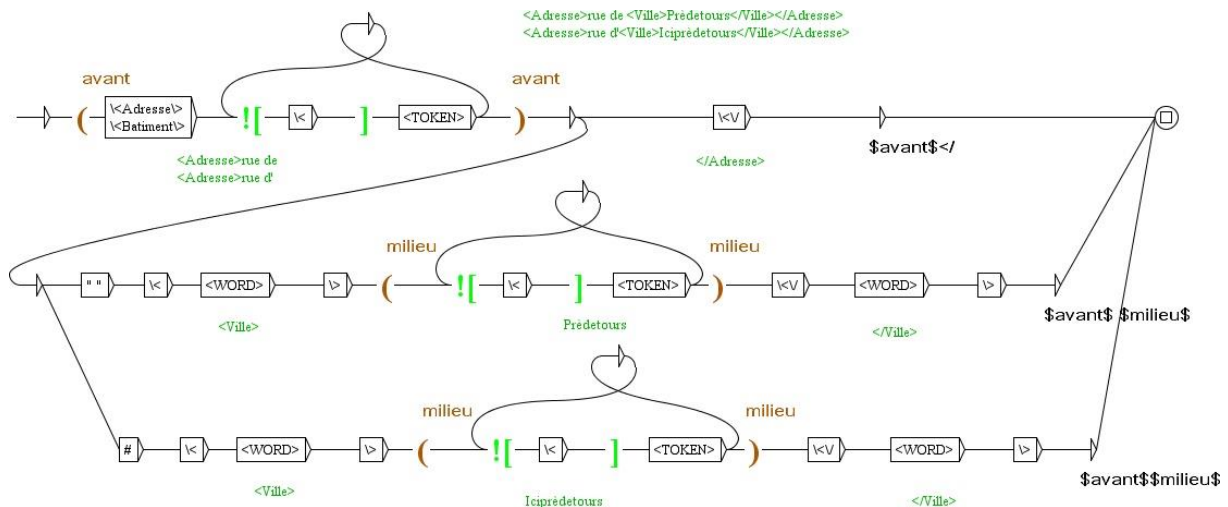
Reprenons la première cascade de synthèse et le graphe *suppressionInterne*. Un problème va surgir s'il n'y a pas d'espace entre les variables \$avant\$ et \$milieu\$, par exemple dans le cas d'une apostrophe. Considérons le même texte où le nom de la ville a été changé en *Iciprêdetours*. Ce texte se trouve²⁶ dans le dossier *French\Corpus\Tutoriel_Unitex_CasSys\TexteAnnexe*.

Pour lancer l'analyse, commençons par ouvrir, toujours dans le menu *Text*, le fichier *French\Corpus\Tutoriel_Unitex_CasSys\TexteAnnexe\Texte.txt*, en répondant *Yes* à la question *Do you want to preprocess the text* et en décochant les deux premières coches pour ne laisser que l'option *Apply all default Dictionaries*. Puis passons la cascade d'analyse. Ouvrons ensuite, dans le menu *Text*, le fichier *French\Corpus\Tutoriel_Unitex_CasSys\TexteAnnexe\Texte_csc.txt*, en répondant *No* à la question *Do you want to preprocess the text*. Passons la cascade de synthèse. Le résultat obtenu est le fichier *Texte_csc_csc.txt*



Une erreur s'est produite : *<Adresse>rue d'Icprêdetours</Adresse>* avec un espace entre l'apostrophe et le nom.

Une première solution pour éviter cela consiste à créer un graphe *suppressionInterneSiPasEspace.grf* qui duplique les chemins en fonction ou non de la présence ou de l'absence de l'espace, en utilisant le symbole # qui signifie "pas d'espace entre les boîtes"²⁷ et le code " " qui désigne un espace.

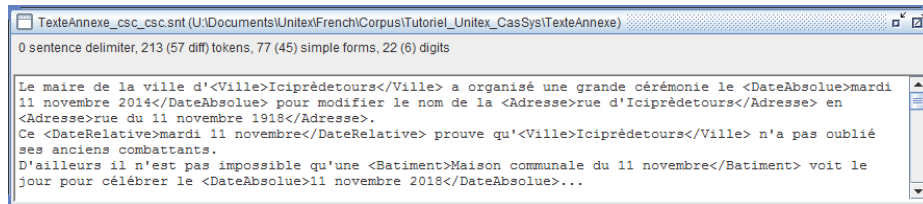


Si le graphe *suppressionInterneSiPasEspace* remplace dans la cascade le graphe *suppressionInterne*, nous obtenons le résultat souhaité, *<Adresse>rue d'Icprêdetours</Adresse>* sans espace entre l'apostrophe et le nom.

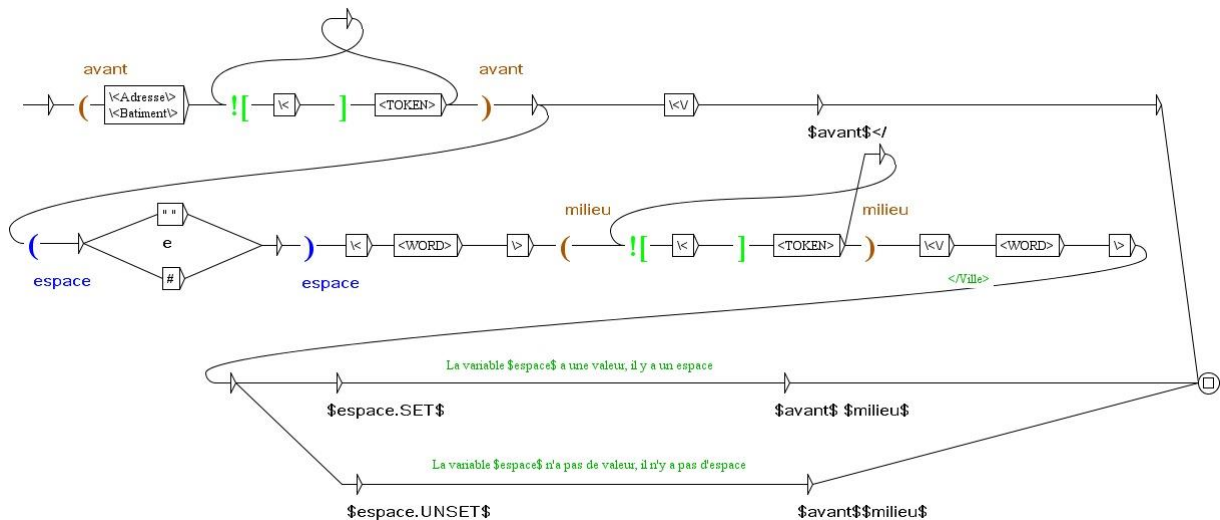
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	<input type="checkbox"/>	balisage.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	suppressionInterne.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	suppressionInterneSiPasEspace.fst2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

²⁶ Si vous n'avez pas pu télécharger la préparation du tutoriel, il faut tout d'abord se placer dans le dossier *Corpus\Tutoriel_Unitex_CasSys* et y créer un dossier nommé *TexteAnnexe*. Puis copier le fichier *texte.txt* du dossier *TexteBrut* et le coller dans le dossier *TexteAnnexe*. Ouvrir ensuite ce fichier via l'éditeur d'Unitex (menu *File Edition/New File*) et remplacer *Prêdetours* par *Icprêdetours*. Et, enfin, enregistrer le fichier et fermer l'éditeur.

²⁷ Voir le manuel, section 4.3.1.



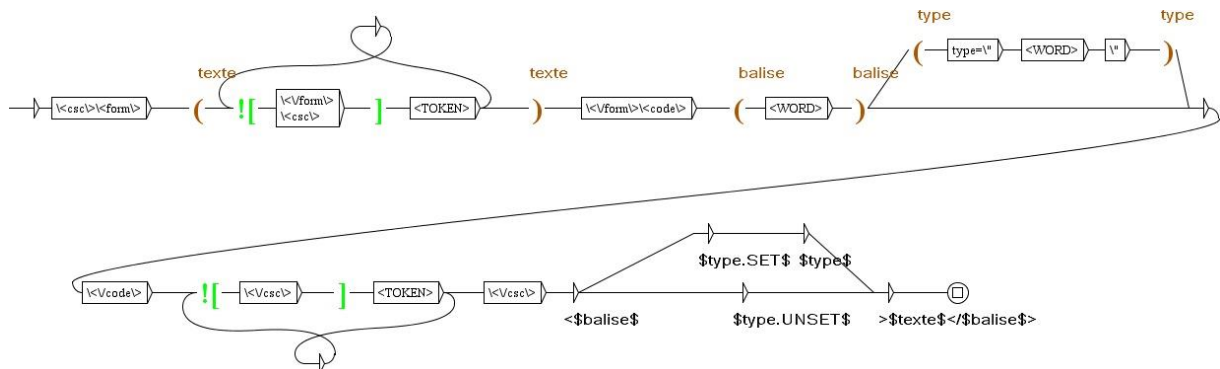
Une autre solution, plus élégante, mais plus complexe, permet d'éviter la duplication des chemins : il s'agit de faire un test sur une variable de sortie²⁸. Le test qui permet de savoir si la variable `$espace` possède une valeur ou non utilise les sorties `$espace.SET$` et `$espace.UNSET$`²⁹.



Si le graphe *suppressionInterneAvecTest* remplace dans la cascade le graphe *suppressionInterneSiPasEspace*, nous obtenons à nouveau le résultat souhaité, `<Adresse>rue d'Iciprèdetours</Adresse>` sans espace entre l'apostrophe et le nom.

4.2 Complément de la section 3.2.3, page 19

De même, on pourrait remplacer le graphe *balise.grf* par un graphe plus élégant, en utilisant un test plutôt qu'une répétition. C'est le graphe *baliseTest.grf*.



4.3 Complément de la section 3.2.5, page 20

Comme à la section précédente, nous pouvons utiliser un test.

²⁸ Voir le manuel, section 6.8.

²⁹ Voir le manuel, section 6.9.1.

